Brief Introduction to Next Generation Sequencing Bioinformatics

Ricardo González Méndez, PhD

UPR School of Medicine

Introduction to Next Generation Sequencing

Alex Sánchez

GRUP DE RECERCA EN ESTADÍSTICA I BIOINFORMÀTICA

Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barelona

UNIVERSITAT DE BARCELONA

Vall d'Hebron Institut de Recerca

Statistics and Bioinformatics Unit
Vall d'Hebron Institut de Recerca

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

From slideshare.net

PITTSBURGH SUPERCOMPUTING CENTER BIOINFORMATICS UNIVERSITY
NATIONAL RESOURCE FOR BIOMEDICAL SUPERCOMPUTING

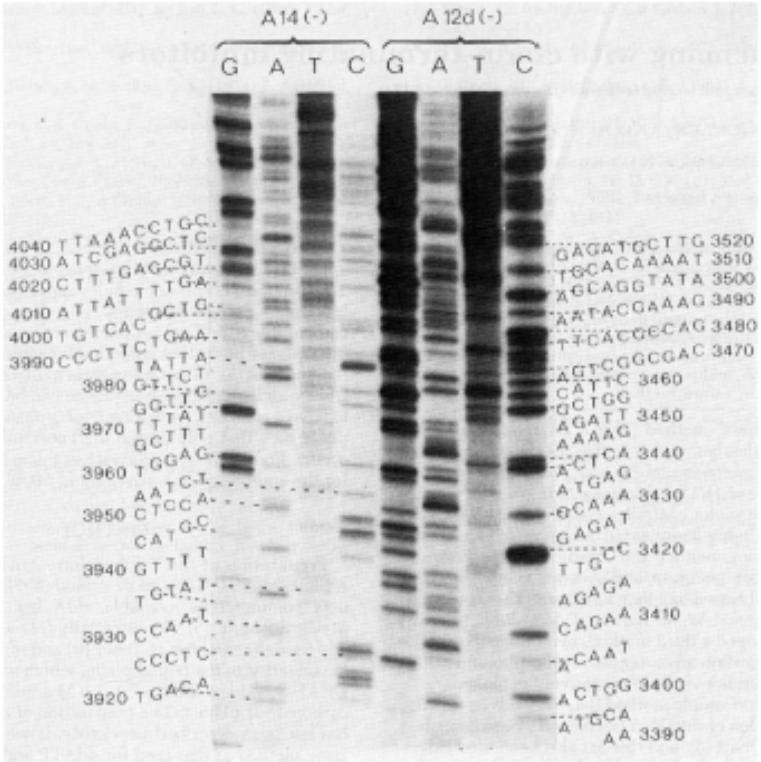
Why is NGS revolutionary?

- NGS has brought high-speed not only to genome sequencing and personal medicine
- It has also changed the way we do genome research
- Got a question on genome organization?
- SEQUENCE IT !!!

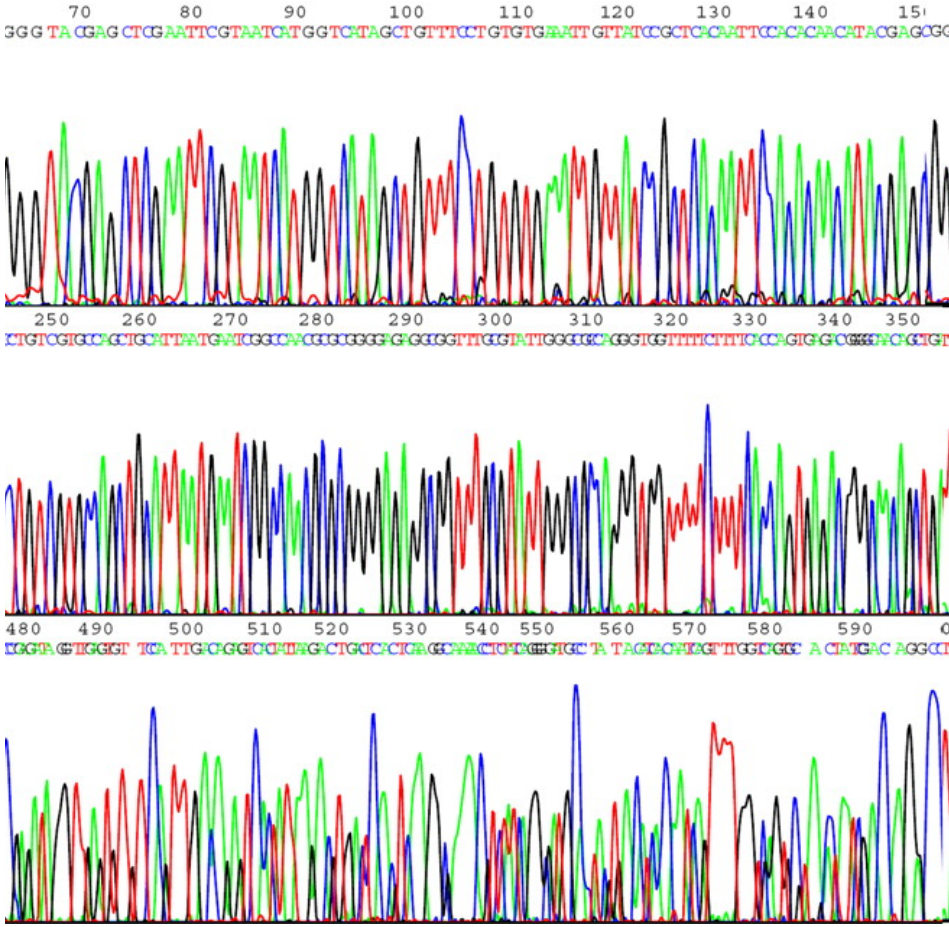
From slideshare.net

1977 (400 bases per week)

1997 (400-500 bases per day)



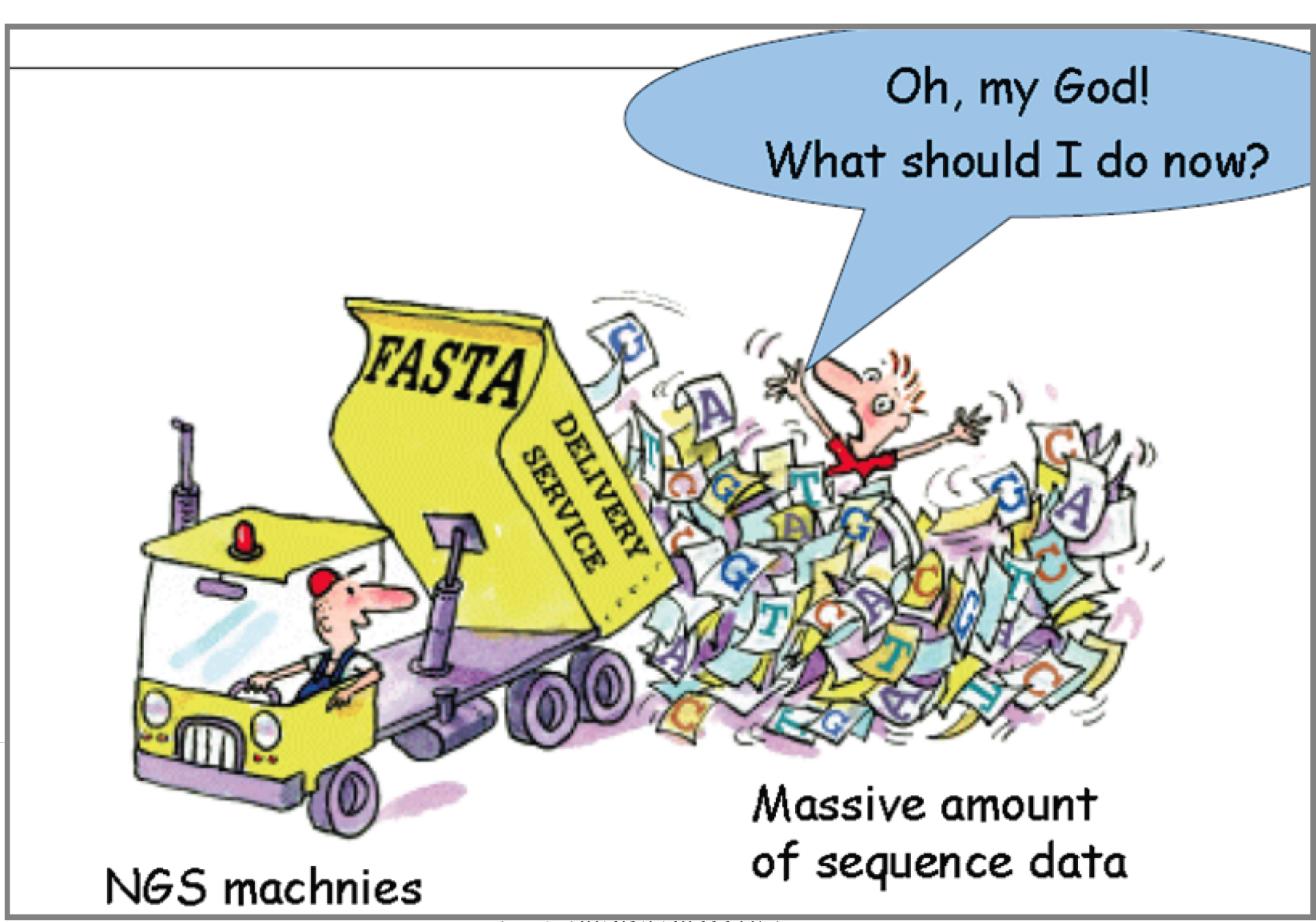
1. Autoradiograph of the acrylamide gel from the sequence determination using restriction fragments A12d and A14 as prime
ementary strand of ϕX174 DNA. The inhibitors used were (left to right) ddGTP, ddATP, ddTTP, and araCTP. Electrophoresi
crylamide gel at 40 mA for 14 hr. The top 10 cm of the gel is not shown. The DNA sequence is written from left to right and upwar
responding bands on the radioautograph. The numbering is as given in ref. 2.



Ion S5 XL System

Simple, rapid workflow for panels, microbes, exomes, and transcriptomes – ideal for higher weekly throughput

		Ion 520 Chip	Ion 530 Chip	Ion 540 Chip	Ion 520 Chip	Ion 530 Chip	Ion 540 Chip
Output*	**200 bp**	0.6–1 Gb	3–4 Gb	10–15 Gb	0.6–1 Gb	3–4 Gb	10–15 Gb
	400 bp	1.2–2 Gb	6–8 Gb	–	1.2–2 Gb	6–8 Gb	–
Reads		3–5 million	15–20 million	60–80 million	3–5 million	15–20 million	60–80 million
Run time	**200 bp**	2.5 hr	2.5 hr	2.5 hr	2.5 hr	2.5 hr	2.5 hr
	400 bp	4 hr	4 hr	–	4 hr	4 hr	–
Analysis time	**200 bp**	5 hr	8 hr	16.5 hr	1 hr	2.5 hr	5 hr
	400 bp	8 hr	17.5 hr	–	2 hr	4 hr	–



From slideshare.net

BIOINFORMATICS CHALLENGES OF NGS

From slideshare.net

NGS pushes bioinformatics needs up!

- Need for large amount of CPU power
  - Informatics groups must manage compute clusters
  - Challenges in parallelizing existing software or redesign of algorithms to work in a parallel environment
  - Another level of software complexity and challenges to interoperability
- VERY large text files (~10 million lines long)
  - Can't do 'business as usual'with familiar tools such as Perl/Python.
  - Impossible memory usage and execution time
  - Impossible to browse for problems
- Need sequence Quality Filtering

From slideshare.net

Data management issues

- Raw data are large. How long should be kept?
- Processed data are manageable for most people
  - 20 million reads (50bp) ~1Gb
- More of an issue for a facility: Hi-Seq recommends 32 CPU cores, each with 4GB RAM
- Certain studies much more data intensive than other
  - Whole genome sequencing
    - A 30X coverage genome pair (tumor/normal) ~500 GB
    - 50 genome pairs ~ 25 TB

From slideshare.net

So what?

- In NGS we have to process really big amounts of data, which is not trivial in computing terms.
- Big NGS projects require supercomputing infrastructures
- Or put another way: it's not the case that anyone can study everything.
  - Small facilities must carefully choose their projects to be scaled with their computing capabilities

From slideshare.net

Computational infrastructure for NGS

There is great variety but a good point to start with:

- Computing cluster
  - Multiple nodes(servers) with of course multiple cores
  - High performance storage (TB, PB level)
  - Fast networks (10Gb ethernet, Infiniband)
- `Enough space and conditions for the equipment ("servers room")
- Skilled people (sysadmin, developers)

From slideshare.net

Starting at $125,000, the knoSYS™100 is based on Knome's big data informatics technology. The system will accept next generation sequence data from leading sequencers, including those sold by Illumina, Life Technologies, and Complete Genomics.

© 2013

Small infrastructure

- Recommended at least 2 machines
  - 8 or 12 cores each machine.
  - 48 Gb RAM minimum each machine.
    - 4 Gb per core minimum
  - BIG local disk. At least 4TB each machine
    - As much local disks as we can afford

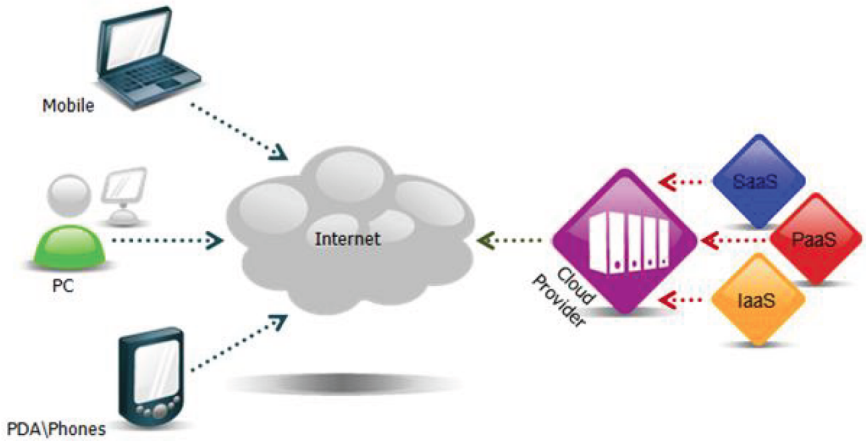
Price range: starting at $12,000 -$18,000

(2 machines)

From slideshare.net

Alternatives (1): Cloud Computing

- Pros
  - Flexibility.
  - You pay what you use.
  - Don´t need to maintain a data center.
- Cons
  - Transfer big datasets over internet is slow.
  - You pay for consumed bandwidth. That is a problem with big datasets.
  - Lower performance, specially in disk read/write.
  - Privacy/security concerns.
  - More expensive for big and long term projects.

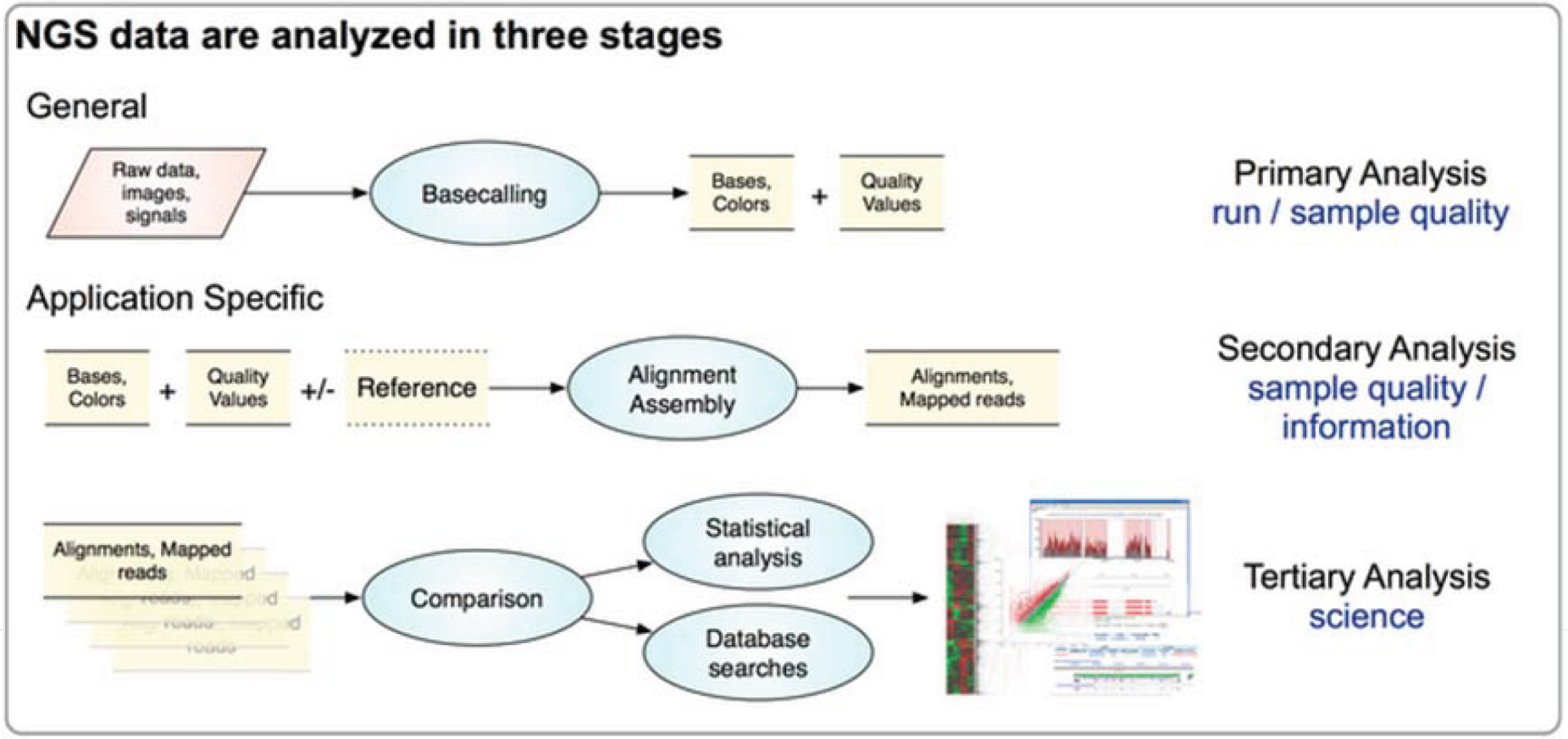


From slideshare.net

NGS DATA ANALYSIS

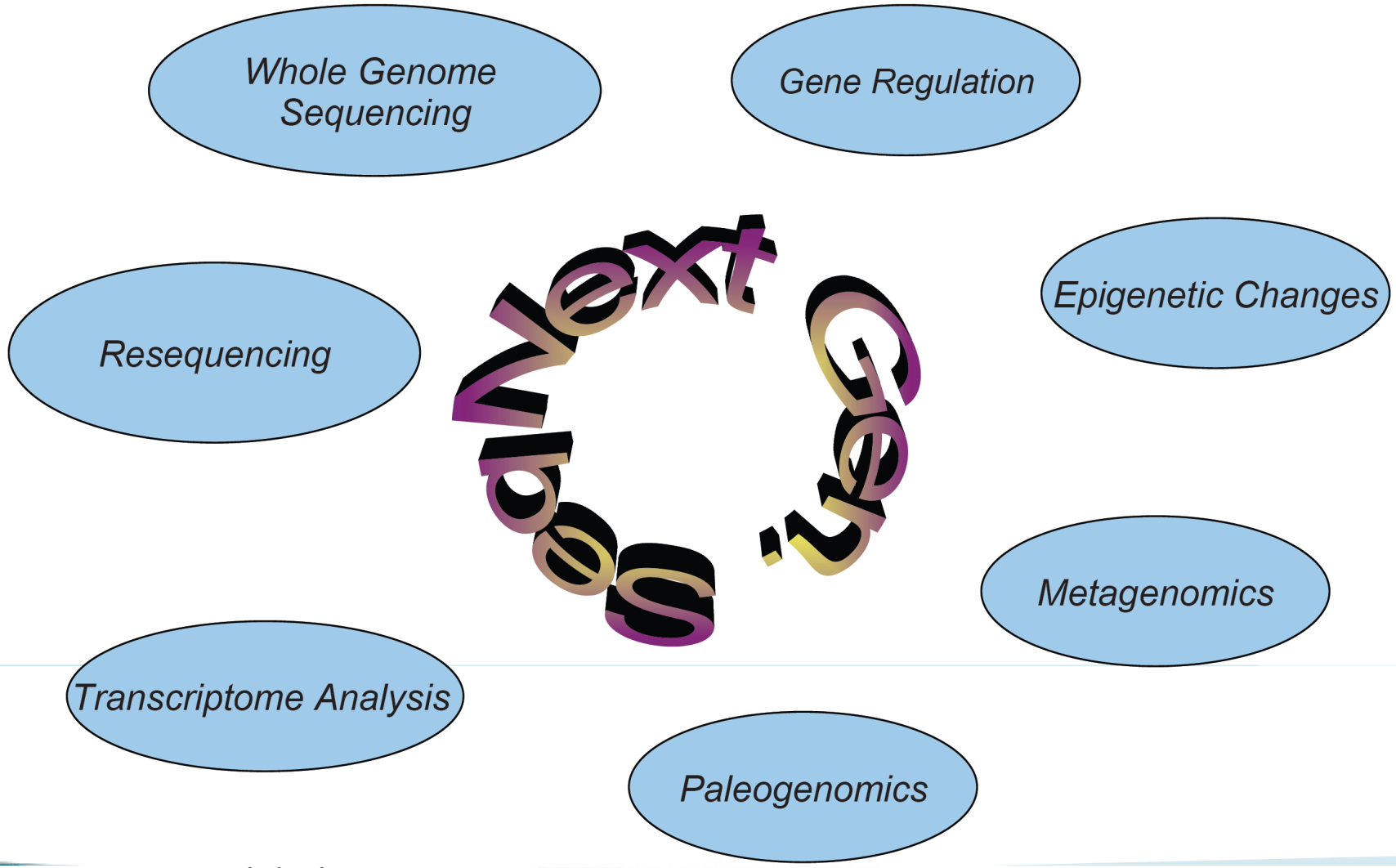
From slideshare.net

NGS data analysis stages



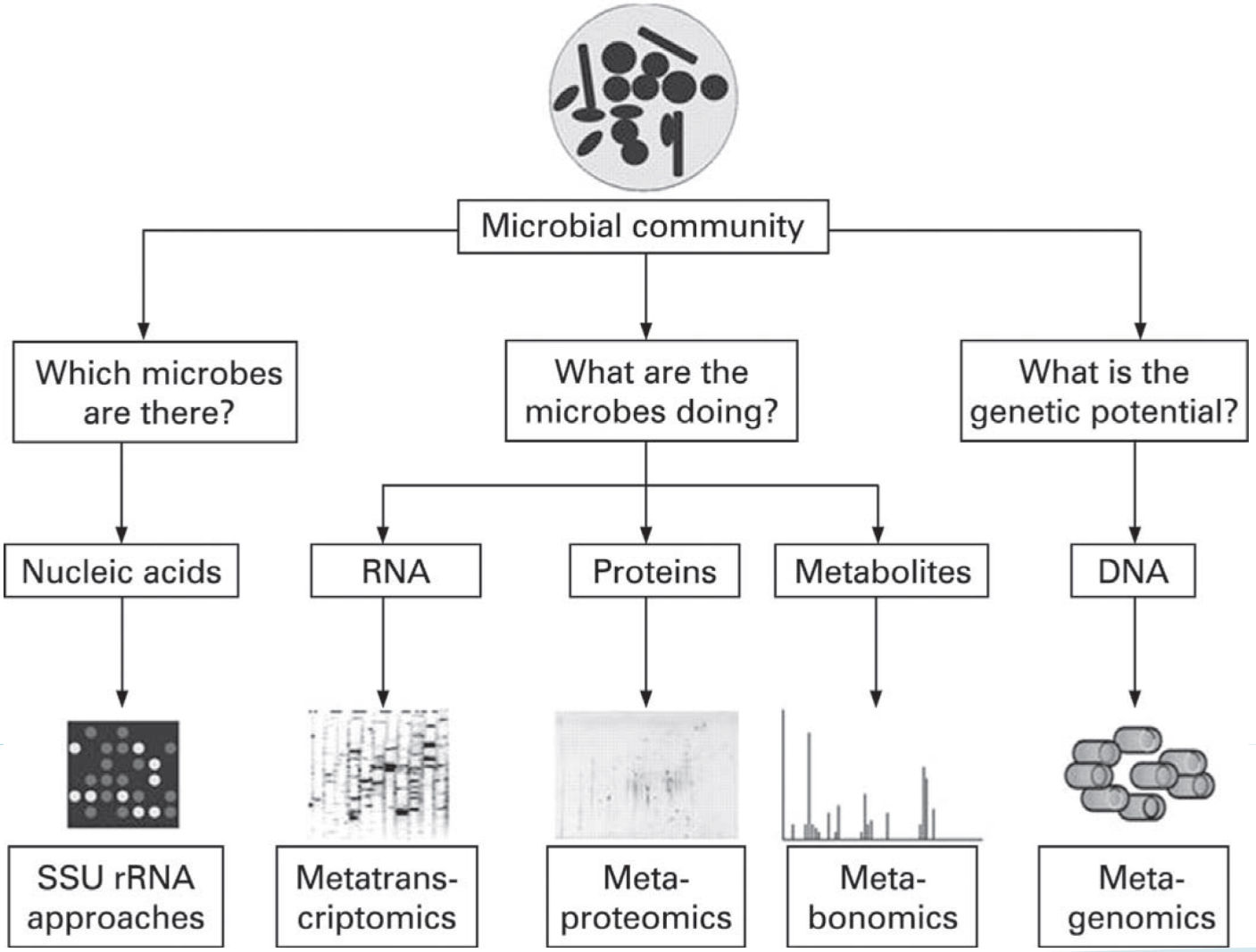
From slideshare.net

NGS Applications are ***sequencing*** applications



From slideshare.net

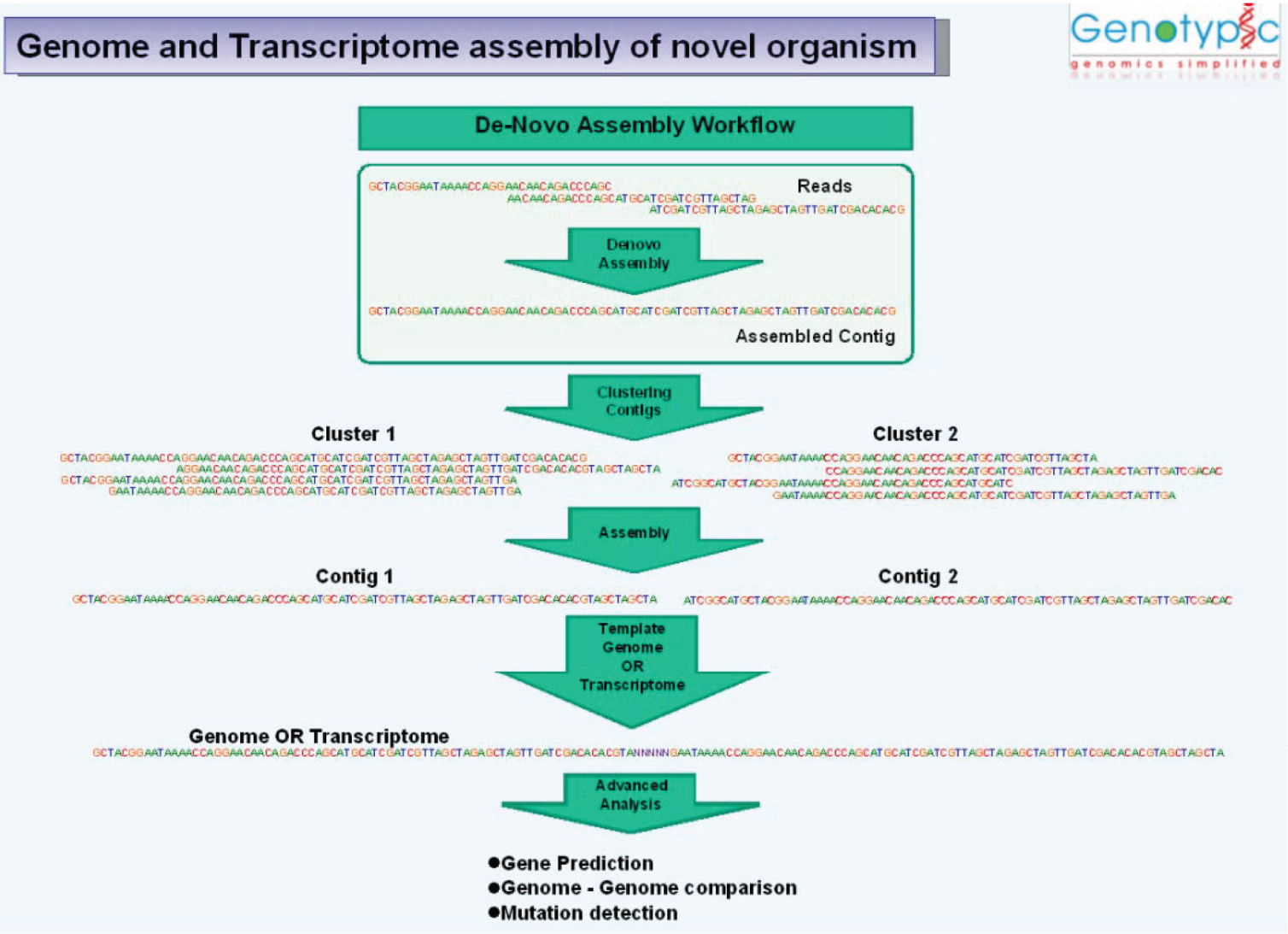
Metagenomics and other community-based "omics"



Zoetendal E G et al.
Gut 2008;57:1605-1615

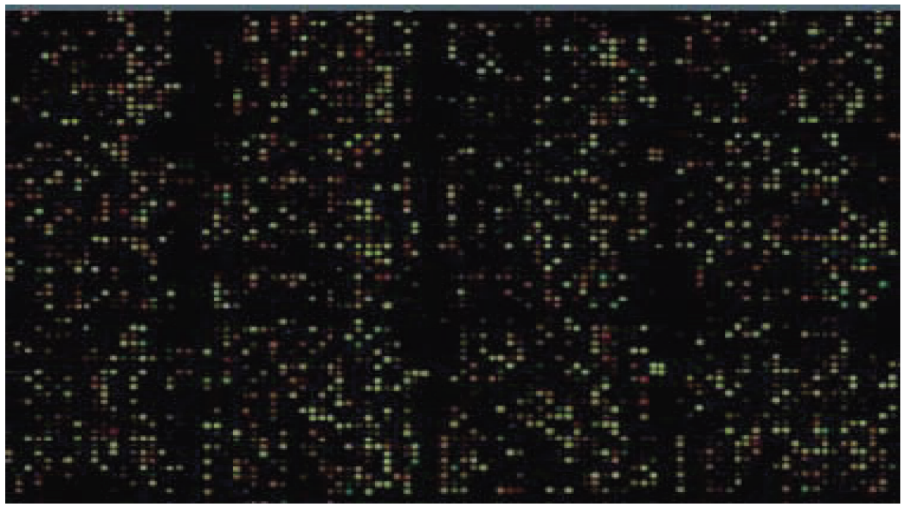
From slideshare.net

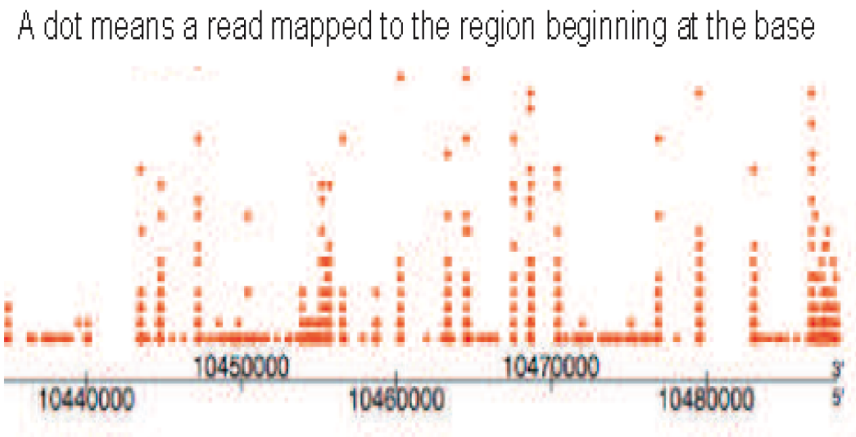
De novo sequencing



From slideshare.net

Transcriptomics by NGS: RNASeq



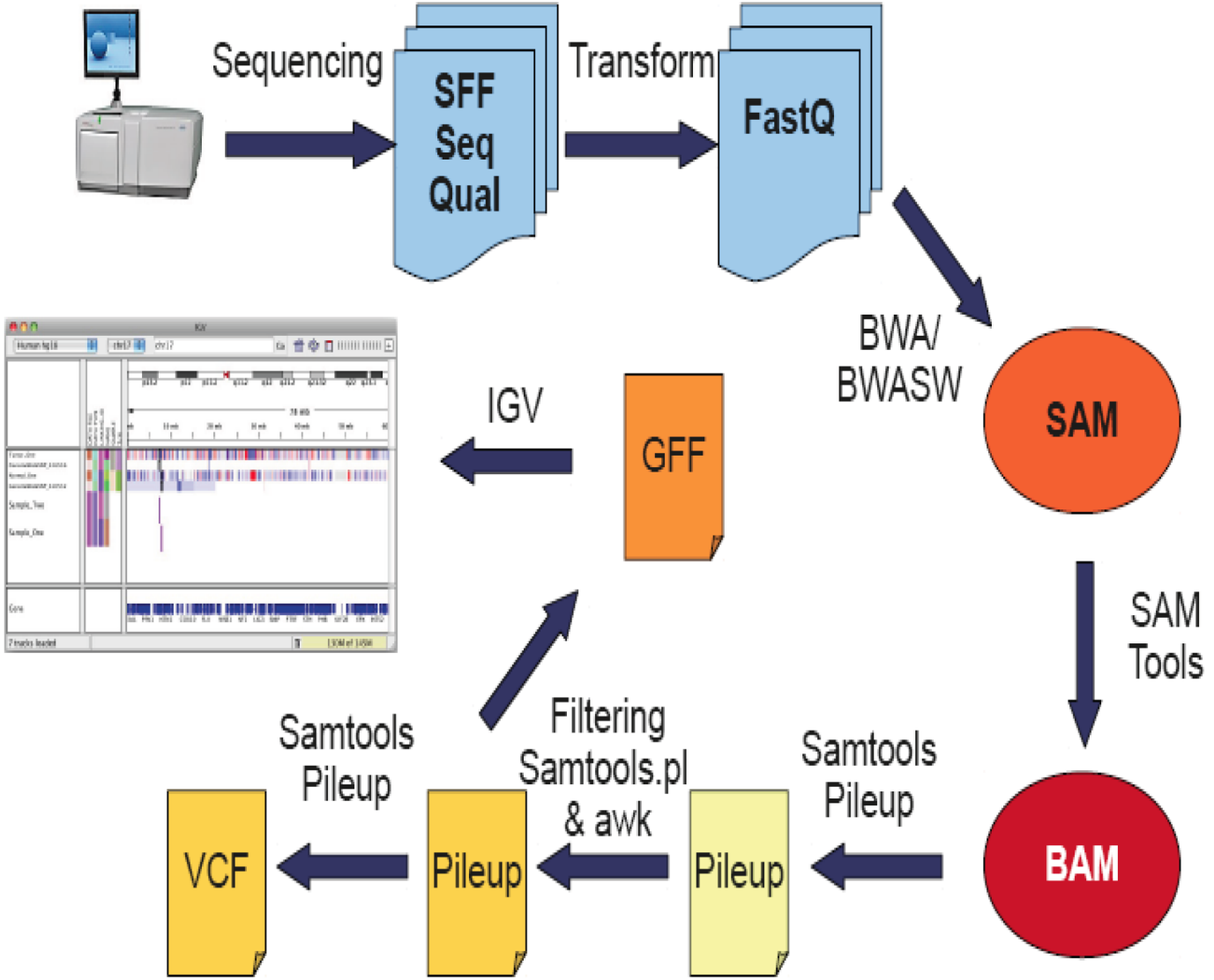


- Analog Signal
  - Easy to convey the signal's information
  - Continuous strength
  - Signal loss and distortion

- Digital Signal
  - Harder to achieve & interpret
  - Reads counts: discrete values
  - Weak background or no noise

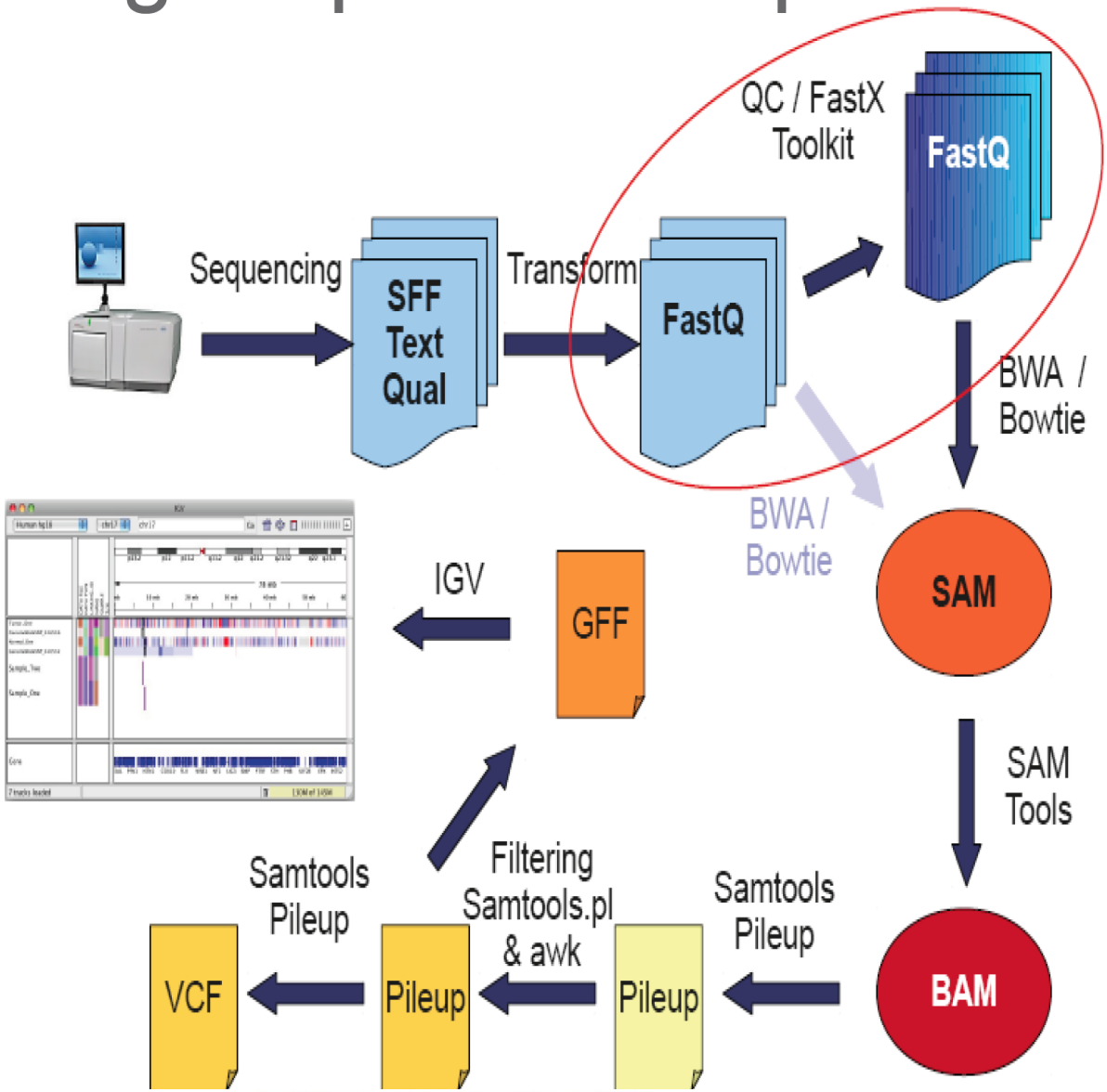
From slideshare.net

A typical workflow (Seq-to-variant wf)



From slideshare.net

Preprocessing sequences improves results



From slideshare.net

Sequence formats

- **FastA** format (everybody knows about it)
  - Header line starts with “>” followed by a sequence ID
  - Sequence (string of nt).

- **FastQ** format (http://maq.sourceforge.net/fastq.shtml)
  - First is the sequence (like Fasta but starting with “@”)
  - Then “+” and sequence ID (optional) and in the following line are QVs encoded as single byte ASCII codes
    - Different quality encode variants

- Nearly all downstream analysis take **FastQ** as input sequence

From slideshare.net

Why QC and preprocessing

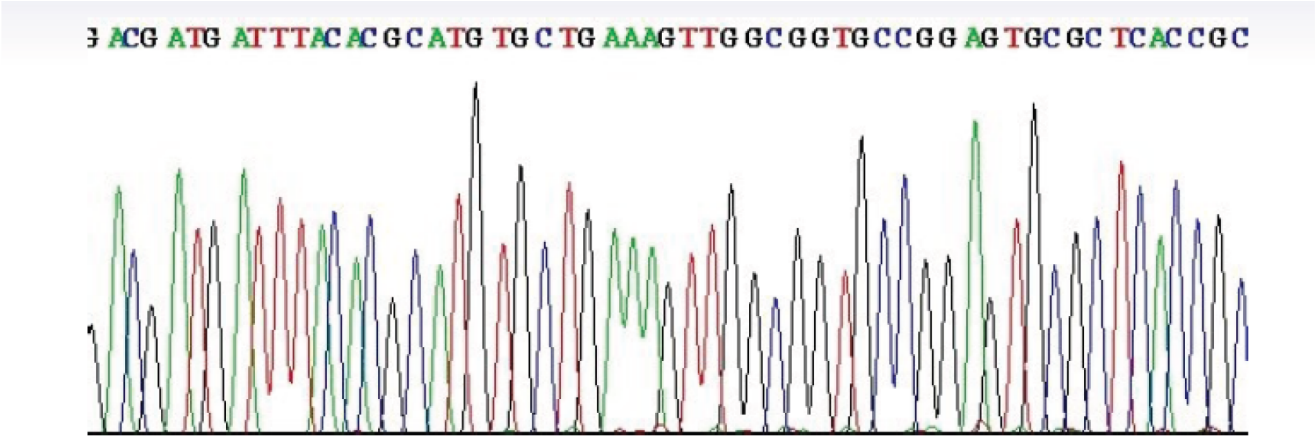
- Sequencer output:
  - Reads + **quality**
- Natural questions
  - Is the quality of my sequenced data OK?
  - If something is wrong can I fix it?
- **Problem**: **HUGE** files… How do they look?
- Files are flat files and are big… tens of Gbs (even hard to browse them)

```
@HWUSI-EAS460:2:1:368:1089#0/1
TACGTACGTACGTACGTACGTAGATCGGAAGAGCGG
+HWUSI-EAS460:2:1:368:1089#0/1
aa[a_a_a^a^a]VZ]R^P[]YNSUTZBBBBBBBBB

@HWUSI-EAS460:2:1:368:528#0/1
CTATTATAATATGACCGACCAGCTAGATCTACAGTC
+HWUSI-EAS460:2:1:368:528#0/1
abbbbaaaabba^aa`Y``aa`aaa``a`a_\_`[_
```

From slideshare.net

How is quality measured?



- Assign quality score to each peak
- The frequently used Phred scores provide log(10)-transformed error
- probability values:
  - score = 20 corresponds to a 1% error rate
  - score = 30 corresponds to a 0.1% error rate
  - score = 40 corresponds to a 0.01% error rate
- The base calling (A, T, G or C) is performed based on Phred scores.
- Ambiguous positions with Phred scores <= 20 are labeled with N.

From slideshare.net

Major categories of high-throughput genomic data analysis

- **Identify differentially expressed (DE) genes**

Purpose: Detect candidate markers related (associated) to clinical outcome for early diagnosis or treatment.

- **Classification (prediction) analysis**

Purpose: Construct a classification model that can predict future patients (diagnosis, prognosis outcome, drug response etc).

- **Cluster analysis**

Purpose: Cluster samples to identify unknown disease subtypes; cluster genes to identify potential gene regulatory modules.

- **Gene set (pathway) analysis**

Purpose: Functional annotation of identified biomarker list by enriched pathways.

- **Meta-analysis**

Purpose: Combine multiple studies to increase statistical power and obtain validated conclusion.

PITTSBURGH SUPERCOMPUTING CENTER
BIOINFORMATICS UNIVERSITY
NATIONAL RESOURCE FOR BIOMEDICAL SUPERCOMPUTING

Statistically significant versus Biologically significant

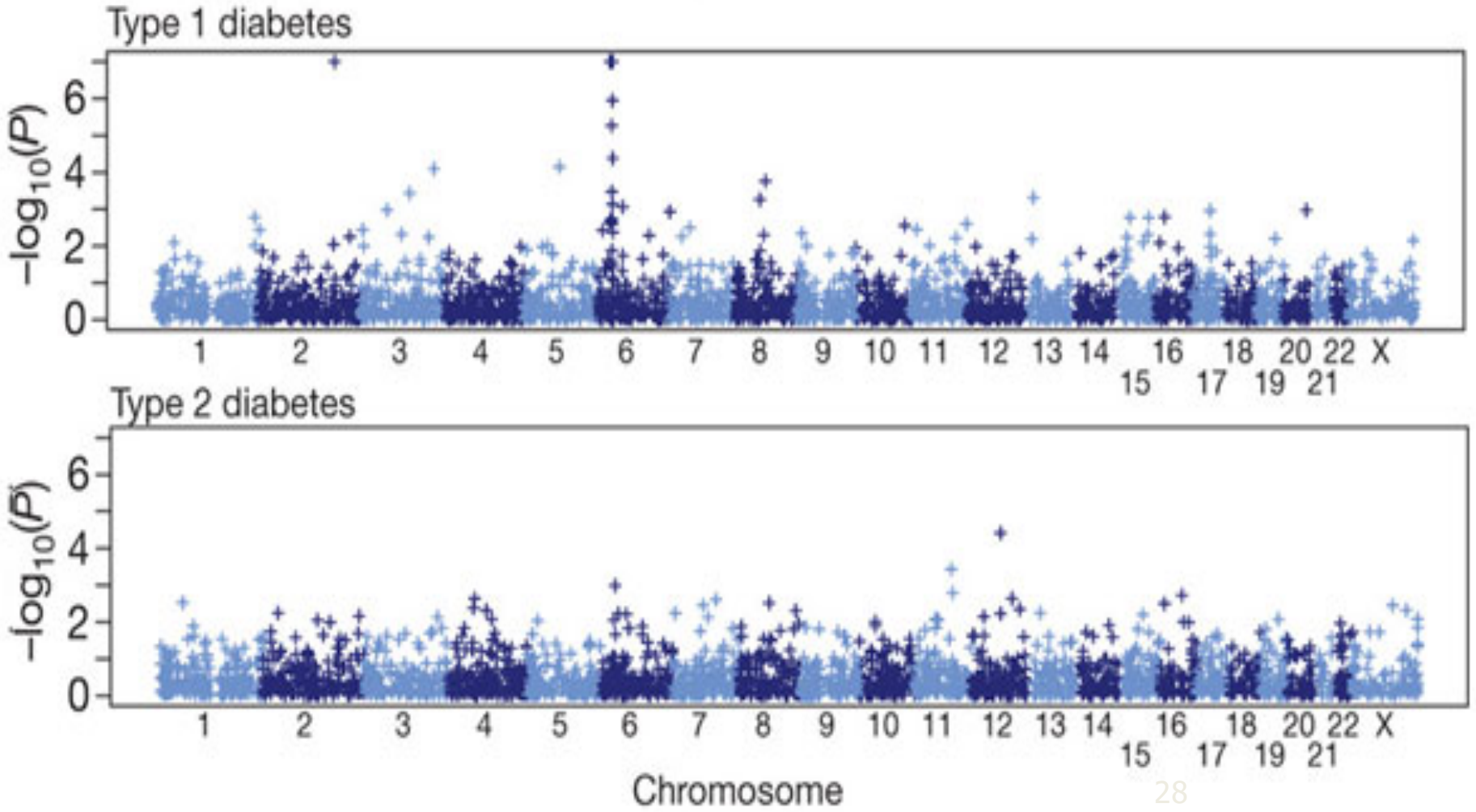
Statistical significance:

the observed difference is unlikely to be due to chance.

Scientific or Biological significance:

the observed level of difference is of sufficient magnitude to be of biological relevance.

Genome Wide Association Studies

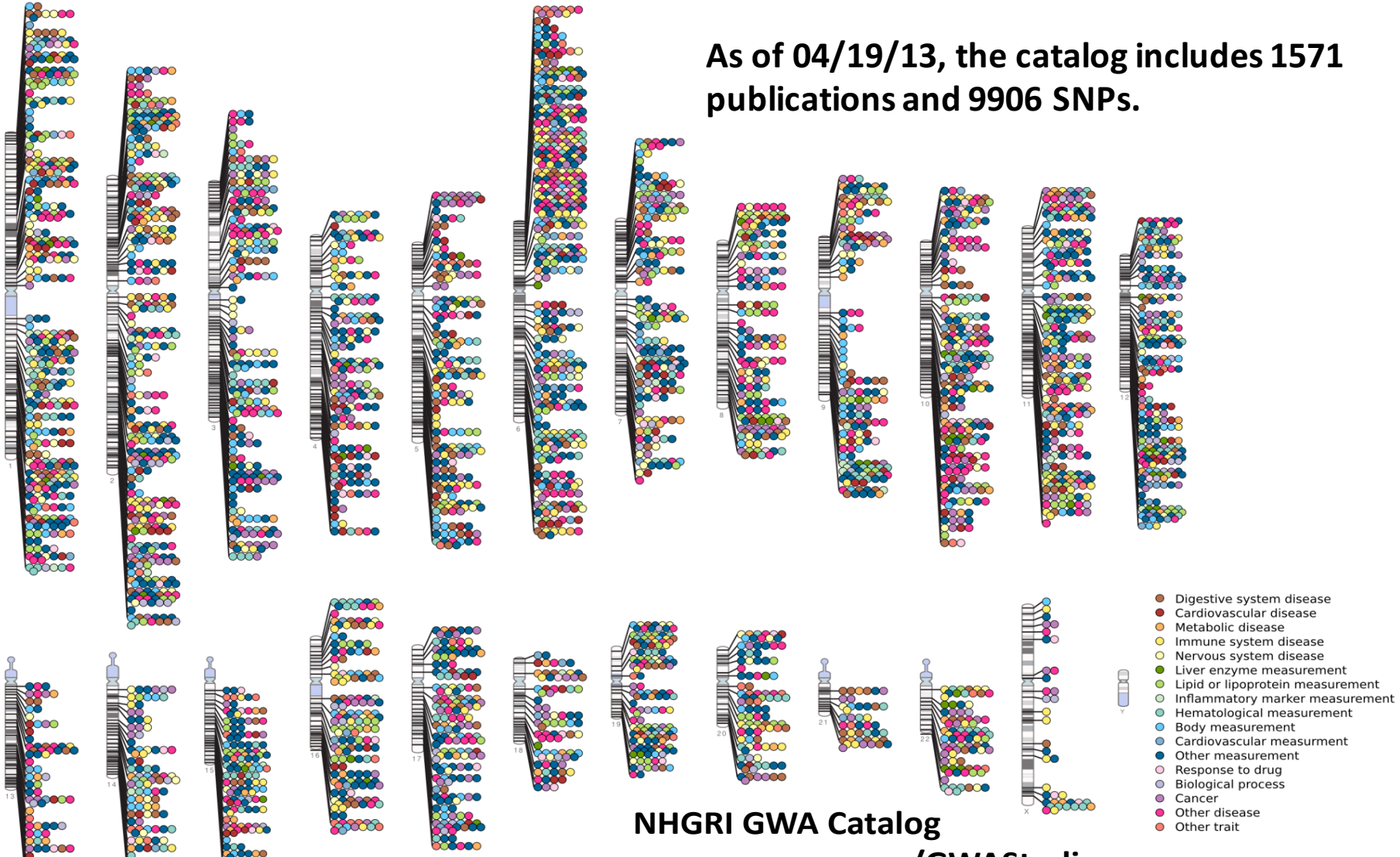


Ricardo Gonzalez Mendez, PhD © 2013

BIOINFORMATICS UNIVERSITY
NATIONAL RESOURCE FOR BIOMEDICAL SUPERCOMPUTING

Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

As of 04/19/13, the catalog includes 1571 publications and 9906 SNPs.



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

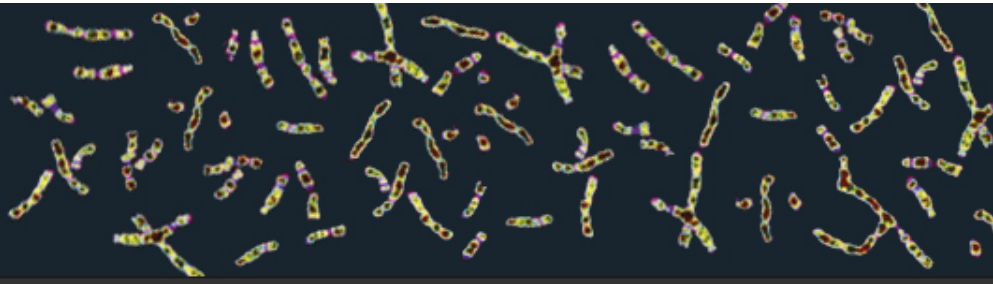
EMBL-EBI

Ricardo Gonzalez Mendez, PhD © 2013

PITTSBURGH SUPERCOMPUTING CENTER
BIOINFORMATICS UNIVERSITY
NATIONAL RESOURCE FOR BIOMEDICAL SUPERCOMPUTING

1000 Genomes

A Deep Catalog of Human Genetic Variation



LATEST ANNOUNCEMENTS

WEDNESDAY OCTOBER 31, 2012

An integrated map of genetic variation from 1092 human genomes

The Phase 1 publication, An Integrated map of genetic variation from 1092 human genomes is now available from Nature and can be downloaded directly from the ftp site. The paper is distributed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. Please share our paper appropriately.

All the data files associated with this paper can be found in our phase1 analysis results directory.

NAVIGATION

- Frequently Asked Questions

LINKS

All Project Announcements

Sample and Project Information

PROJECT OVERVIEW

The 1000 Genomes Project is an international collaboration to produce an extensive public catalog of human genetic variation, including SNPs and structural variants, and their haplotype contexts. This resource will support genome-wide association studies and other medical research studies.

The genomes of about 2500 unidentified people from about 25 populations around the world will be sequenced using next-generation sequencing technologies. The results of the study will be freely and publicly accessible to researchers worldwide.

Further information about the project is available in the About tab. Information about downloading, browsing or using the 1000 Genomes data is available in the Data tab.

IN THIS ISSUE

- This Week In JAMA
- Viewpoint
- A Piece Of My Mind
- Original Contribution
- Special Communication
- Editorial
- Letters
- Medical News And Perspectives
- From JAMA's Daily News Site
- Capitol Health Call
- From The Centers For Disease Control And Prevention
- Poetry And Medicine
- JAMA 100 Years Ago
- Correction
- JAMA Patient Page
- JAMA Masthead
- Appendix

Variants in the ATP-Binding Cassette Transporter (*ABCA7*), Apolipoprotein E ε4,and the Risk of Late-Onset Alzheimer Disease in African Americans PDF

Christiane Reitz, MD, PhD; Gyungah Jun, PhD; Adam Naj, PhD; Ruchita Rajbhandary, MPH; Badri Narayan Vardarajan, PhD; Li-San Wang, PhD; Otto Valladares, MS; Chiao-Feng Lin, PhD; Eric B. Larson, MD, MPH; Neill R. Graff-Radford, MD; Denis Evans, MD; Philip L. De Jager, MD, PhD; Paul K. Crane, MD, MPH; Joseph D. Buxbaum, PhD; Jill R. Murrell, PhD; Towfique Raj, PhD; Nilufer Ertekin-Taner, MD, PhD; Mark Logue, PhD; Clinton T. Baldwin, PhD; Robert C. Green, MD, MPH; Lisa L. Barnes, PhD; Laura B. Cantwell, MPH; M. Daniele Fallin, PhD; Rodney C. P. Go, PhD; Patrick Griffith, MD; Thomas O. Obisesan, MD; Jennifer J. Manly, PhD; Kathryn L. Lunetta, PhD; M. Ilyas Kamboh, PhD; Oscar L. Lopez, MD; David A. Bennett, MD; Hugh Hendrie, MB, ChB, DSc; Kathleen S. Hall, PhD; Alison M. Goate, PhD; Goldie S. Byrd, PhD; Walter A. Kukull, PhD; Tatiana M. Foroud, PhD; Jonathan L. Haines, PhD; Lindsay A. Farrer, PhD; Margaret A. Pericak-Vance, PhD; Gerard D. Schellenberg, PhD; Richard Mayeux, MD, MSc; for the Alzheimer Disease Genetics Consortium

Includes: Supplemental Content

Association Between *BRAF* V600E Mutation and Mortality in Patients With Papillary Thyroid Cancer PDF

Mingzhao Xing, MD, PhD; Ali S. Alzahrani, MD; Kathryn A. Carson, ScM; David Viola, MD; Rossella Elisei, MD; Bela Bendlova, PhD; Linwah Yip, MD; Caterina Mian, MD; Federica Vianello, MD; R. Michael Tuttle, MD; Eyal Robenshtok, MD; James A. Fagin, MD; Efisio Puxeddu, MD, PhD; Laura Fugazzola, MD; Agnieszka Czarniecka, MD; Barbara Jarzab, MD, PhD; Christine J. O'Neill, MBBS(Hons), MS; Mark S. Sywak, MD; Alfred K. Lam, MD, PhD; Garcilaso Riesco-Eizaguirre, MD, PhD; Pilar Santisteban, PhD; Hirotaka Nakayama, MD; Ralph P. Tufano, MD; Sara I. Pai, M.D, PhD; Martha A. Zeiger, MD; William H. Westra, MD; Douglas P. Clark, MD; Roderick Clifton-Bligh, PhD; David Sidransky, MD; Paul W. Ladenson, MD; Vlasta Sykorova, PhD

Includes: Supplemental Content

A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic *Escherichia coli* O104:H4 PDF

Nicholas J. Loman, MBBS, PhD; Chrystala Constantinidou, PhD; Martin Christner, MD; Holger Rohde, MD; Jacqueline Z.-M. Chan, PhD; Joshua Quick, BSc; Jacqueline C. Weir, MSci; Christopher Quince, PhD; Geoffrey P. Smith, PhD; Jason R. Betley, PhD; Martin Aepfelbacher, MD; Mark J. Pallen, MA, MD, PhD

Includes: Supplemental Content

ABSTRACT

Importance Genetic variants associated with susceptibility to late-onset Alzheimer disease are known for individuals of European ancestry, but whether the same or different variants account for the genetic risk of Alzheimer disease in African American individuals is unknown. Identification of disease-associated variants helps identify targets for genetic testing, prevention, and treatment.

Objective To identify genetic loci associated with late-onset Alzheimer disease in African Americans.

Design, Setting, and Participants The Alzheimer Disease Genetics Consortium (ADGC) assembled multiple data sets representing a total of 5896 African Americans (1968 case participants, 3928 control participants) 60 years or older that were collected between 1989 and 2011 at multiple sites. The association of Alzheimer disease with genotyped and imputed single-nucleotide polymorphisms (SNPs) was assessed in case-control and in family-based data sets. Results from individual data sets were combined to perform an inverse variance–weighted meta-analysis, first with genome-wide analyses and subsequently with gene-based tests for previously reported loci.

Main Outcomes and Measures Presence of Alzheimer disease according to standardized criteria.

Results Genome-wide significance in fully adjusted models (sex, age, *APOE* genotype, population stratification) was observed for a SNP in *ABCA7* (rs115550680, allele = G; frequency, 0.09 cases and 0.06 controls; odds ratio [OR], 1.79 [95% CI, 1.47-2.12]; $P = 2.2 \times 10^{-9}$), which is in linkage disequilibrium with SNPs previously associated with Alzheimer disease in Europeans ($0.8<D'<0.9$). The effect size for the SNP in *ABCA7* was comparable with that of the *APOE* ε4–determining SNP rs429358 (allele = C; frequency, 0.30 cases and 0.18 controls; OR, 2.31 [95% CI, 2.19-2.42]; $P = 5.5 \times 10^{-47}$). Several loci previously associated with Alzheimer disease but not reaching significance in genome-wide analyses were replicated in gene-based analyses accounting for linkage disequilibrium between markers and correcting for number of tests performed per gene (*CR1, BIN1, EPHA1, CD33;* $0.0005<$empirical $P < .001$).

Conclusions and Relevance In this meta-analysis of data from African American participants, Alzheimer disease was significantly associated with variants in *ABCA7* and with other genes that have been associated with Alzheimer disease in individuals of European ancestry. Replication and functional validation of this finding is needed before this information is used in clinical settings.

ABSTRACT

Importance Identification of the bacterium responsible for an outbreak can aid in disease management. However, traditional culture-based diagnosis can be difficult, particularly if no specific diagnostic test is available for an outbreak strain.

Objective To explore the potential of metagenomics, which is the direct sequencing of DNA extracted from microbiologically complex samples, as an open-ended clinical discovery platform capable of identifying and characterizing bacterial strains from an outbreak without laboratory culture.

Design, Setting, and Patients In a retrospective investigation, 45 samples were selected from fecal specimens obtained from patients with diarrhea during the 2011 outbreak of Shiga-toxigenic *Escherichia coli* (STEC) O104:H4 in Germany. Samples were subjected to high-throughput sequencing (August-September 2012), followed by a 3-phase analysis (November 2012-February 2013). In phase 1, a de novo assembly approach was developed to obtain a draft genome of the outbreak strain. In phase 2, the depth of coverage of the outbreak strain genome was determined in each sample. In phase 3, sequences from each sample were compared with sequences from known bacteria to identify pathogens other than the outbreak strain.

Main Outcomes and Measures The recovery of genome sequence data for the purposes of identification and characterization of the outbreak strain and other pathogens from fecal samples.

Results During phase 1, a draft genome of the STEC outbreak strain was obtained. During phase 2, the outbreak strain genome was recovered from 10 samples at greater than 10-fold coverage and from 26 samples at greater than 1-fold coverage. Sequences from the Shiga-toxin genes were detected in 27 of 40 STEC-positive samples (67%). In phase 3, sequences from *Clostridium difficile, Campylobacter jejuni, Campylobacter concisus,* and *Salmonella enterica* were recovered.

Conclusions and Relevance These results suggest the potential of metagenomics as a culture-independent approach for the identification of bacterial pathogens during an outbreak of diarrheal disease. Challenges include improving diagnostic sensitivity, speeding up and simplifying workflows, and reducing costs.

Using Galaxy on BioU

Alexander J. Ropelewski

ropelews@psc.edu

What is Galaxy

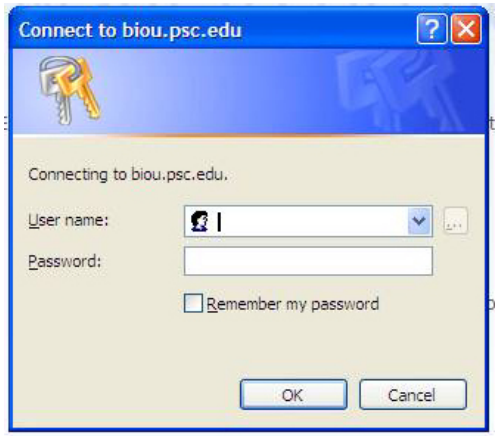
- Galaxy is an open-source web tool (portal) that organizes bioinformatics work-flows.
- Provides methods to:
  - Record workflows (history)
  - Share common sequence data (shared data libraries)
  - Import and use (private) sequence data
  - Share results with others (shared histories)
  - Integrate new tools into the interface
  - Run long analyses using batch queues
- For more information on Galaxy see the Galaxy wiki: http://wiki.g2.bx.psu.edu/

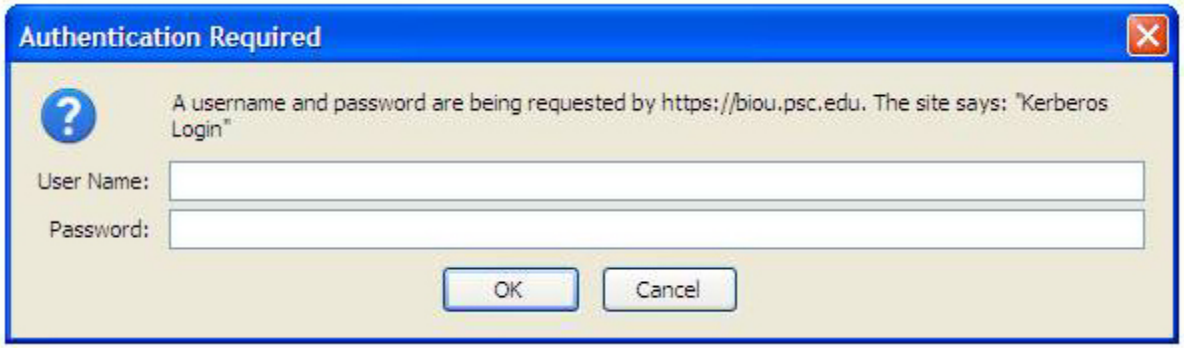
Galaxy vs. Writing Your Own PBS Scripts

Item Evaluated	Galaxy	PBS Scripts
Ease of Use	Intuitive for beginners	More difficult to learn
Amount of tools	Limited	Any tool that runs on Unix
Variety of tool options	Limited	Any option that the tool accepts
Length, size and configuration of run	Limited	Any valid configuration, length, size
Workflow log	Entire analysis	Job level log only
Style of use	Interactive, web page	Batch, command line

Using Galaxy on BioU

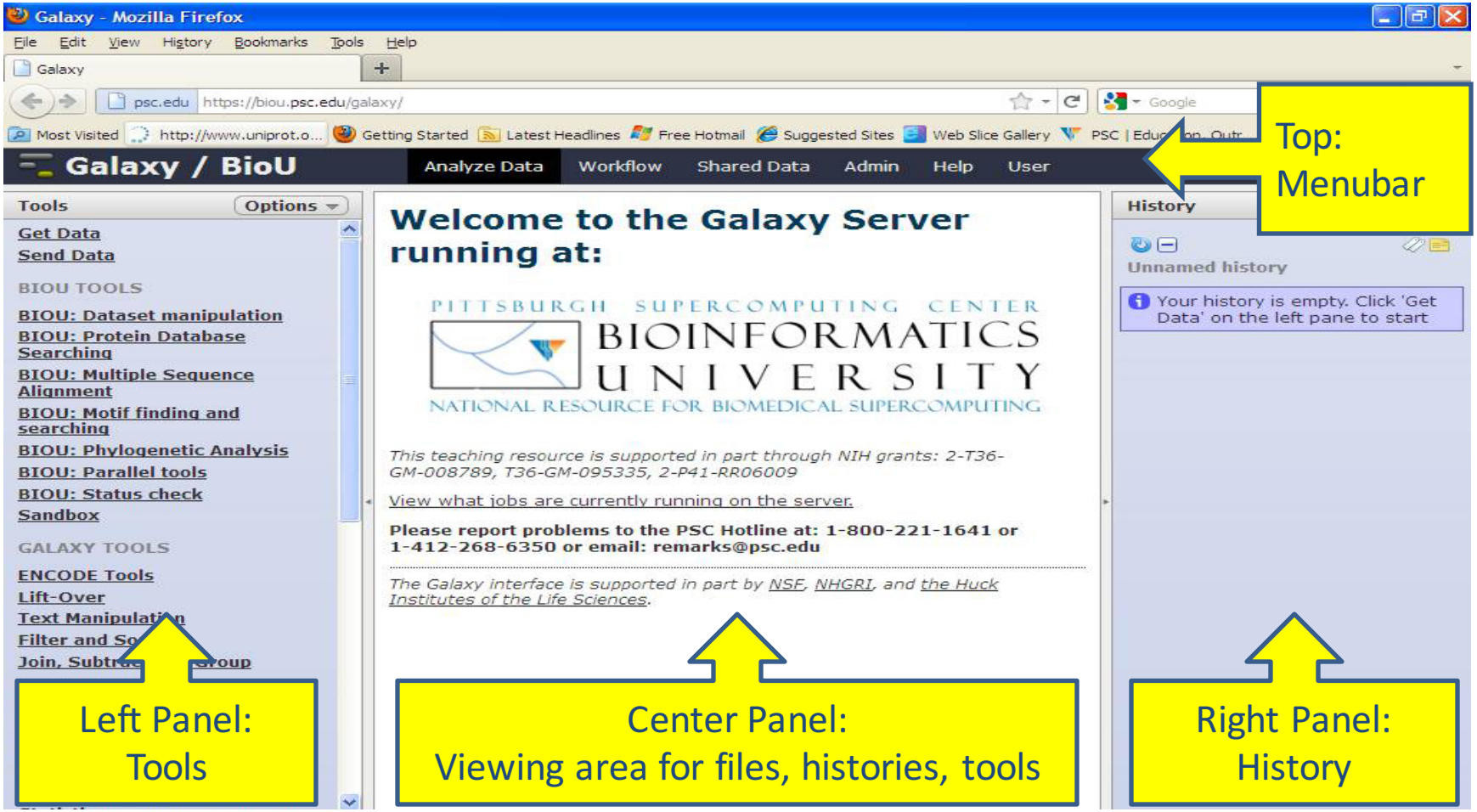
- Step 1: Open a web browser such as Firefox
- Step 2: Go to: https://biou.psc.edu/galaxy
- Step 3: Enter your PSC/Kerberos Username and Password when prompted by the web browser





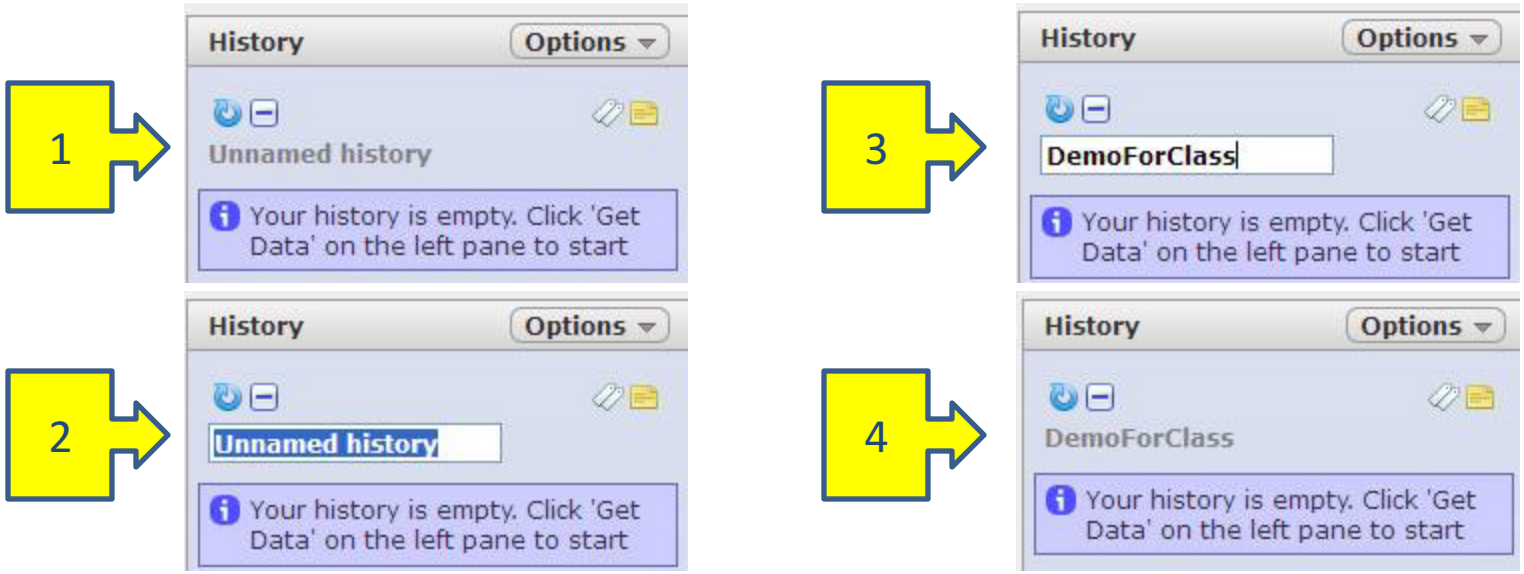
Successful Login View

Main Galaxy Window



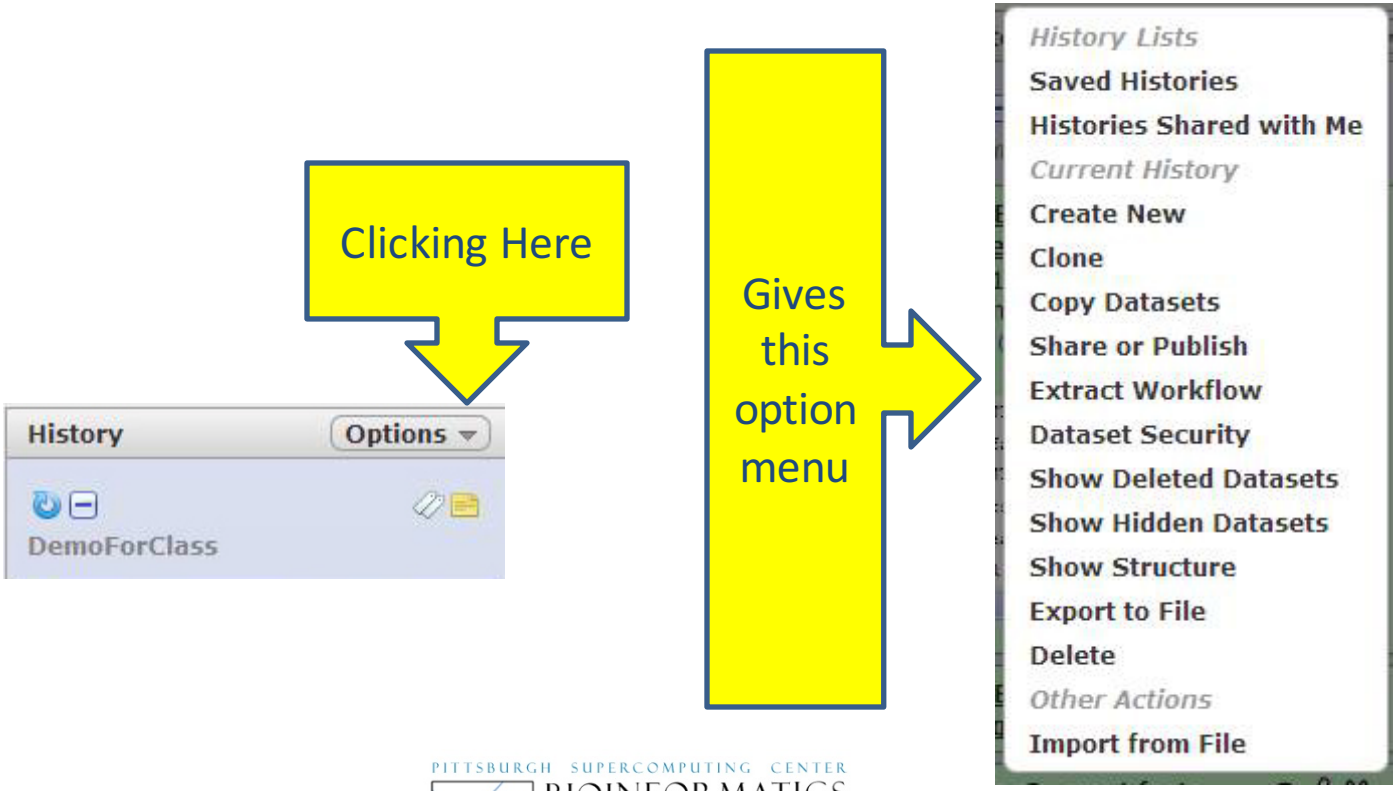
Galaxy History Names

- A Galaxy history is a recording (log) of what you did
- Use a descriptive name to clarify your history
- To change name: 1) click on old name; 2) delete old name; 3) enter new name; 4) hit ENTER key.



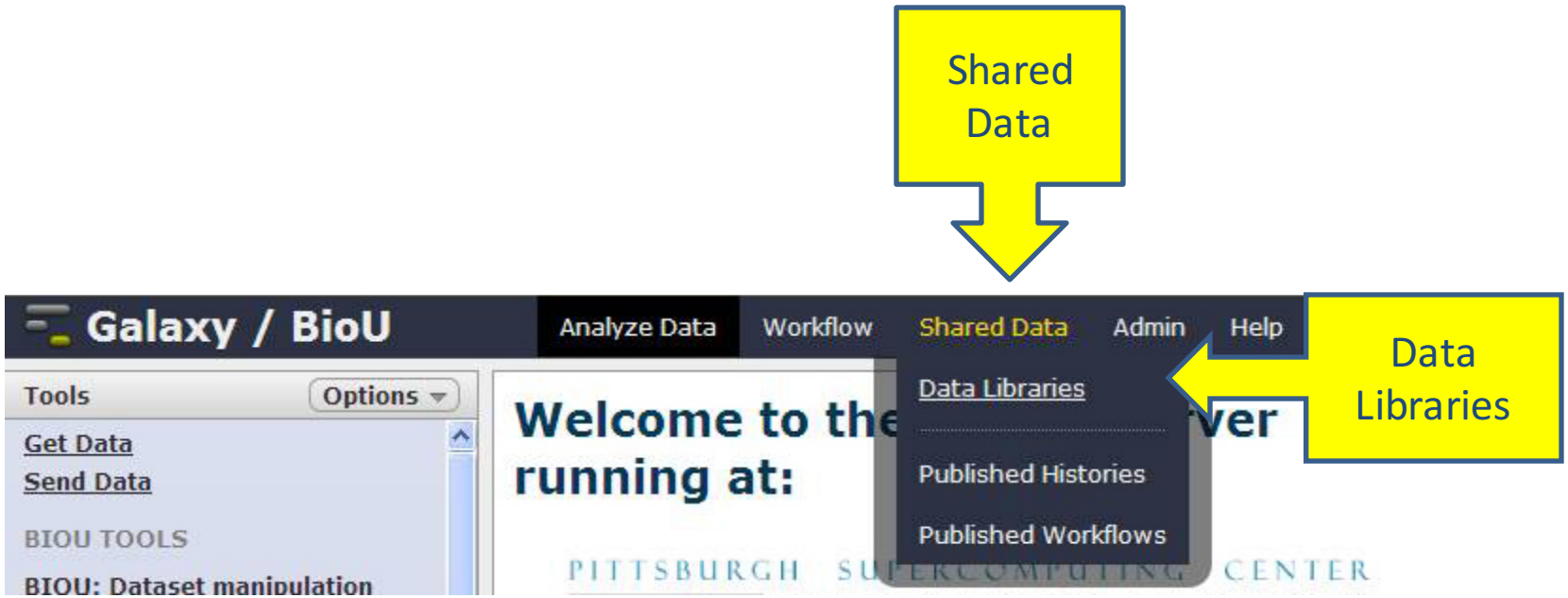
Galaxy History Options

- To view options, click on the triangle next to the Options menu on the history panel



Copying Shared Data to Galaxy History

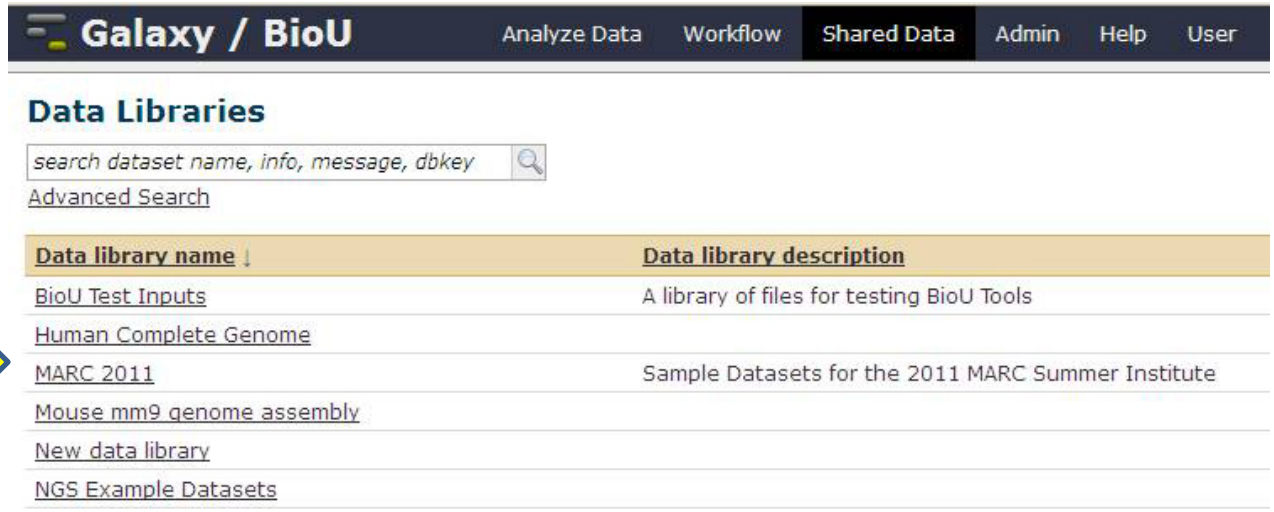
- Step 1: On top Galaxy Menubar choose Shared Data then select Data Libraries



Copying Shared Data to Galaxy History

- Step 2: Click on the specific Data Library name to select it

Click on a name to select



Copying Shared Data to Galaxy History

- Step 3: Click on triangle next to data file name.
- Step 4: Select Import this dataset into selected histories

Choose Import

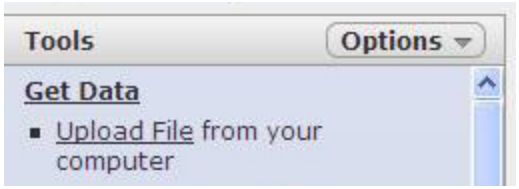
PITTSBURGH SUPERCOMPUTING CENTER
BIOINFORMATICS UNIVERSITY
NATIONAL RESOURCE FOR BIOMEDICAL SUPERCOMPUTING

Copying Shared Data to Galaxy History

- Step 5: Select the destination history.
- Step 6: Click on Import library datasets
- The shared dataset will now be in your history

Uploading Your Own Data

- Under Tools, click on Get Data then Upload File from your computer



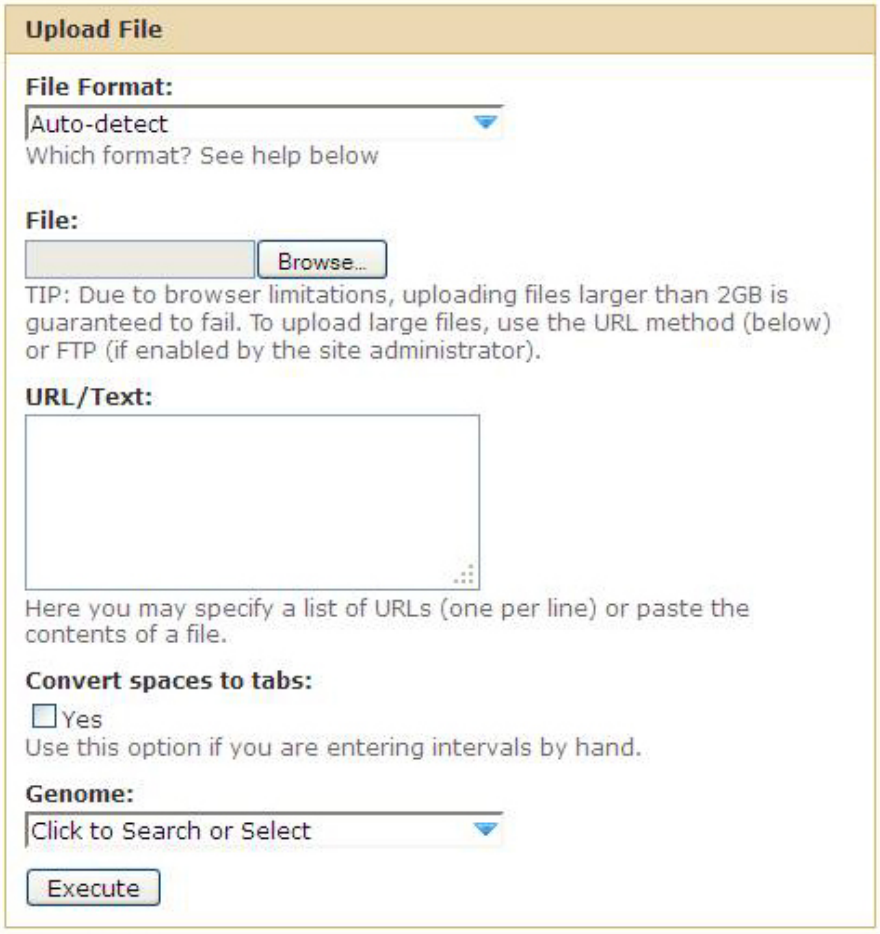
- A form will appear in the center panel to:
  - Select the file format of the data to be uploaded
  - Enter the file name
  - Upload the file

Uploading Your Own Data

Choose File Format

Choose File Name

Click Execute to Upload

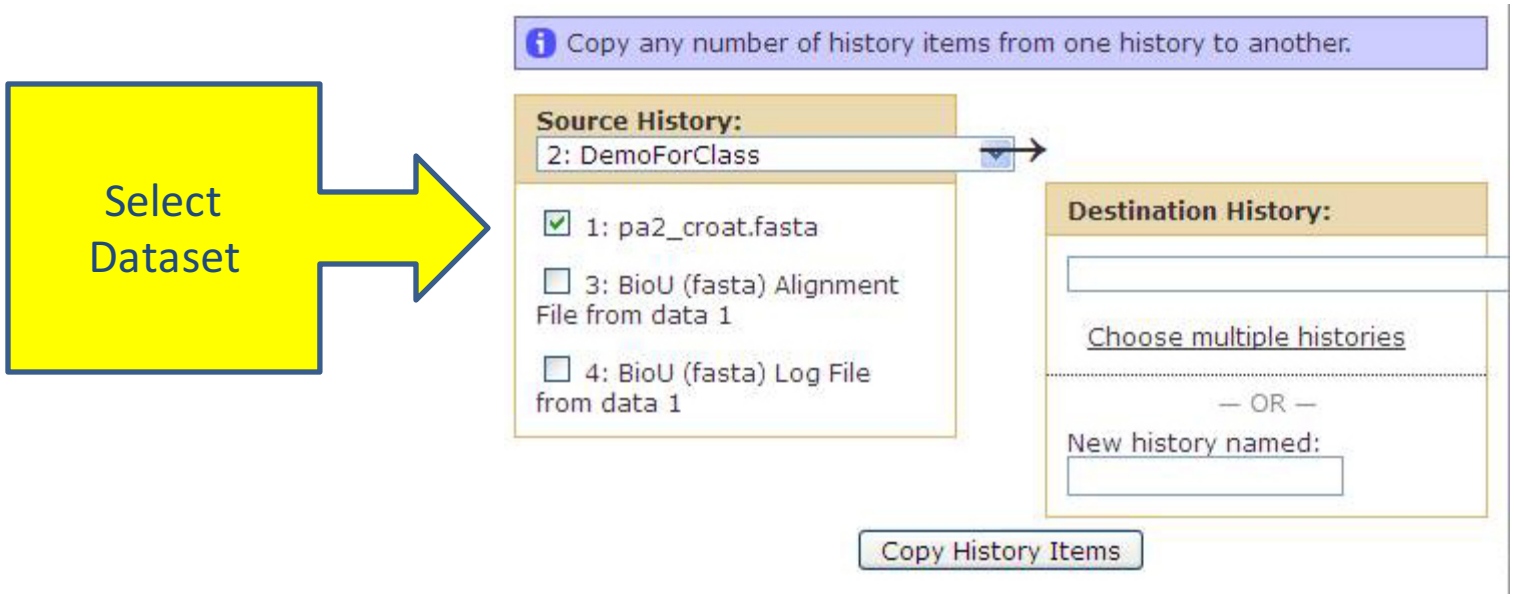


Copy Galaxy Datasets From History

- Step 1: Select history Options triangle, then select copy datasets from the submenu
- Step 2: Select the history that contains the dataset you will copy. (Click on triangle for list)

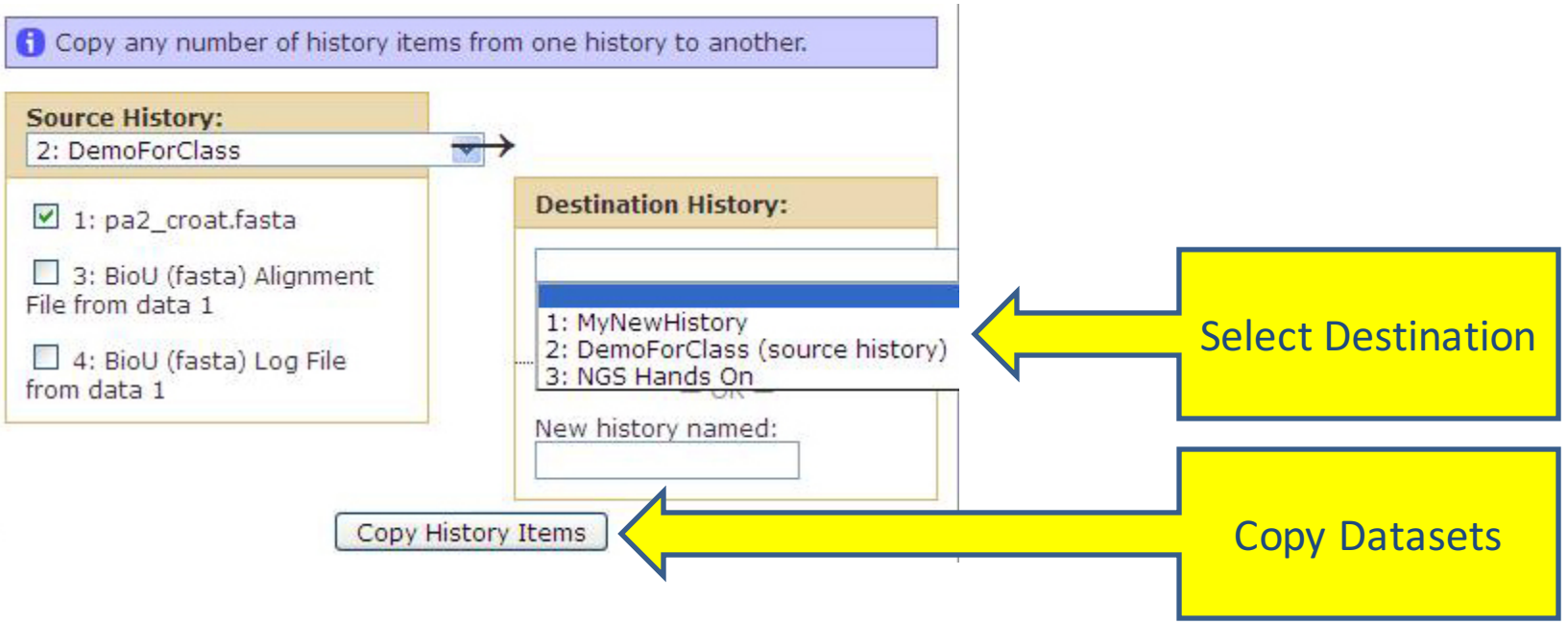
Copy Galaxy Datasets From History

- Step3: Select the dataset that you want to copy.



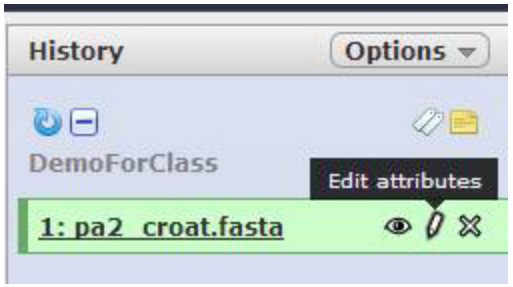
Copy Galaxy Datasets From History

- Step 4: Select the destination history.
- Step 5: Click on Copy History Items



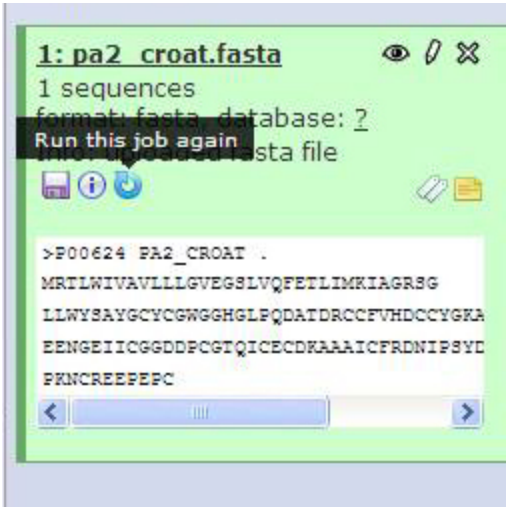
Symbols on Galaxy History Lists

- Eyeball: Display Data in Center panel
- Pen: Edit data attributes
- X: Delete



Symbols on Galaxy History Lists

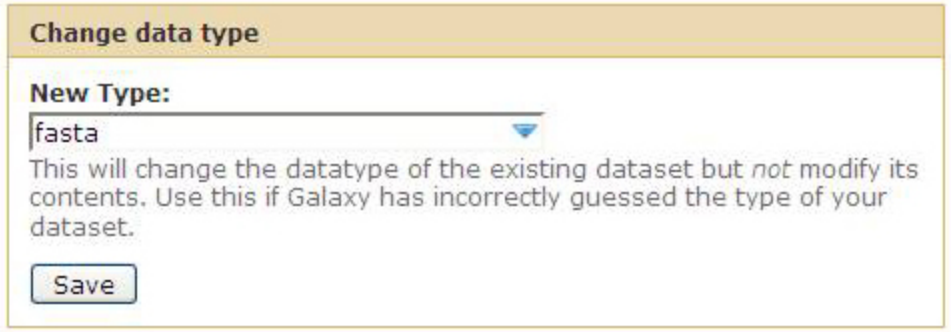
- Click on name for expanded view
- Disk: Download Data
- Circle Arrow: Rerun (or refresh)



PITTSBURGH SUPERCOMPUTING CENTER
BIOINFORMATICS UNIVERSITY
NATIONAL RESOURCE FOR BIOMEDICAL SUPERCOMPUTING

Changing Data Type Attribute

- Galaxy interface uses a data type attribute to sets the format of data inside Galaxy.
- Sometimes this value is set incorrectly, yet the data itself is in correct format for the tool.
- To correct, edit the attribute (click on pen) and change the type

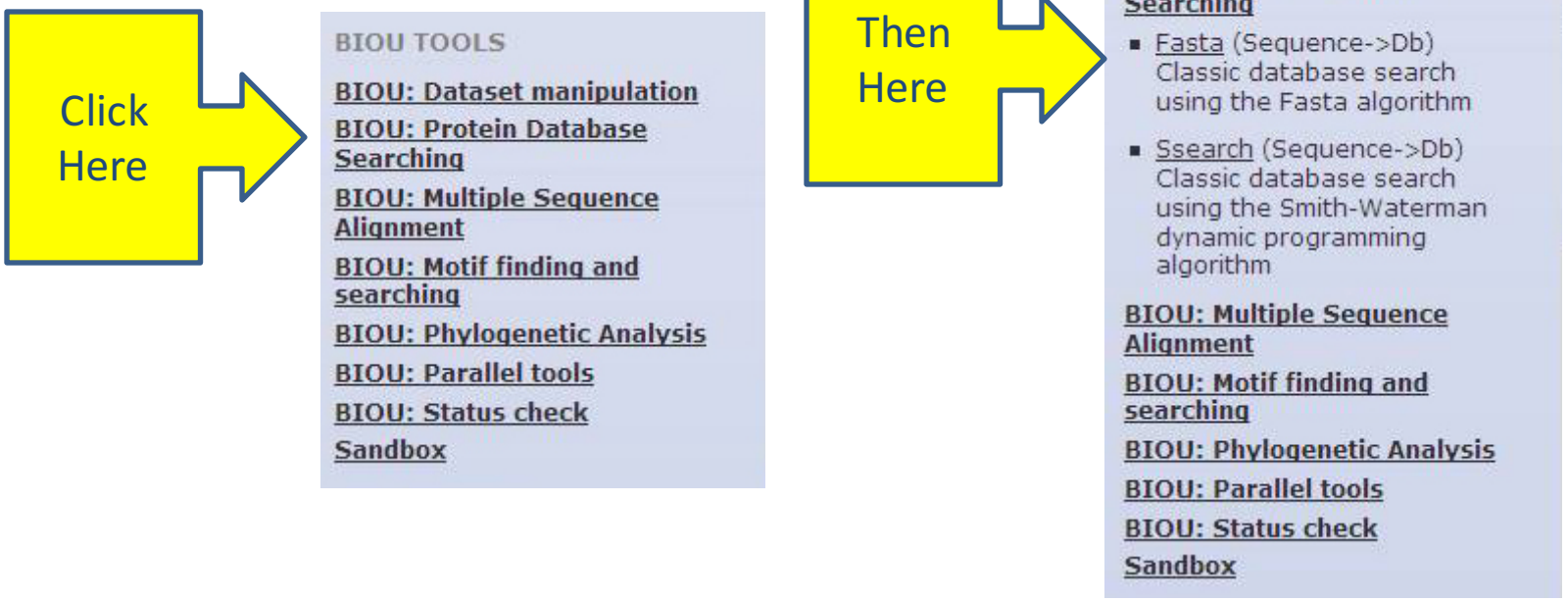


Running Galaxy Tools

- Example: Use Fasta tool with the sample dataset *pa2_croat.fasta* to search the *UniProt database (Swissprot Section)*
- Step 1: Copy the dataset pa2_croat.fasta from the Shared Data – Data Libraries - BioU Test Inputs library into your current history

Running Galaxy Tools

- Step 2: Under the Galaxy BIOU TOOLS section, locate and click on BIOU Protein Database Searching then Fasta

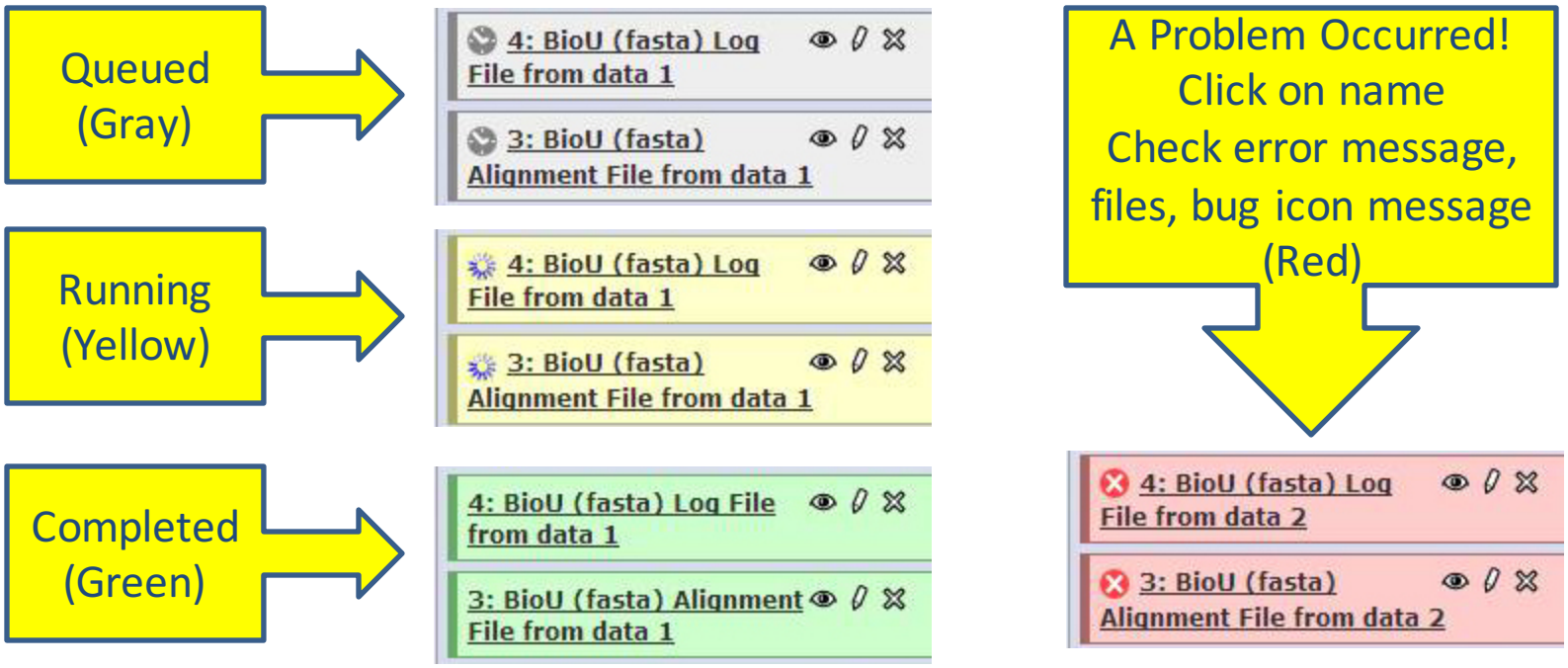


Running Galaxy Tools

- Step 3: Change any parameters; search with pa2_croat.fasta against UniProt (SwissProt Section)
- Step 4: Click on Execute

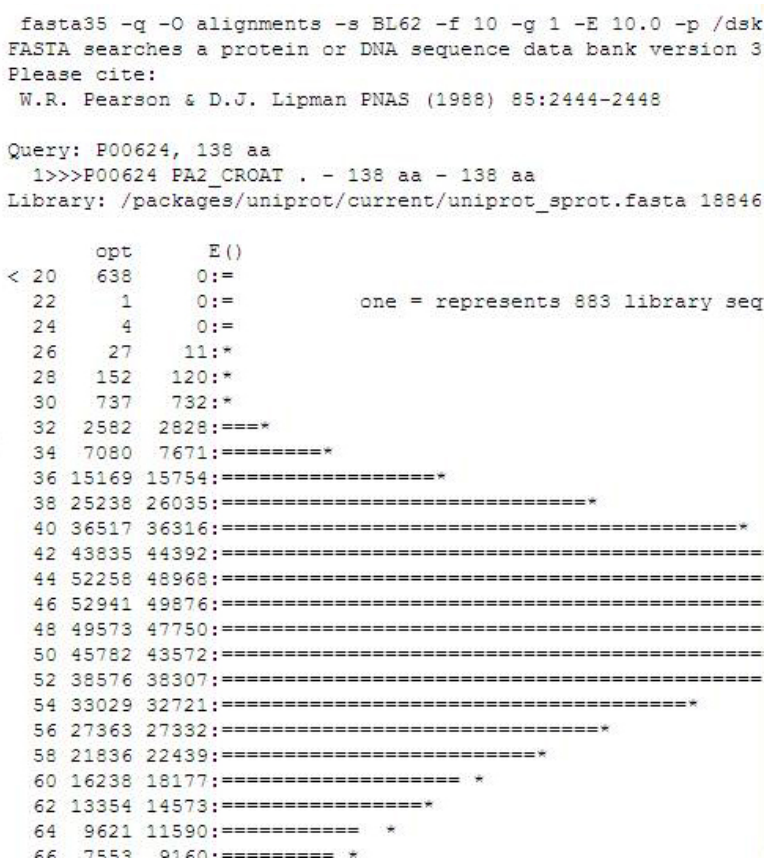
Running Galaxy Tools

- Step 5: Additional items will appear in history and will change colors according to status



Running Galaxy Tools

- Step 6: Click on Eyeballs to view the results



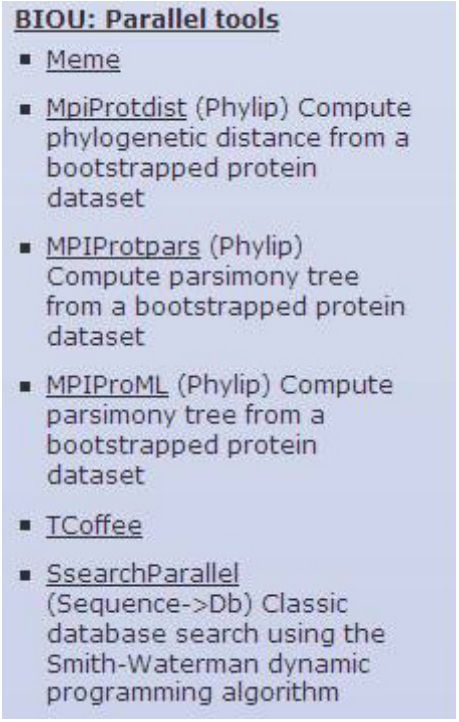
```
 fasta35 -q -O alignments -s BL62 -f 10 -g 1 -E 10.0 -p /dsk
FASTA searches a protein or DNA sequence data bank version 3
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query: P00624, 138 aa
  1>>>P00624 PA2_CROAT . - 138 aa - 138 aa
Library: /packages/uniprot/current/uniprot_sprot.fasta 18846

       opt      E()
< 20   638      0:=
  22     1      0:=          one = represents 883 library seq
  24     4      0:=
  26    27     11:*
  28   152    120:*
  30   737    732:*
  32  2582   2828:===*
  34  7080   7671:========*
  36 15169  15754:==================*
  38 25238  26035:==============================*
  40 36517  36316:==========================================*
  42 43835  44392:==================================================
  44 52258  48968:==================================================
  46 52941  49876:==================================================
  48 49573  47750:==================================================
  50 45782  43572:==================================================
  52 38576  38307:==============================================
  54 33029  32721:=======================================*
  56 27363  27332:================================*
  58 21836  22439:==========================*
  60 16238  18177:==================== *
  62 13354  14573:================*
  64  9621  11590:============  *
  66  7553   9160:========== *
```

Running Galaxy Tools

- Galaxy runs are currently limited to 48 hours (walltime).
- For long runs, there are parallel versions of some tools in Galaxy
- These parallel tools are listed under BIOU: Parallel Tools



Galaxy Requests

- Let us know if there are tools that have not been implemented in Galaxy that you need for your bioinformatics teaching
- Not all command-line options for tools are implemented in Galaxy; if there are options for tools that you really need please let us know!
- Galaxy runs are limited to 48 hours. If you have specific needs for your class that require longer runs of certain tools please let us know.