High Performance Computing in Life Sciences

Part I

HPC Introduction Introduction

PartII

BioComputing Sofware

Oleksandr Moskalenko
om@ufl.edu

Matt Gitzendanner
magitz@ufl.edu

UF Research Computing
Information Technology

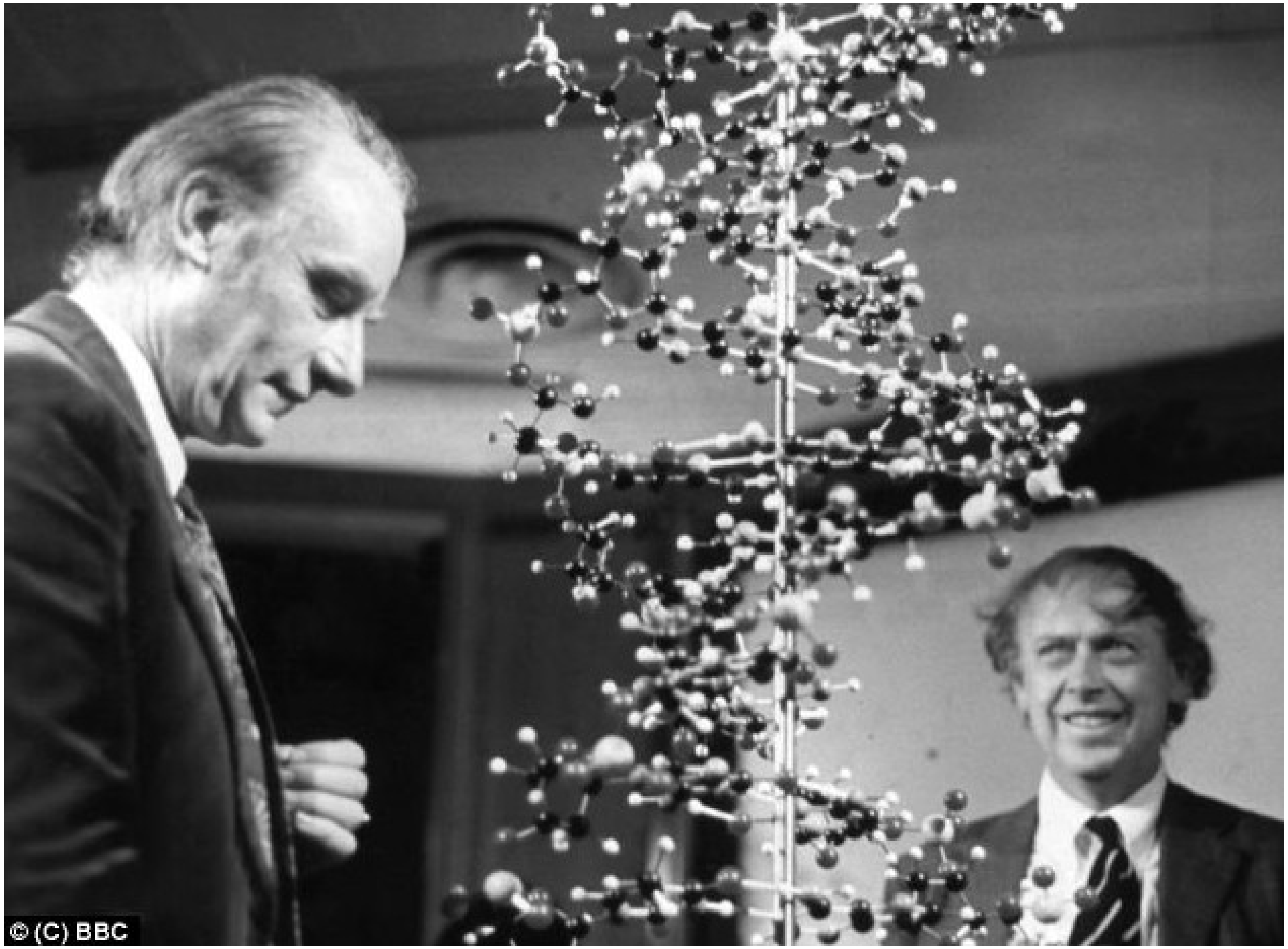
Summary

- The scale of biocomputing challenges
- The evolution of High-Performance Computing
- Current state of the traditional computing
- Parallelizing analyses
  - Traditional multiprocessing
  - Hadoop
  - Specialized approaches
- The interfaces
  - GUI vs. Web vs. Batch (comman-line)
- Biocomputing Software (Part II)

Historical Perspective

From a molecule to millions of genomes

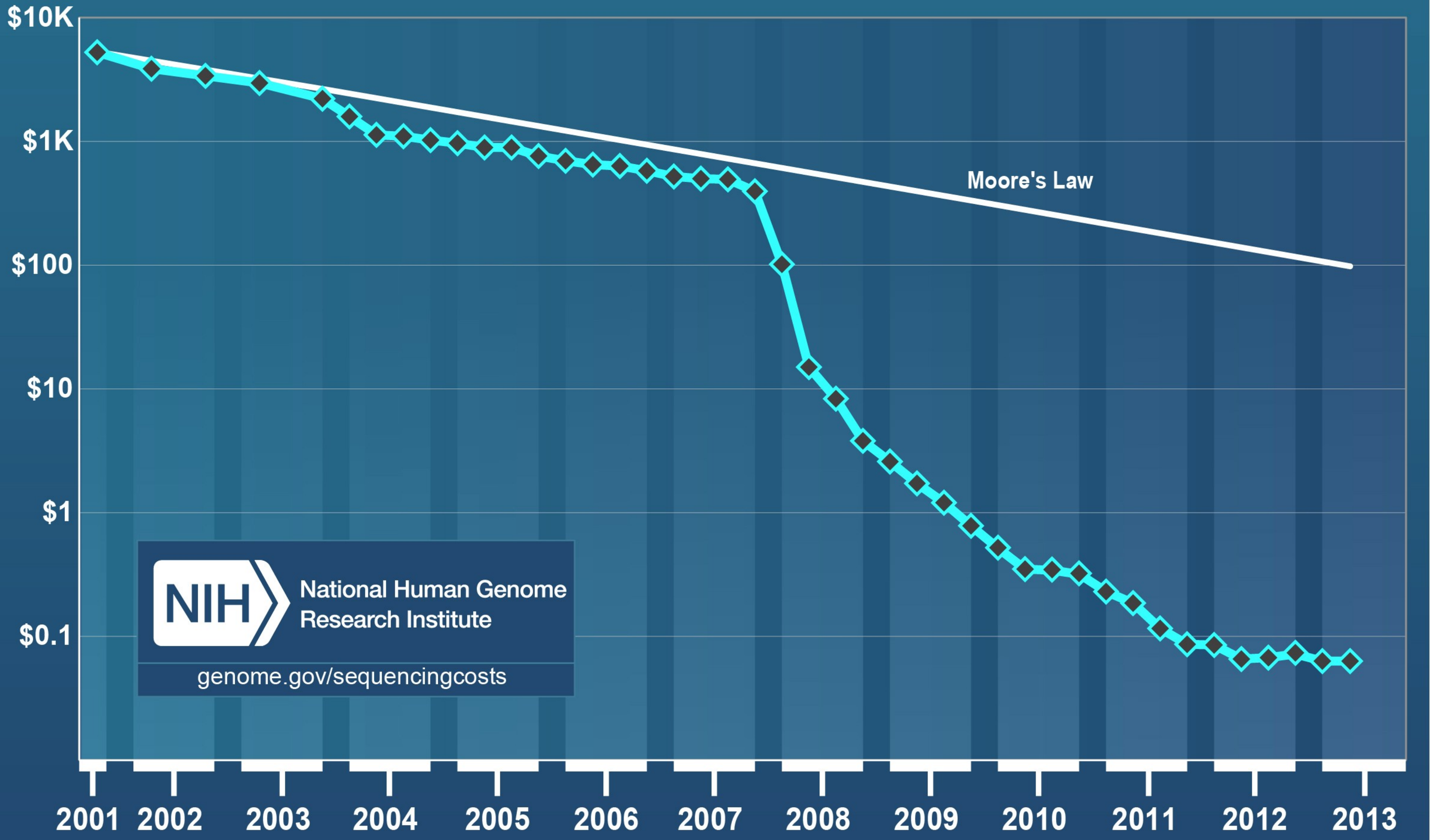
The Beginning



Sequencing Data Scaling

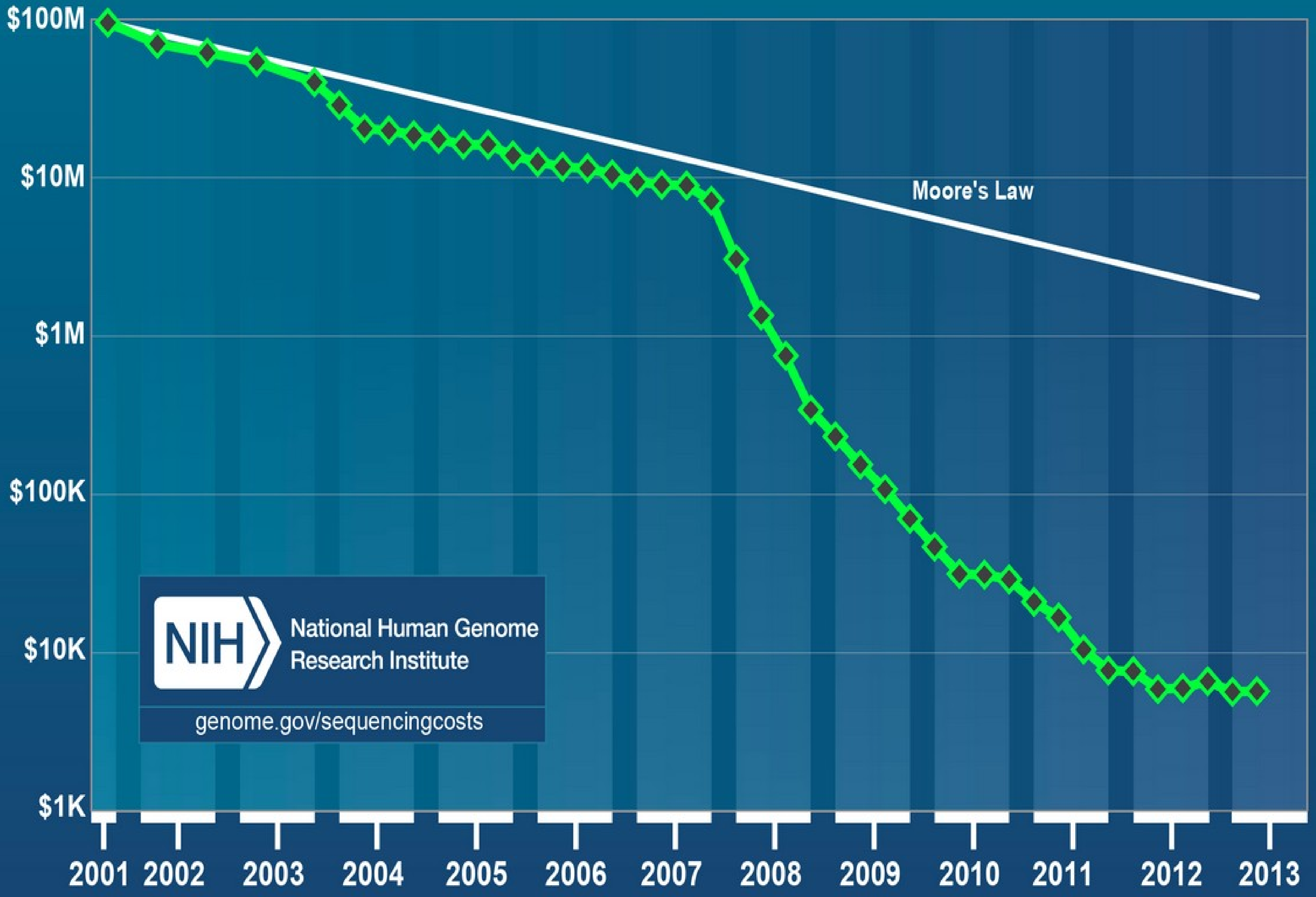
- Genome Size * Coverage
  - Viral – 1-100kbp
  - Bacteria, Archaea – 1-10Mbp
  - Simple Eukaryotes – 10-100 Mbp
  - Animals, Plants – 100Mbp - > 100Gbp
- Sequencing Coverage
  - ~10x in the Sanger Shotgun WGS times
  - ~30x for an average analysis
  - ~100x for metagenomic studies
  - Up to ~1000x for low-frequency SNP analysis in mixed samples

Cost per Raw Megabase of DNA Sequence

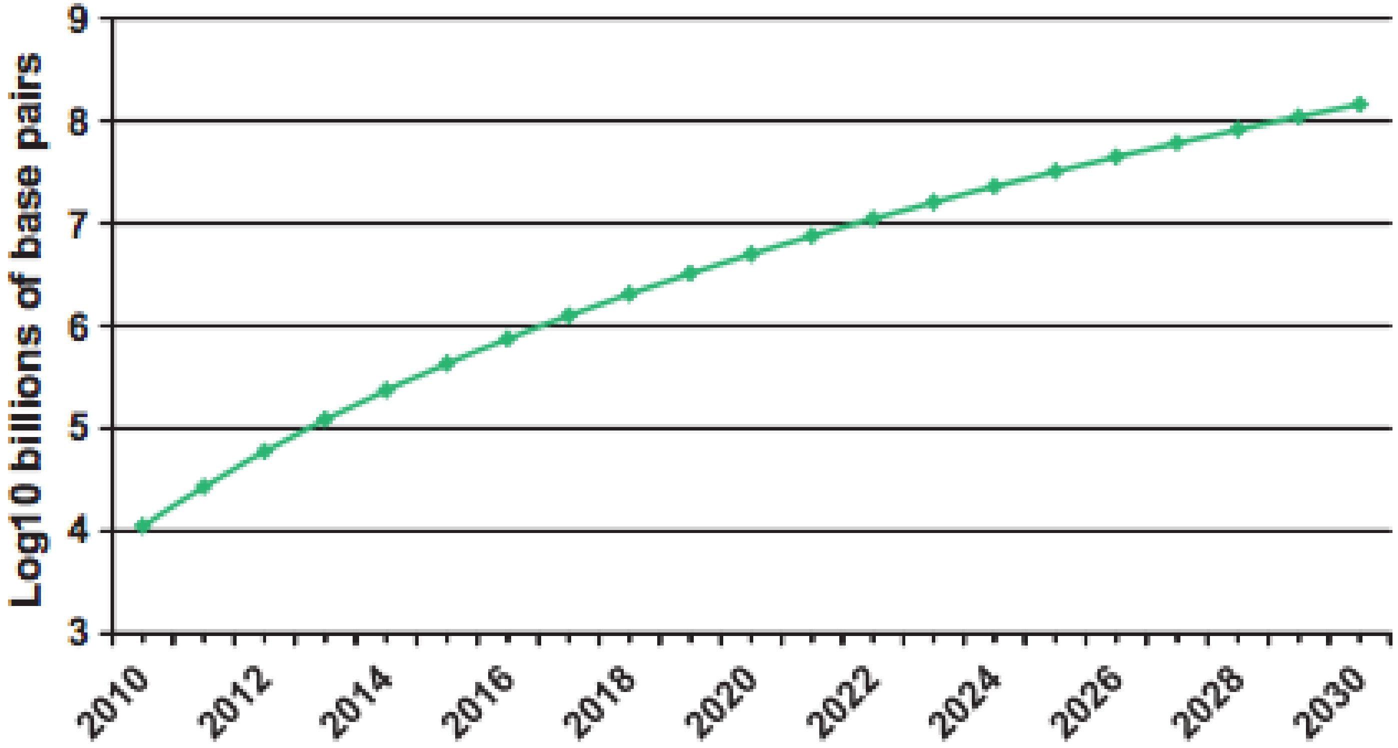


NIH National Human Genome Research Institute

genome.gov/sequencingcosts


$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts

Growth of Sequencing Data



106 (Mb) -> 109 (Gb) -> 1012 (Tb) -> 1015 (Pb) -> 1018 (Eb) -> 1021 (Zb)

Grossman et al. (2011)

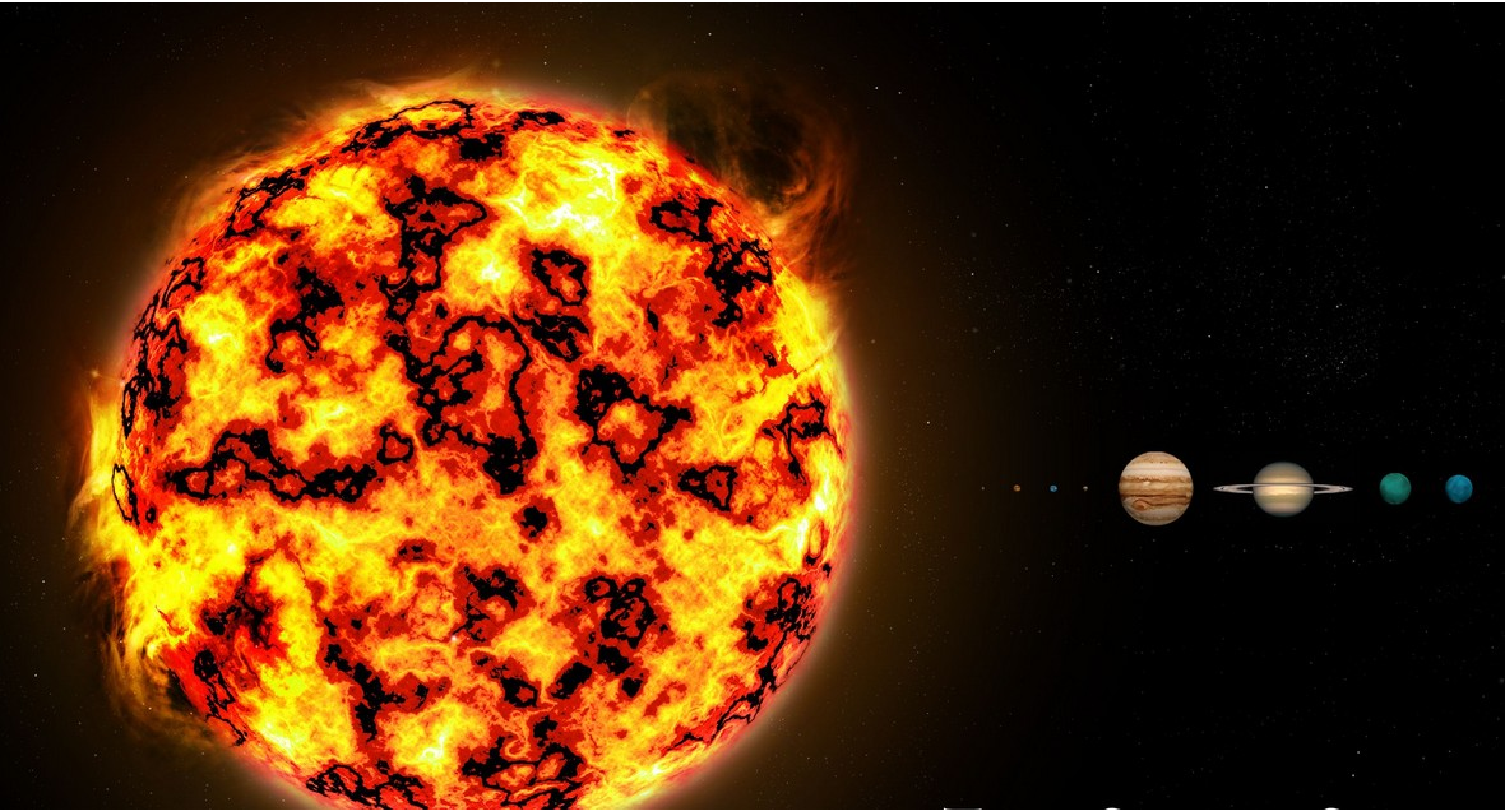
Growth of Sequencing Data

- 1 Gigabyte: A pickup truck filled with paper OR A symphony in high-fidelity sound OR A movie at TV quality
- 10 Terabytes: The printed collection of the US Library of Congress
- 2 Petabytes: All US academic research libraries
- 5 Exabytes: All words ever spoken by human beings.
- 2.7 Zettabytes: the total amount of global data in 2012 (IDC).

106 (Mb) -> 109 (Gb) -> 1012 (Tb) -> 1015 (Pb) -> 1018 (Eb) -> 1021 (Zb)

Grossman et al. (2011)

BioComputing Growth - NGS



Evolution of HPC

From Local to Global

“Local” BioComputing

Early Grid BioComputing

UNIVERSITY OF FLORIDA | High-Performance Computing

HiPerGator

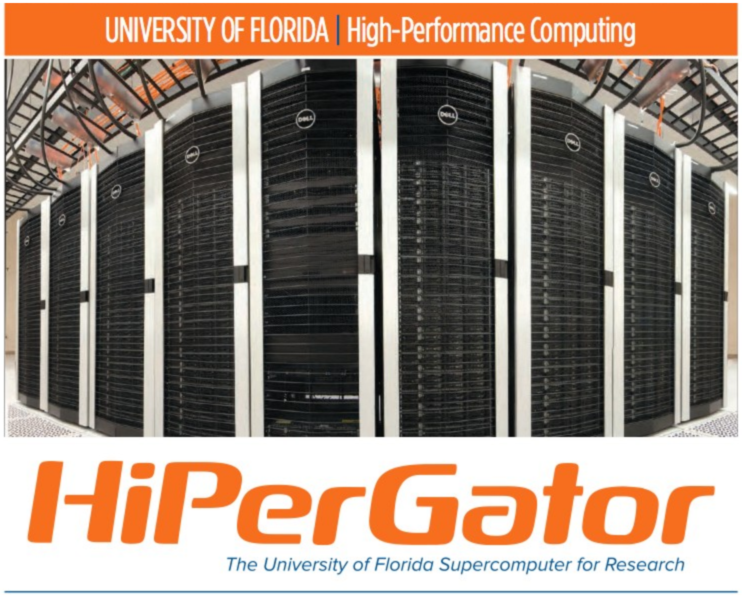
The University of Florida Supercomputer for Research

Contemporary Cluster Specs

- Storage and Networking:
  - 2Pb – Lustre parallel file system
  - 100Gbit networking, Infiniband Fabric
- Computing nodes:
  - 64 x 2.4GHz AMD Abu Dhabi cores
  - 254gb of usable memory
  - 1TB of local storage
- Big memory nodes:
  - 512Gb and 1TB memory with 48-80 cores
- GPU nodes:
  - Tesla, Fermi, Kepler GPU classes

HPC Considerations

- Scale


UNIVERSITY OF FLORIDA | High-Performance Computing
HiPerGator
The University of Florida Supercomputer for Research

HPC Considerations

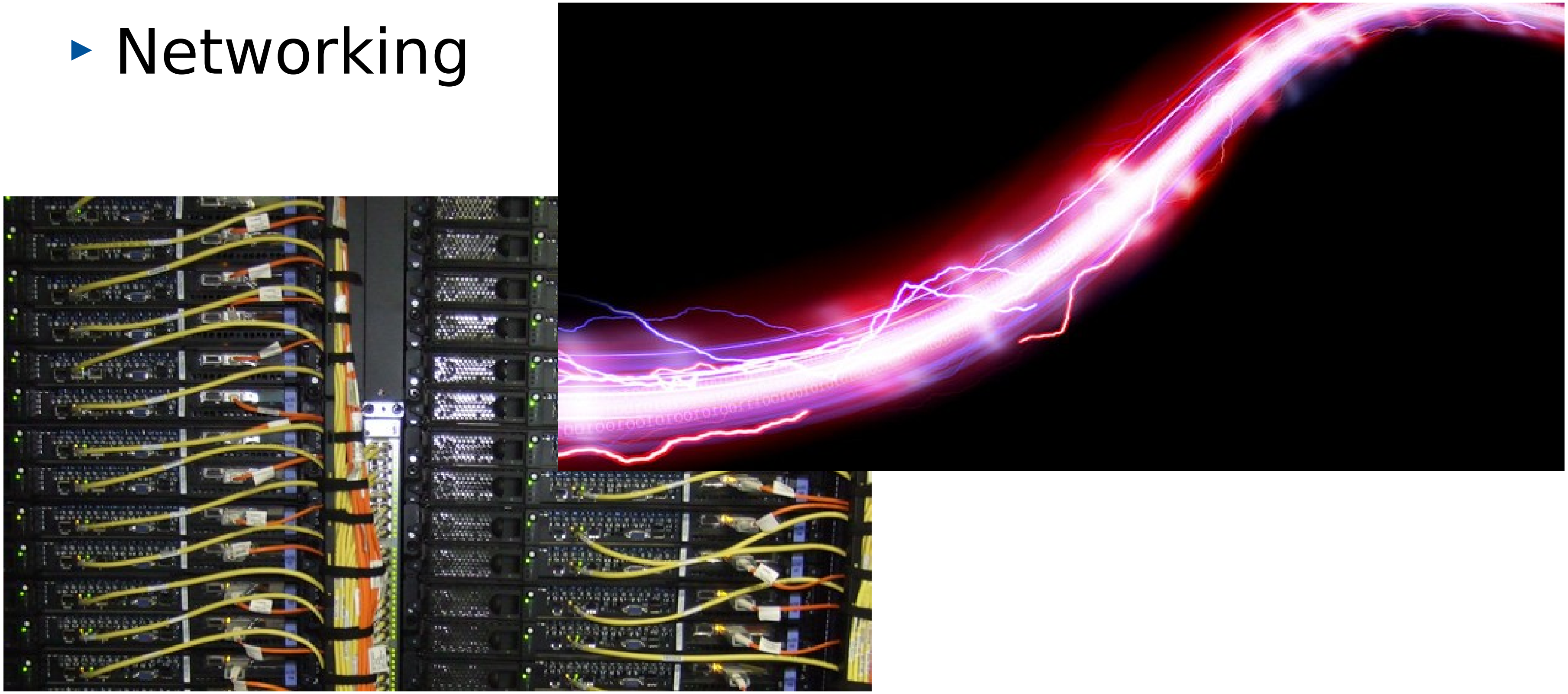
- Computational capacity vs. power and cooling

UF Data Center

- UF Data Center on Eastside Campus
  - 10,000 sq.ft and 1.75 MW total
  - 5,000 sq. ft. space for Research Computing

HPC Considerations

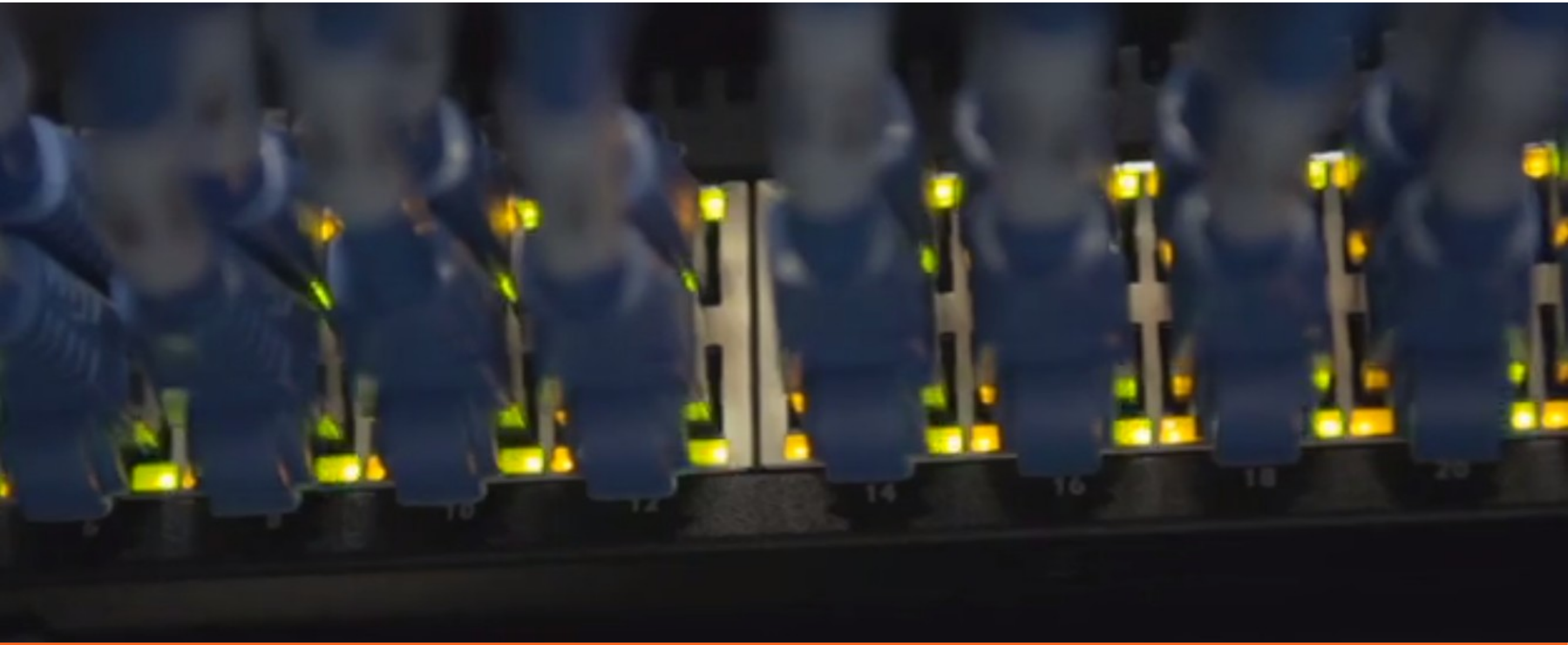
- Interconnects
- Networking



Internet2 Network

UF was 1st in the nation

- Internet2 Innovation Platform
  - 100 Gpbs connectivity
  - Campus Research Network now 200 Gbps



HPC Considerations

- Storage
- Parallel file systems
- High I/O storage
- Distributed storage

Scaling the HPC

The power of many

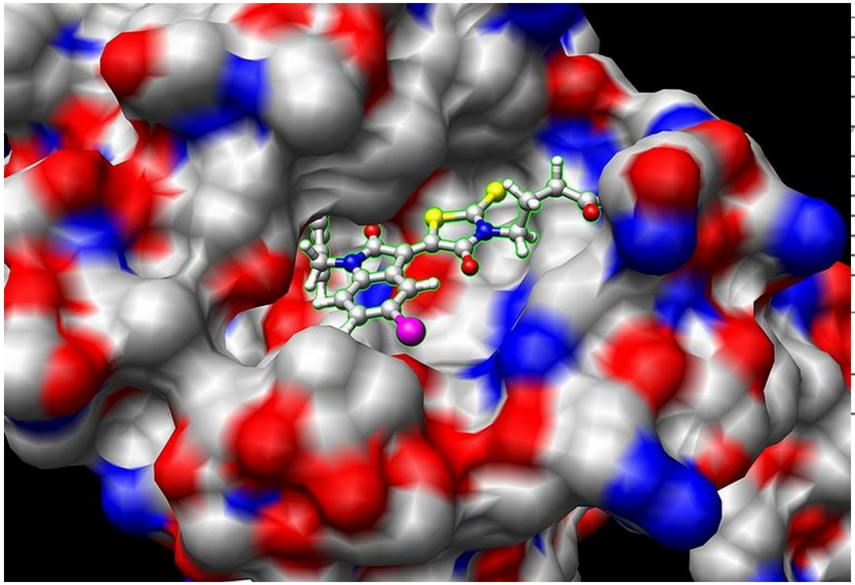
UNIVERSITY OF FLORIDA | High-Performance Computing

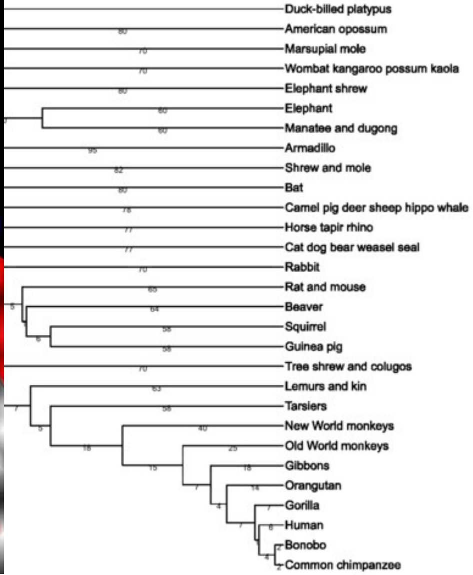
HiPerGator

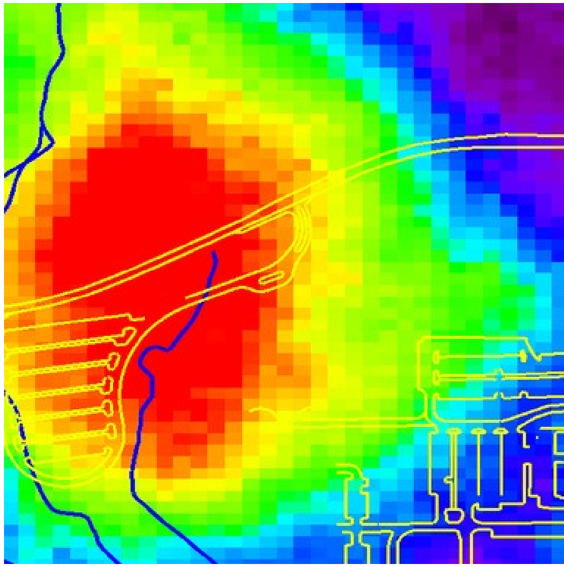
The University of Florida Supercomputer for Research

Computational Power

- Modeling, phylogenetics, simulations







Traditional Computation

- *De*-novo genome assembly
- Short-read mapping
- RNA-Seq
- BS-Seq
- CHIP-Seq
- SNP calling
- Pathway analysis
- …

- Why? Poor parallelization

Circumventing the Moore’s Law

Divide and conquer

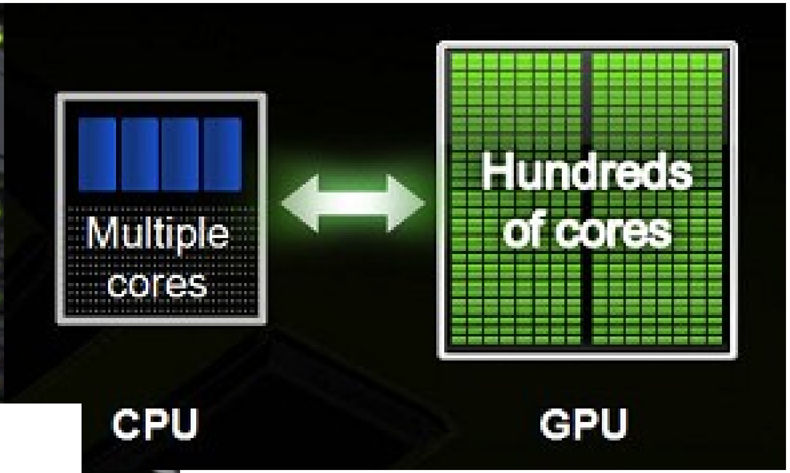
Traditional Parallel Computing

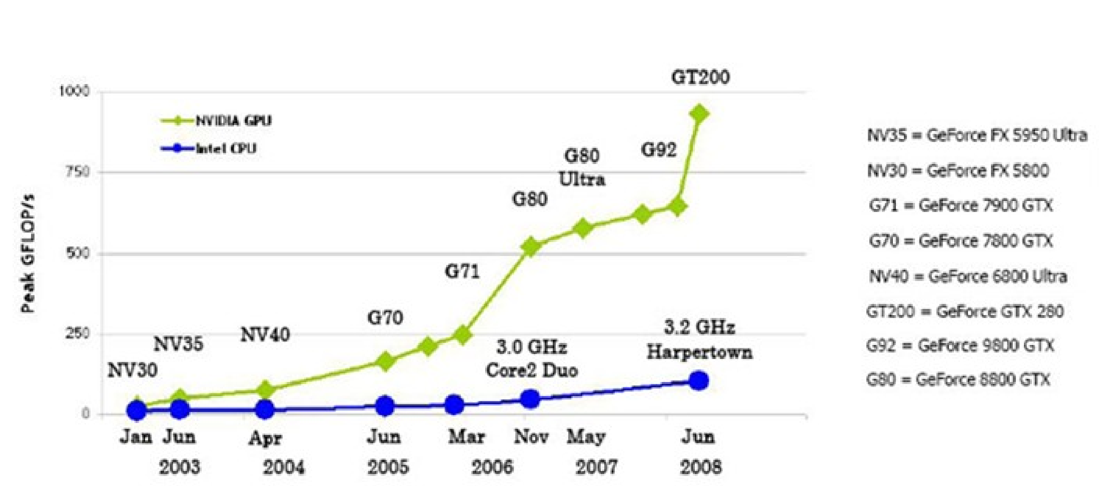
- Split analyses manually, run separately
- Multi-core (SMP) analyses with enabled software
- Multi-node (MPI) analyses with specially constructed software

GPU Computing

- Highly Parallelizable

Tesla™ GPU Parallel High Performance Computing (HPC) Solutions
Accelerate Your Performance - With Supermicro Massively Multithreaded.



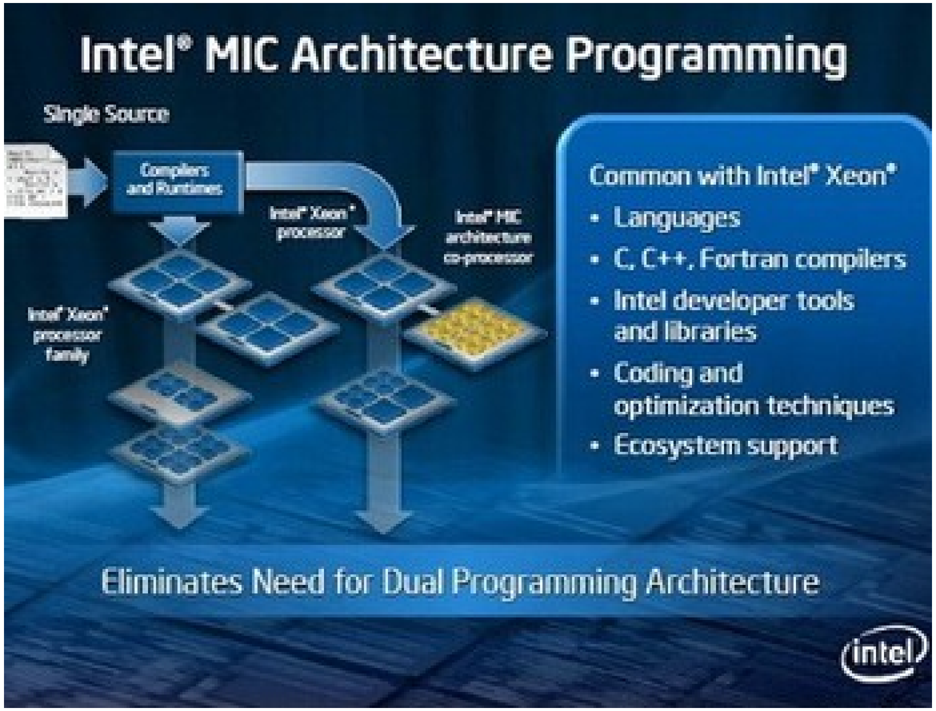


Need the code!

CUDA

MIC Computing

- Highly Parallelizable
- Standard x86 cores
- No need for learning a different programming paradigm ???



Intel® MIC Architecture Programming

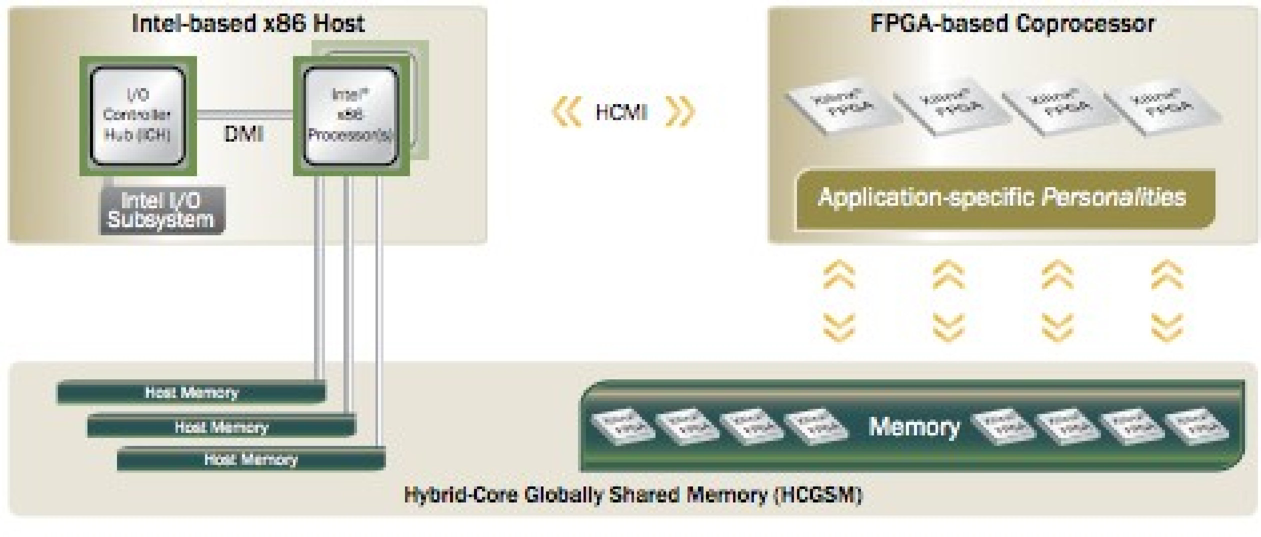
Single Source
Compilers and Runtimes
Intel® Xeon® processor
Intel® MIC architecture co-processor
Intel® Xeon® processor family

Common with Intel® Xeon®

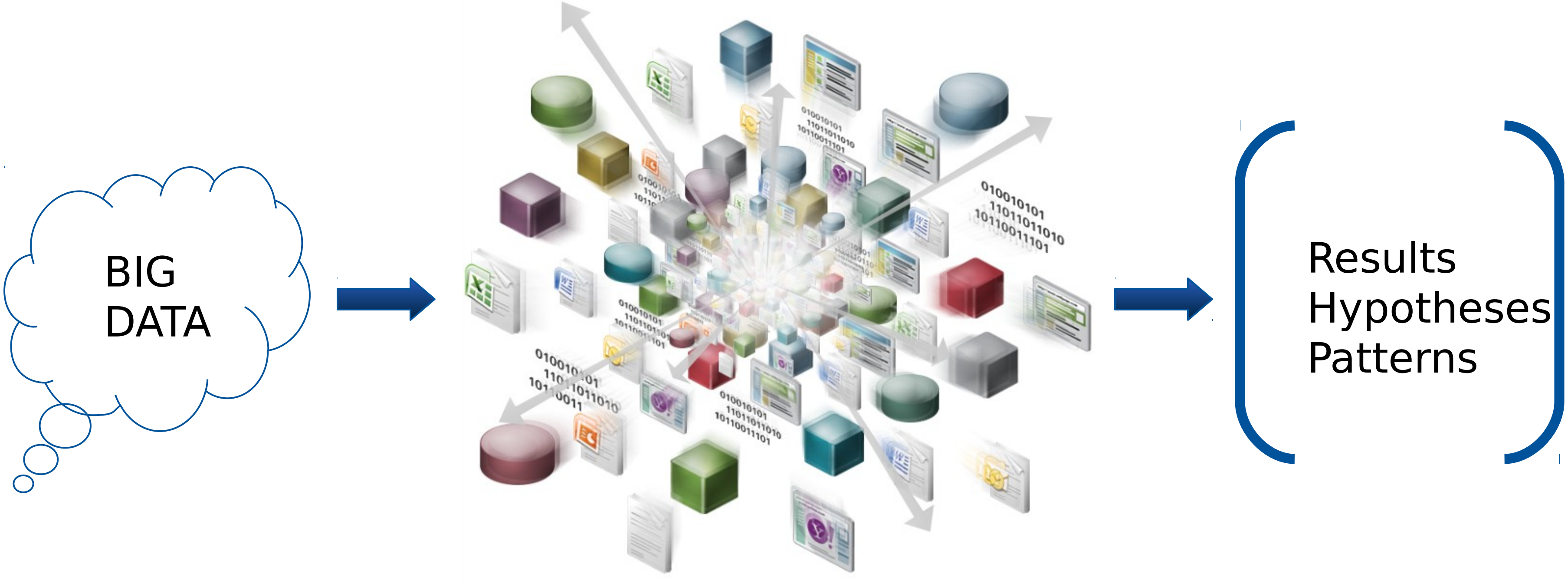
- Languages
- C, C++, Fortran compilers
- Intel developer tools and libraries
- Coding and optimization techniques
- Ecosystem support

Eliminates Need for Dual Programming Architecture

Specialized Processing



Distributed Computation (Hadoop)

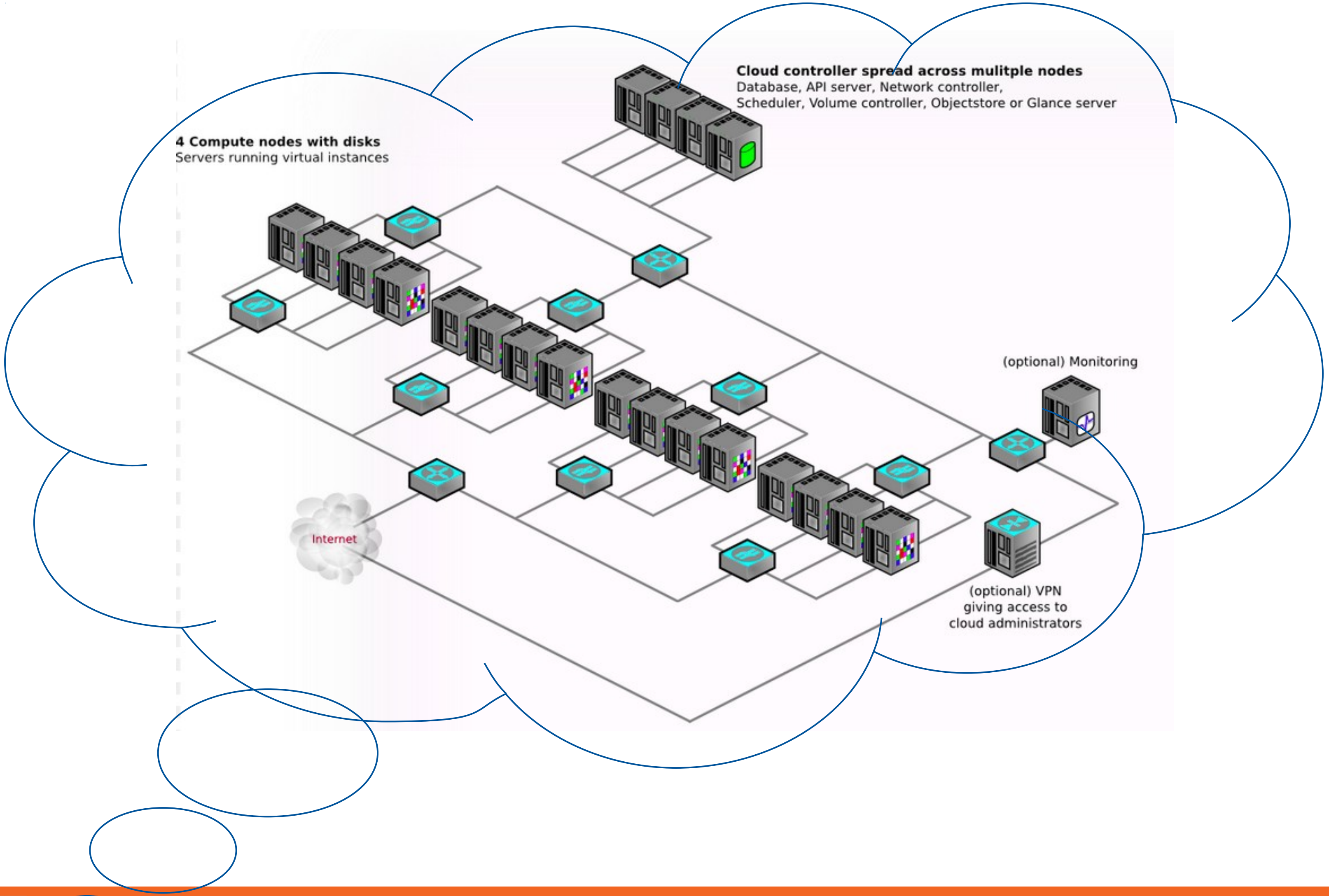


BIG DATA

Results
Hypotheses
Patterns

Map-Reduce Approach

Biocomputing Cloud 9 ???

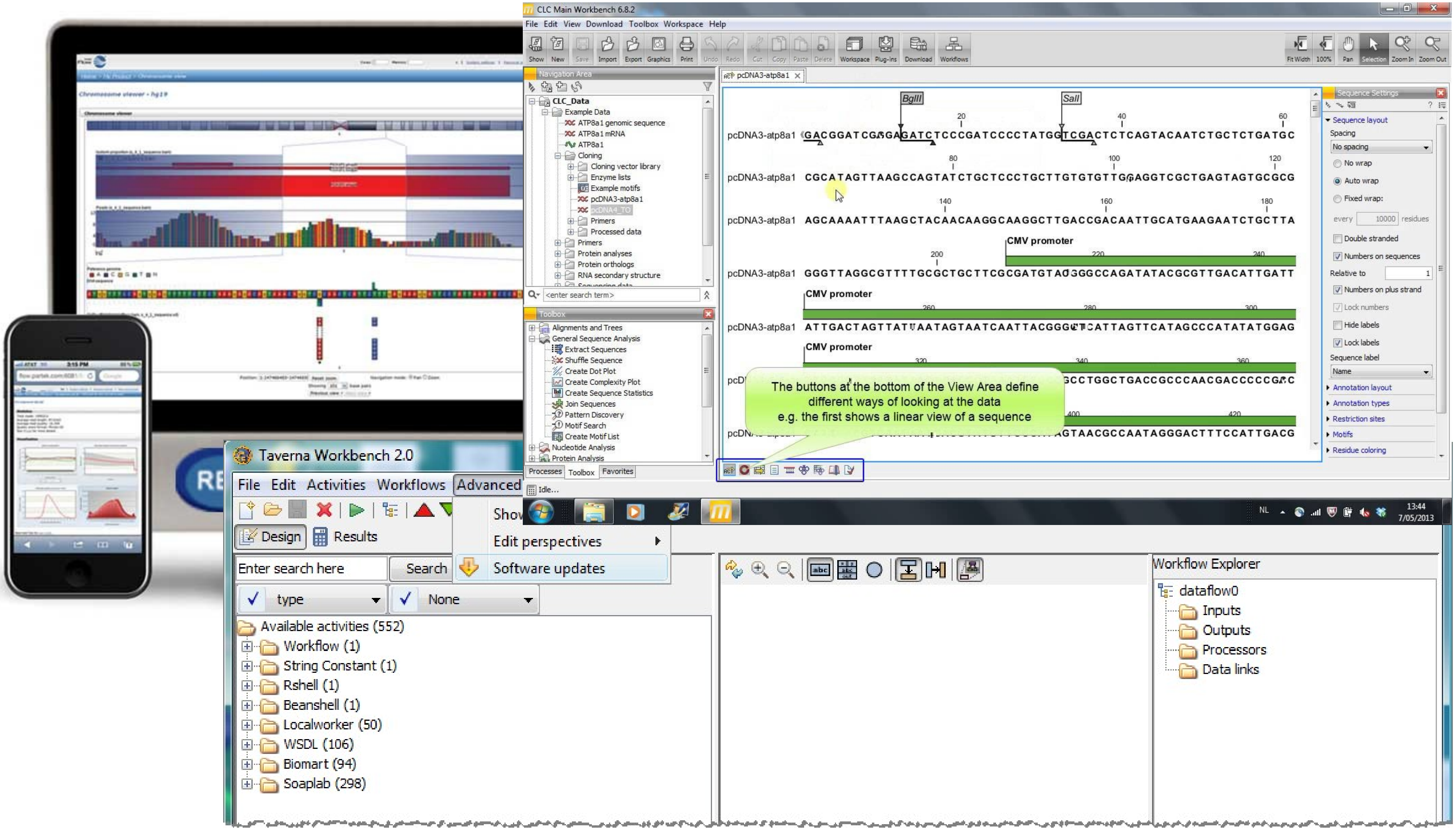

Cloud controller spread across mulitple nodes
Database, API server, Network controller,
Scheduler, Volume controller, Objectstore or Glance server
4 Compute nodes with disks
Servers running virtual instances
(optional) Monitoring
Internet
(optional) VPN
giving access to
cloud administrators

Interfaces

Interfaces, Interfaces, Interfaces!!!

What the Future May Bring

Graphical User Interfaces



The buttons at the bottom of the View Area define different ways of looking at the data e.g. the first shows a linear view of a sequence

Graphical User Interfaces

- Proprietary applications
  - Graphical User Interface
  - Integrate multiple tools, pipelines
  - User friendly-wizards for analyses
  - Many can tie into servers or clusters
  - Often highly optimized
  - Expensive
  - Limited flexibility
  - Limited scalability
  - Proprietary algorithms

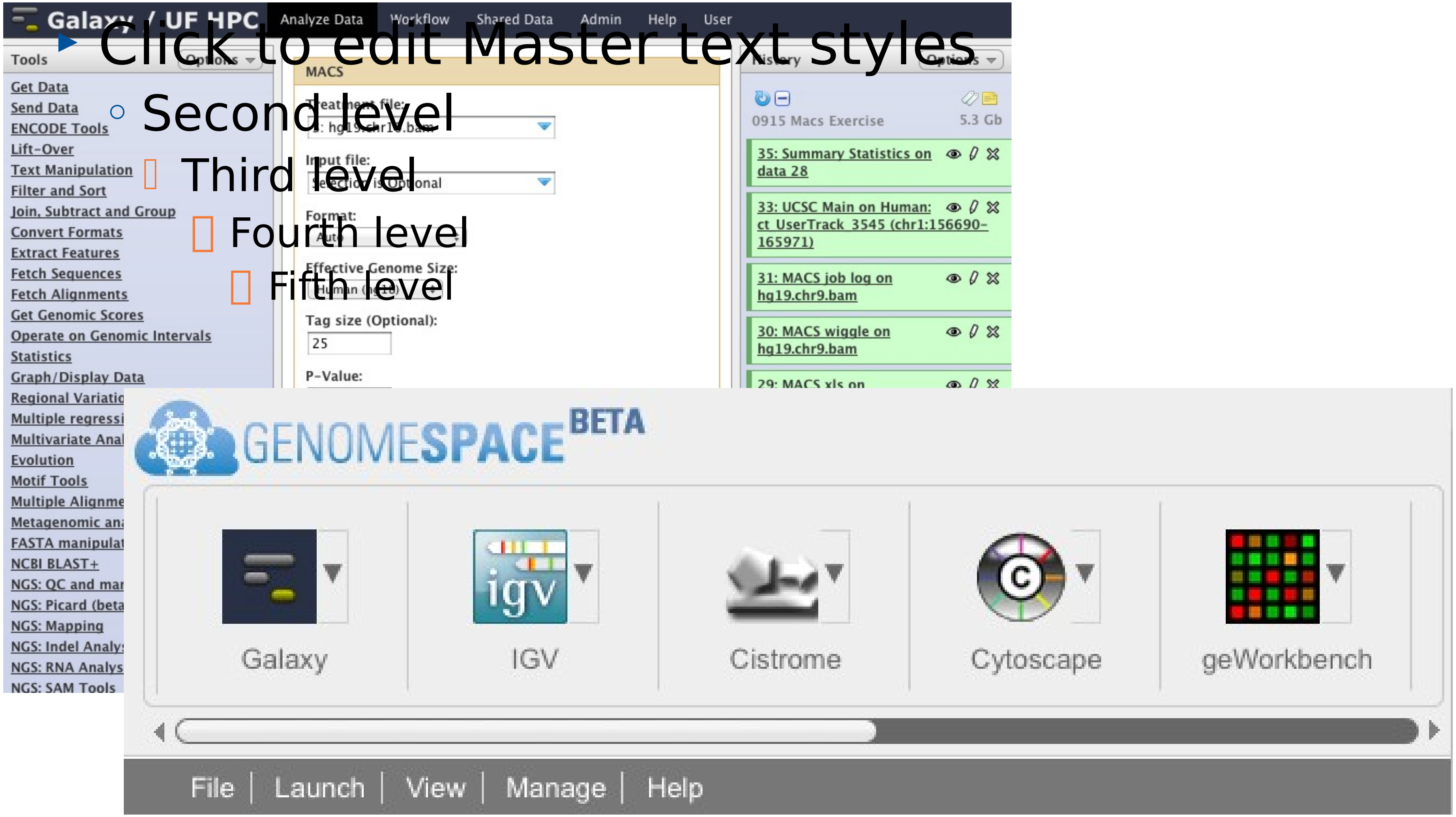
CLC bio
Accelerating Scientific Research

NextGENe®
2nd Generation Sequence Analysis Software

Partek®
Genomics Suite™

Web Interfaces

- Click to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level

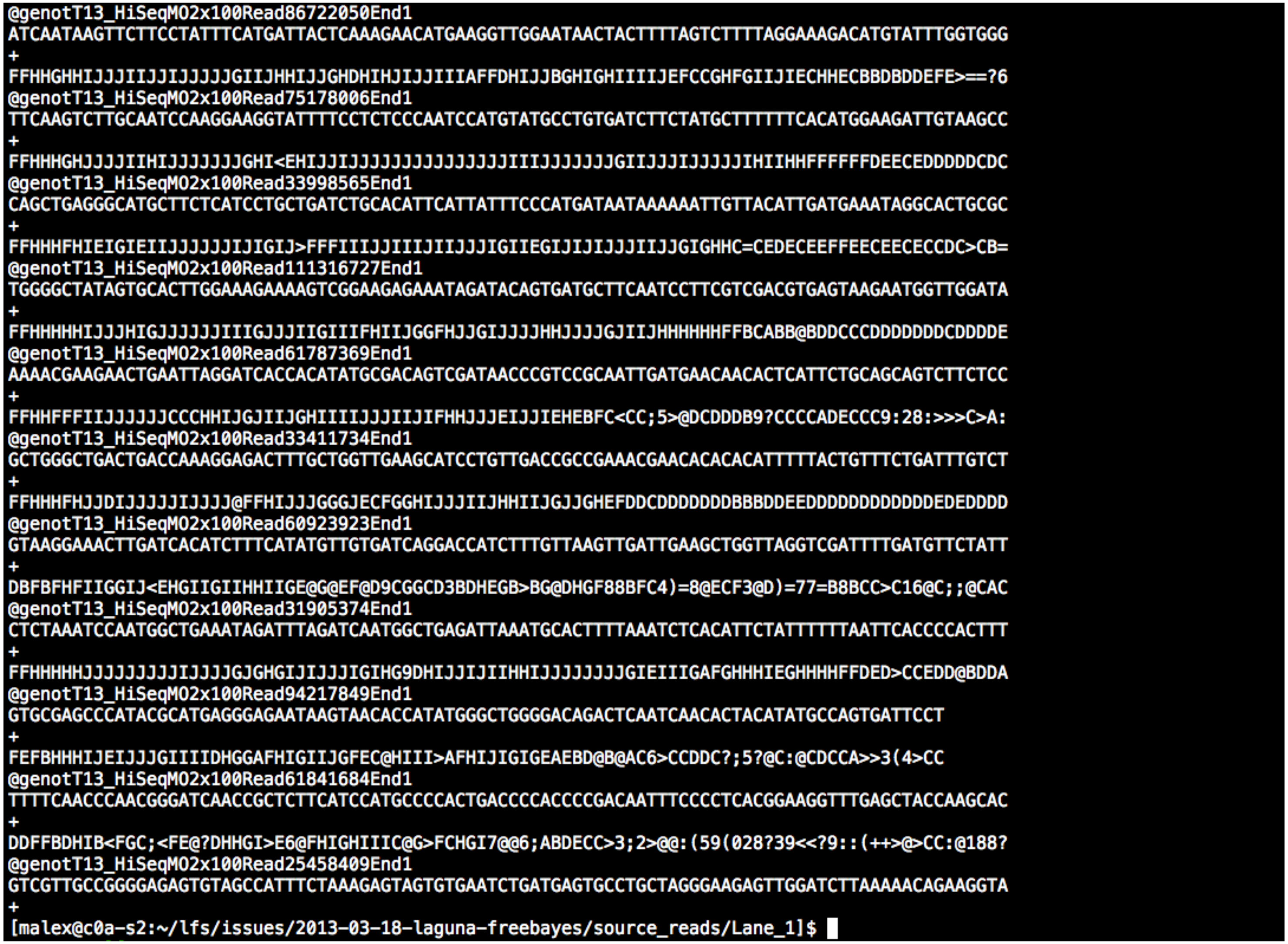


Web Interfaces

- Galaxy
  - Free, Open Source
  - Public or private instance, physical or cloud-based
  - Web interface
  - Most applications can be integrated
  - User made pipelines
  - Moderately scalable
  - Integrating applications time consuming
  - User made pipelines—where to start? reliability?

Galaxy

Batch Processing

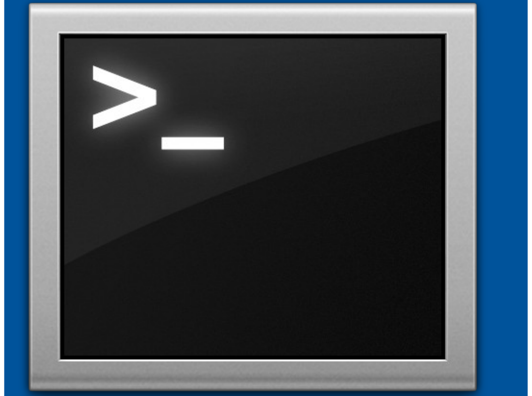


```
@genotT13_HiSeqM02x100Read86722050End1
ATCAATAAGTTCTTCCTATTTCATGATTACTCAAAGAACATGAAGGTTGGAATAACTACTTTTAGTCTTTTAGGAAAGACATGTATTTGGTGGG
+
FFHHGHHIJJJIIJJIJJJJJGIIJHHIJJGHDHIHJIJJIIIAFFDHIJJBGHIGHIIIIJEFCCGHFGIIJIECHHECBBDBDDEFE>==?6
@genotT13_HiSeqM02x100Read75178006End1
TTCAAGTCTTGCAATCCAAGGAAGGTATTTTCCTCTCCCAATCCATGTATGCCTGTGATCTTCTATGCTTTTTTCACATGGAAGATTGTAAGCC
+
FFHHHGHJJJJIIHIJJJJJJJJGHI<EHIJJIJJJJJJJJJJJJJJJIIIJJJJJJJGIIJJJIJJJJJIHIIHHFFFFFFDEECEDDDDDCDC
@genotT13_HiSeqM02x100Read33998565End1
CAGCTGAGGGCATGCTTCTCATCCTGCTGATCTGCACATTCATTATTTCCCATGATAATAAAAAATTGTTACATTGATGAAATAGGCACTGCGC
+
FFHHHFHIEIGIEIIJJJJJJIJIGIJ>FFFIIIJJIIIJIIJJJIGIIEGIJIJIJJJIIJJGIGHHC=CEDECEEFFEECEECECCDC>CB=
@genotT13_HiSeqM02x100Read111316727End1
TGGGGCTATAGTGCACTTGGAAAGAAAAGTCGGAAGAGAAATAGATACAGTGATGCTTCAATCCTTCGTCGACGTGAGTAAGAATGGTTGGATA
+
FFHHHHHIJJJHIGJJJJJJIIIGJJJIIGIIIFHIIJGGFHJJGIJJJJHHJJJJGJIIJHHHHHHFFBCABB@BDDCCCDDDDDDDCDDDDE
@genotT13_HiSeqM02x100Read61787369End1
AAAACGAAGAACTGAATTAGGATCACCACATATGCGACAGTCGATAACCCGTCCGCAATTGATGAACAACACTCATTCTGCAGCAGTCTTCTCC
+
FFHHFFFIIJJJJJJCCCHHIJGJIIJGHIIIIJJJIIJIFHHJJJEIJJIEHEBFC<CC;5>@DCDDDB9?CCCCADECCC9:28:>>>C>A:
@genotT13_HiSeqM02x100Read33411734End1
GCTGGGCTGACTGACCAAAGGAGACTTTGCTGGTTGAAGCATCCTGTTGACCGCCGAAACGAACACACACATTTTTACTGTTTCTGATTTGTCT
+
FFHHHFHJJDIJJJJJIJJJJ@FFHIJJJGGGJECFGGHIJJJIIJHHIIJGJJGHEFDDCDDDDDDDBBBDDEEDDDDDDDDDDDDEDEDDDD
@genotT13_HiSeqM02x100Read60923923End1
GTAAGGAAACTTGATCACATCTTTCATATGTTGTGATCAGGACCATCTTTGTTAAGTTGATTGAAGCTGGTTAGGTCGATTTTGATGTTCTATT
+
DBFBFHFIIGGIJ<EHGIIGIIHHIIGE@G@EF@D9CGGCD3BDHEGB>BG@DHGF88BFC4)=8@ECF3@D)=77=B8BCC>C16@C;;@CAC
@genotT13_HiSeqM02x100Read31905374End1
CTCTAAATCCAATGGCTGAAATAGATTTAGATCAATGGCTGAGATTAAATGCACTTTTAAATCTCACATTCTATTTTTTAATTCACCCCACTTT
+
FFHHHHHJJJJJJJJJIJJJJGJGHGIJIJJJIGIHG9DHIJJIJIIHHIJJJJJJJJGIEIIIGAFGHHHIEGHHHHFFDED>CCEDD@BDDA
@genotT13_HiSeqM02x100Read94217849End1
GTGCGAGCCCATACGCATGAGGGAGAATAAGTAACACCATATGGGCTGGGGACAGACTCAATCAACACTACATATGCCAGTGATTCCT
+
FEFBHHHIJEIJJJGIIIIDHGGAFHIGIIJGFEC@HIII>AFHIJIGIGEAEBD@B@AC6>CCDDC?;5?@C:@CDCCA>>3(4>CC
@genotT13_HiSeqM02x100Read61841684End1
TTTTCAACCCAACGGGATCAACCGCTCTTCATCCATGCCCCACTGACCCCACCCCGACAATTTCCCCTCACGGAAGGTTTGAGCTACCAAGCAC
+
DDFFBDHIB<FGC;<FE@?DHHGI>E6@FHIGHIIIC@G>FCHGI7@@6;ABDECC>3;2>@@:(59(028?39<<?9::(++>@>CC:@188?
@genotT13_HiSeqM02x100Read25458409End1
GTCGTTGCCGGGGAGAGTGTAGCCATTTCTAAAGAGTAGTGTGAATCTGATGAGTGCCTGCTAGGGAAGAGTTGGATCTTAAAAACAGAAGGTA
+
[malex@c0a-s2:~/lfs/issues/2013-03-18-laguna-freebayes/source_reads/Lane_1]$ 
```

Batch Processing

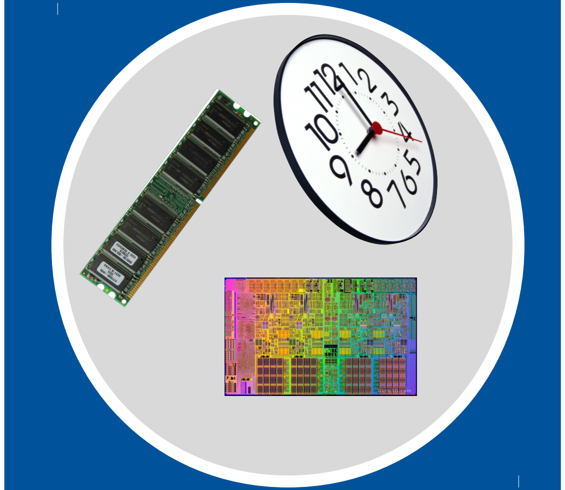
User interaction

Galaxy



Login node (Head node)

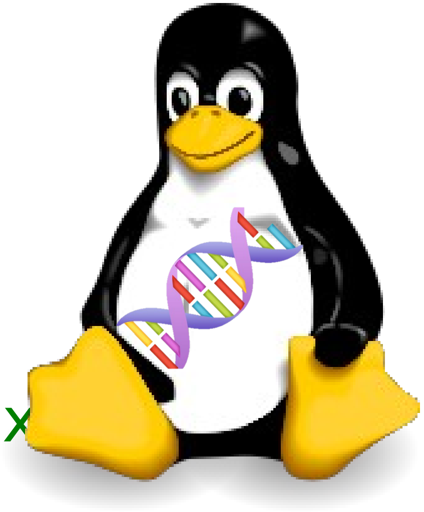
Scheduler

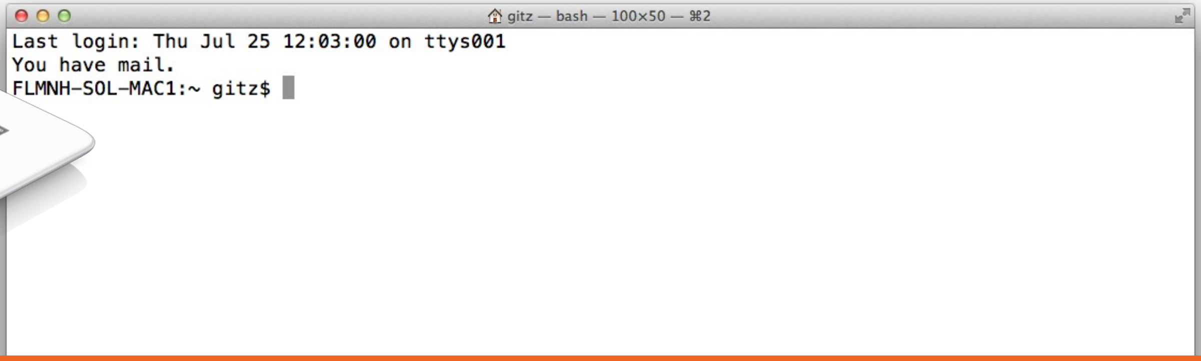


Tell the scheduler what you want to do

Batch Processing

- The Linux Command Line
  - Maximum flexibility
  - Most informatics tools run under Linux
  - Write your own tool, or script
  - Maximum scalability
  - Learning barrier of entry





```
Last login: Thu Jul 25 12:03:00 on ttys001
You have mail.
FLMNH-SOL-MAC1:~ gitz$ 
```

Batch processing

- Submission Script

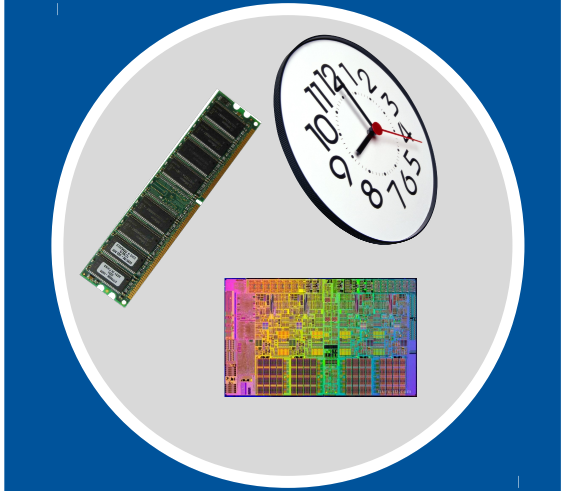
```
#!/bin/bash
#PBS -N My_Job_Name
#PBS -M Joe_Shmoe@ufl.edu
#PBS -m abe
#PBS -o My_Job.log
#PBS –e My_Job.err
#PBS -l nodes=1:ppn=1
#PBS -l walltime=00:05:00
#PBS –l pmem=900mb

cd $PBS_O_WORKDIR
date
module load test_app
test_app –i file.txt
```

Compute resources

Your job runs on the cluster

Scheduler



Tell the scheduler what you want to do

Accessing software via environment modules

- `module load trinity`

- Automatically:
  - Sets, `$HPC_TRINITY_DIR`
    - To run Inchworm, simply type

```
inchworm --reads reads.fa --run_inchworm [opts]
```

  - Loads Bowtie and Allpaths, two Trinity dependencies
    - You don’t need to hunt those down, or worry if they are in your path or not

It’s all in the software!

Matt Gitzendanner

UF Research Computing

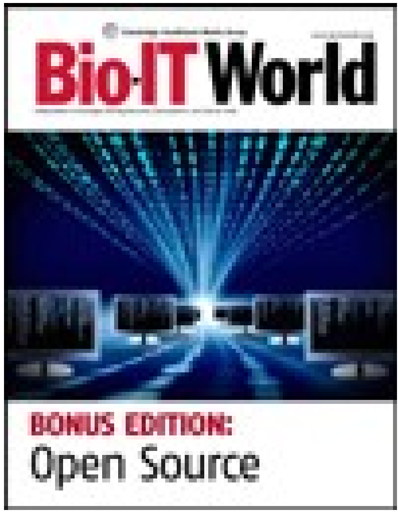
- Click to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level

Galaxy: Data intensive biology *for everyone*

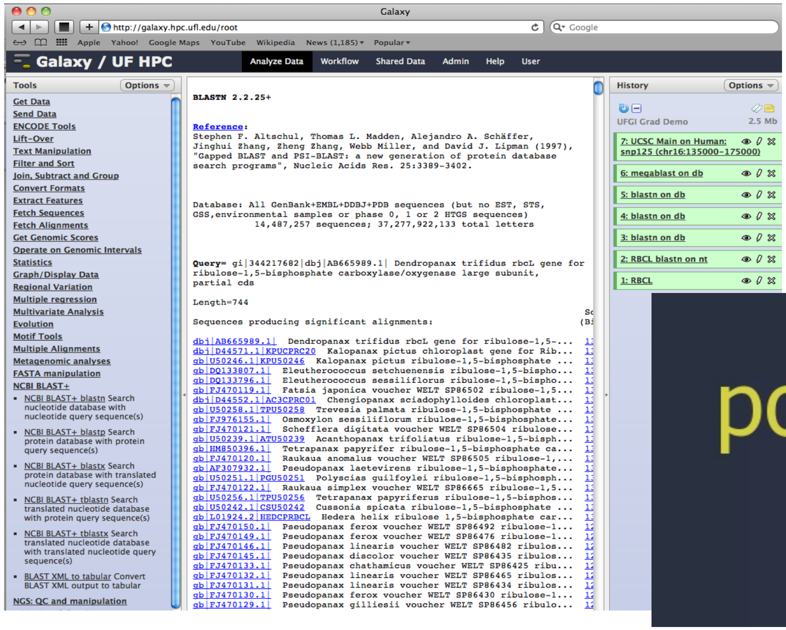
- Accessible, reproducible, transparent computational biology

- galaxy.hpc.ufl.edu
  - Local instance of Galaxy
    - Faster access to storage, easier upload
    - Local compute resources
    - Local control

What is Galaxy?



Galaxy Provides Life Support for NGS Exploration



Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NCBI BLAST+

- NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)
- NCBI BLAST+ blastp Search protein database with protein query sequence(s)
- NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
- NCBI BLAST+ tblastn Search translated nucleotide database with protein query sequence(s)
- NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)
- BLAST XML to tabular Convert BLAST XML output to tabular

NGS: QC and manipulation

```
BLASTN 2.2.25+


Reference:
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database
search programs", Nucleic Acids Res. 25:3389-3402.



Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           14,487,257 sequences; 37,277,922,133 total letters



Query= gi|344217682|dbj|AB665989.1| Dendropanax trifidus rbcL gene for
ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit,
partial cds

Length=744
                                                                      Sc
Sequences producing significant alignments:                          (Bi

dbj|AB665989.1|  Dendropanax trifidus rbcL gene for ribulose-1,5-...  13
dbj|D44571.1|KPUCPRC20  Kalopanax pictus chloroplast gene for Rib...  13
gb|U50246.1|KPU50246  Kalopanax pictus ribulose-1,5-bisphosphate ...  13
gb|DQ133807.1|  Eleutherococcus setchuenensis ribulose-1,5-bispho...  13
gb|DQ133796.1|  Eleutherococcus sessiliflorus ribulose-1,5-bispho...  13
gb|FJ470119.1|  Fatsia japonica voucher WELT SP86502 ribulose-1,5...  13
dbj|D44552.1|AC3CPRC01  Chengiopanax sciadophylloides chloroplast...  13
gb|U50258.1|TPU50258  Trevesia palmata ribulose-1,5-bisphosphate ...  13
gb|FJ976155.1|  Osmoxylon sessiliflorum ribulose-1,5-bisphosphate...  13
gb|FJ470121.1|  Schefflera digitata voucher WELT SP86504 ribulose...  13
gb|U50239.1|ATU50239  Acanthopanax trifoliatus ribulose-1,5-bisph...  13
gb|HM850396.1|  Tetrapanax papyrifer ribulose-1,5-bisphosphate ca...  13
gb|FJ470120.1|  Raukaua anomalus voucher WELT SP86505 ribulose-1,...  13
gb|AF307932.1|  Pseudopanax laetevirens ribulose-1,5-bisphosphate...  13
gb|U50251.1|PGU50251  Polyscias guilfoylei ribulose-1,5-bisphosph...  13
gb|FJ470122.1|  Raukaua simplex voucher WELT SP86665 ribulose-1,5...  13
gb|U50256.1|TPU50256  Tetrapanax papyriferus ribulose-1,5-bisphos...  13
gb|U50242.1|CSU50242  Cussonia spicata ribulose-1,5-bisphosphate ...  13
gb|L01924.2|HEDCPRBCL  Hedera helix ribulose 1,5-bisphosphate car...  13
gb|FJ470150.1|  Pseudopanax ferox voucher WELT SP86492 ribulose-1...  12
gb|FJ470149.1|  Pseudopanax ferox voucher WELT SP86476 ribulose-1...  12
gb|FJ470146.1|  Pseudopanax linearis voucher WELT SP86482 ribulos...  12
gb|FJ470145.1|  Pseudopanax discolor voucher WELT SP86435 ribulos...  12
gb|FJ470133.1|  Pseudopanax chathamicus voucher WELT SP86425 ribu...  12
gb|FJ470132.1|  Pseudopanax linearis voucher WELT SP86465 ribulos...  12
gb|FJ470131.1|  Pseudopanax linearis voucher WELT SP86434 ribulos...  12
gb|FJ470130.1|  Pseudopanax ferox voucher WELT SP86430 ribulose-1...  12
gb|FJ470129.1|  Pseudopanax gilliesii voucher WELT SP86456 ribulo...  12
```

UFGI Grad Demo 2.5 Mb

7: UCSC Main on Human: snp125 (chr16:135000-175000)
6: megablast on db
5: blastn on db
4: blastn on db
3: blastn on db
2: RBCL blastn on nt
1: RBCL

powered by Galaxy

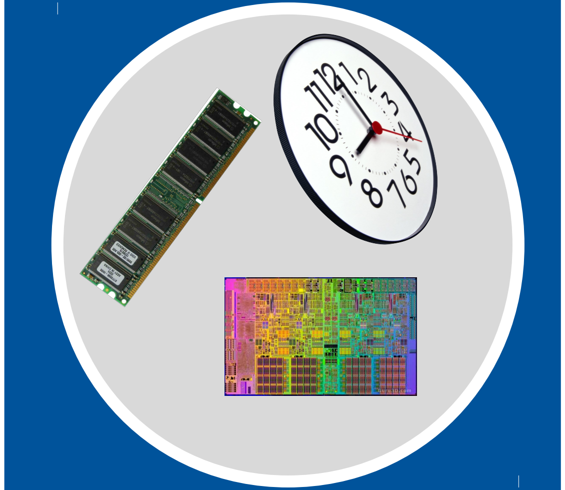
Cluster basics

User interaction

Galaxy

Login node (Head node)

Scheduler

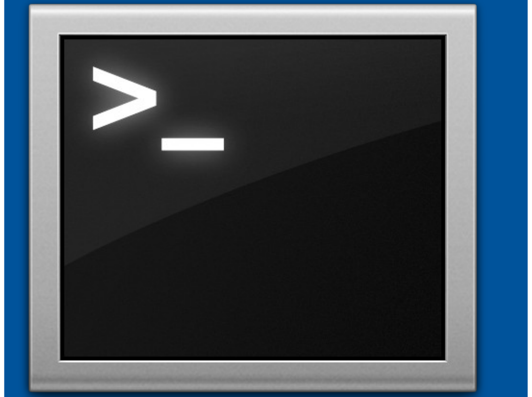


Tell the scheduler what you want to do

Cluster basics

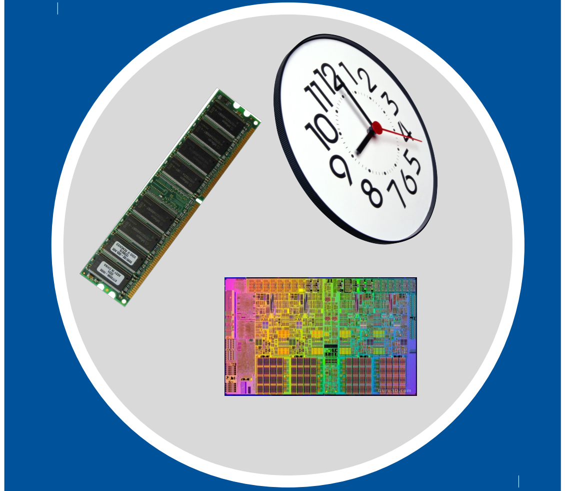
User interaction

Galaxy



Login node (Head node)

Scheduler



Tell the scheduler what you want to do

Pond *et al*. 2009, *Genome Research*

Resource

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond,[1,2,6,9] Samir Wadhawan,[3,6,7] Francesca Chiaromonte,[4] Guruprasad Ananda,[1,3] Wen-Yu Chung,[1,3,8] James Taylor,[1,5,9] Anton Nekrutenko,[1,3,9] and The Galaxy Team[1]

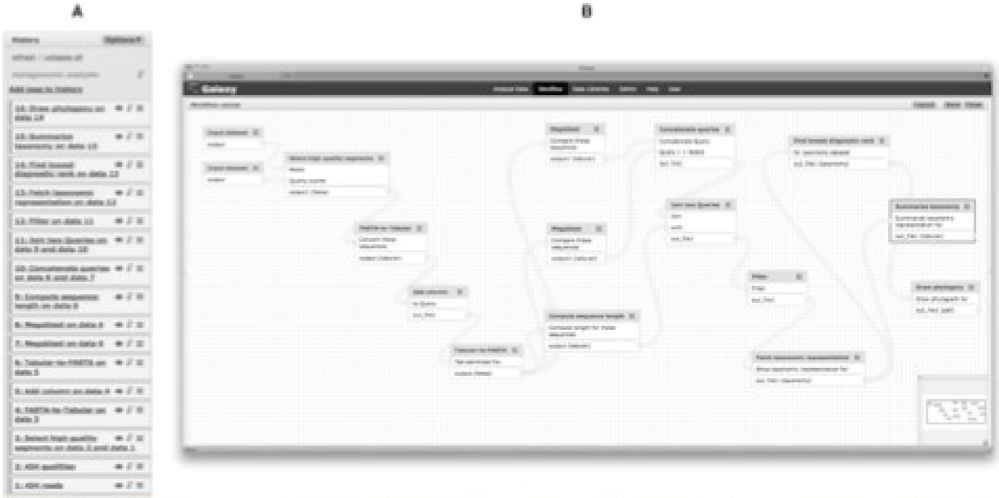
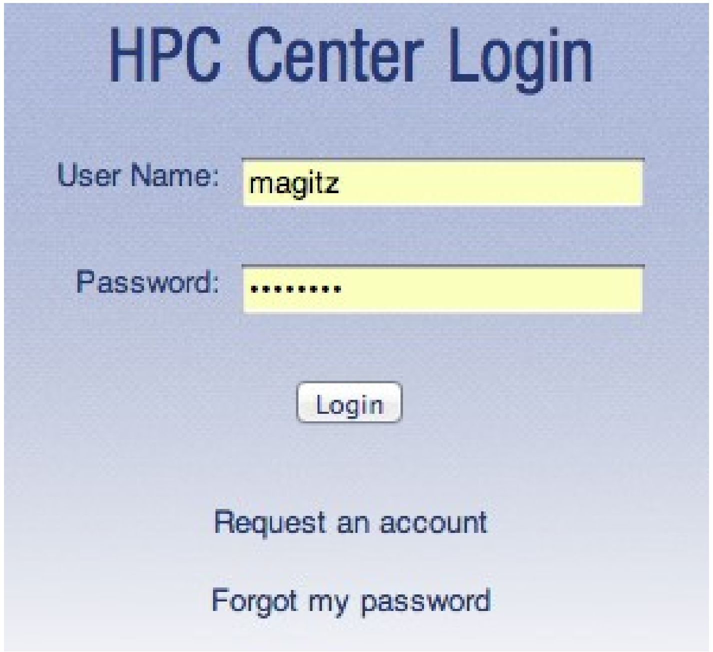


Figure 3. (*A*) Galaxy history pane showing all steps of a metagenomic analysis described in the study. (*B*) Workflow representation of analysis. Using workflow functionality, the user can rerun analyses in their entirety.

Galaxy demo

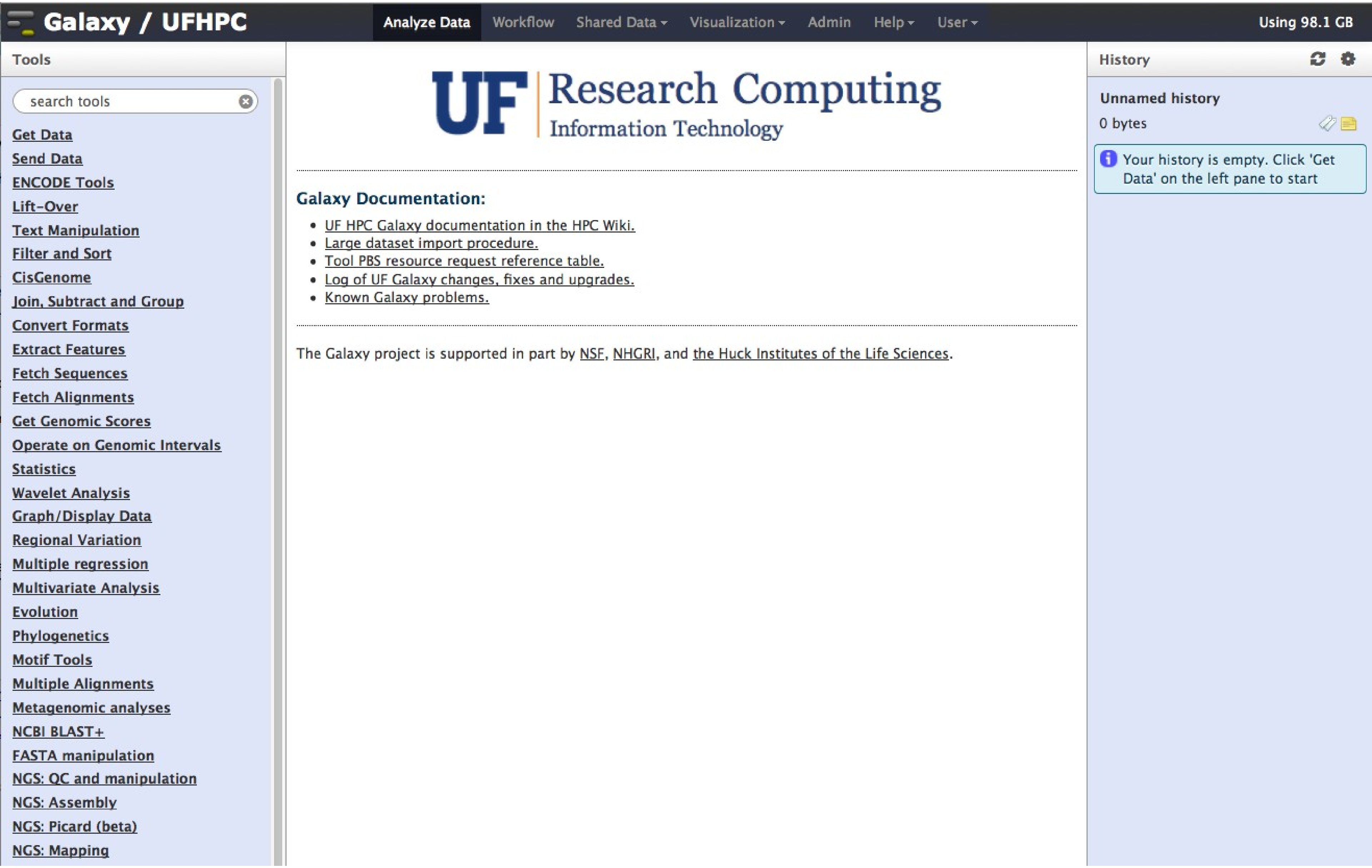
http://galaxy.hpc.ufl.edu



HPC Center Login

User Name: magitz

Password: •••••••••



UF | Research Computing
Information Technology

Galaxy Documentation:

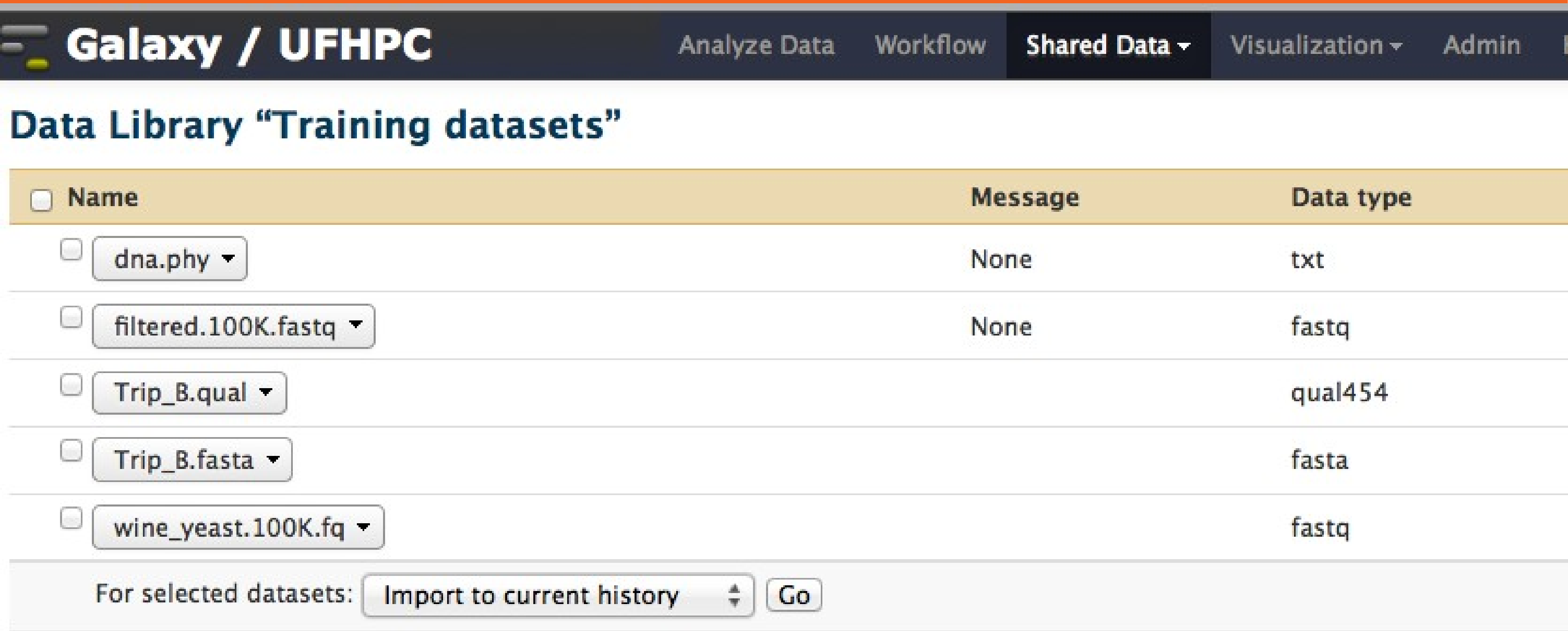
- UF HPC Galaxy documentation in the HPC Wiki.
- Large dataset import procedure.
- Tool PBS resource request reference table.
- Log of UF Galaxy changes, fixes and upgrades.
- Known Galaxy problems.

The Galaxy project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

Unnamed history

0 bytes

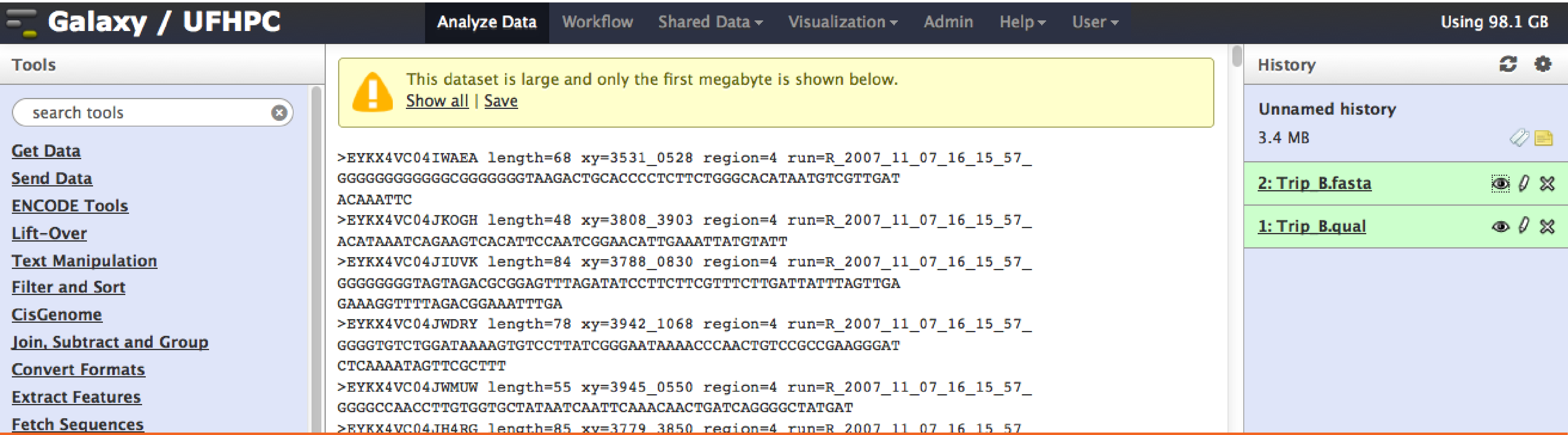
Your history is empty. Click 'Get Data' on the left pane to start



Data Library "Training datasets"

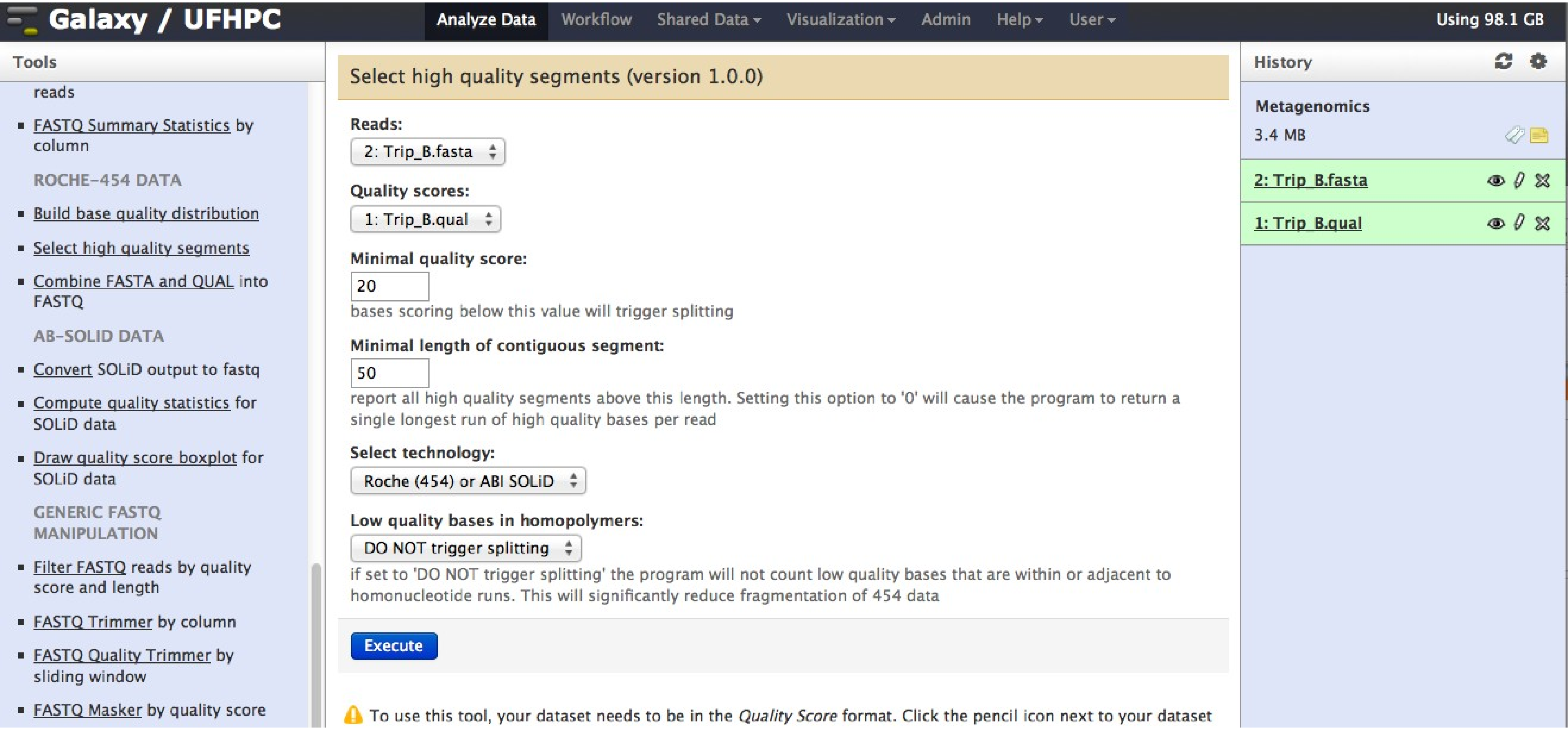
□ Name	Message	Data type
□ dna.phy	None	txt
□ filtered.100K.fastq	None	fastq
□ Trip_B.qual		qual454
□ Trip_B.fasta		fasta
□ wine_yeast.100K.fq		fastq

For selected datasets: Import to current history Go



This dataset is large and only the first megabyte is shown below.
Show all | Save

```
>EYKX4VC04IWAEA length=68 xy=3531_0528 region=4 run=R_2007_11_07_16_15_57_
GGGGGGGGGGGGCGGGGGGGTAAGACTGCACCCCTCTTCTGGGCACATAATGTCGTTGAT
ACAAATTC
>EYKX4VC04JKOGH length=48 xy=3808_3903 region=4 run=R_2007_11_07_16_15_57_
ACATAAATCAGAAGTCACATTCCAATCGGAACATTGAAATTATGTATT
>EYKX4VC04JIUVK length=84 xy=3788_0830 region=4 run=R_2007_11_07_16_15_57_
GGGGGGGGTAGTAGACGCGGAGTTTAGATATCCTTCTTCGTTTCTTGATTATTTAGTTGA
GAAAGGTTTTAGACGGAAATTTGA
>EYKX4VC04JWDRY length=78 xy=3942_1068 region=4 run=R_2007_11_07_16_15_57_
GGGGTGTCTGGATAAAAGTGTCCTTATCGGGAATAAAACCCAACTGTCCGCCGAAGGGAT
CTCAAAATAGTTCGCTTT
>EYKX4VC04JWMUW length=55 xy=3945_0550 region=4 run=R_2007_11_07_16_15_57_
GGGGCCAACCTTGTGGTGCTATAATCAATTCAAACAACTGATCAGGGGCTATGAT
>EYKX4VC04JH4RG length=85 xy=3779_3850 region=4 run=R_2007_11_07_16_15_57
```


Galaxy / UFHPC
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 98.1 GB
Tools
reads
FASTQ Summary Statistics by column
ROCHE-454 DATA
Build base quality distribution
Select high quality segments
Combine FASTA and QUAL into FASTQ
AB-SOLID DATA
Convert SOLiD output to fastq
Compute quality statistics for SOLiD data
Draw quality score boxplot for SOLiD data
GENERIC FASTQ MANIPULATION
Filter FASTQ reads by quality score and length
FASTQ Trimmer by column
FASTQ Quality Trimmer by sliding window
FASTQ Masker by quality score
History
Metagenomics
3.4 MB
2: Trip_B.fasta
1: Trip_B.qual

Select high quality segments (version 1.0.0)

Reads: 2: Trip_B.fasta

Quality scores: 1: Trip_B.qual

Minimal quality score: 20

bases scoring below this value will trigger splitting

Minimal length of contiguous segment: 50

report all high quality segments above this length. Setting this option to '0' will cause the program to return a single longest run of high quality bases per read

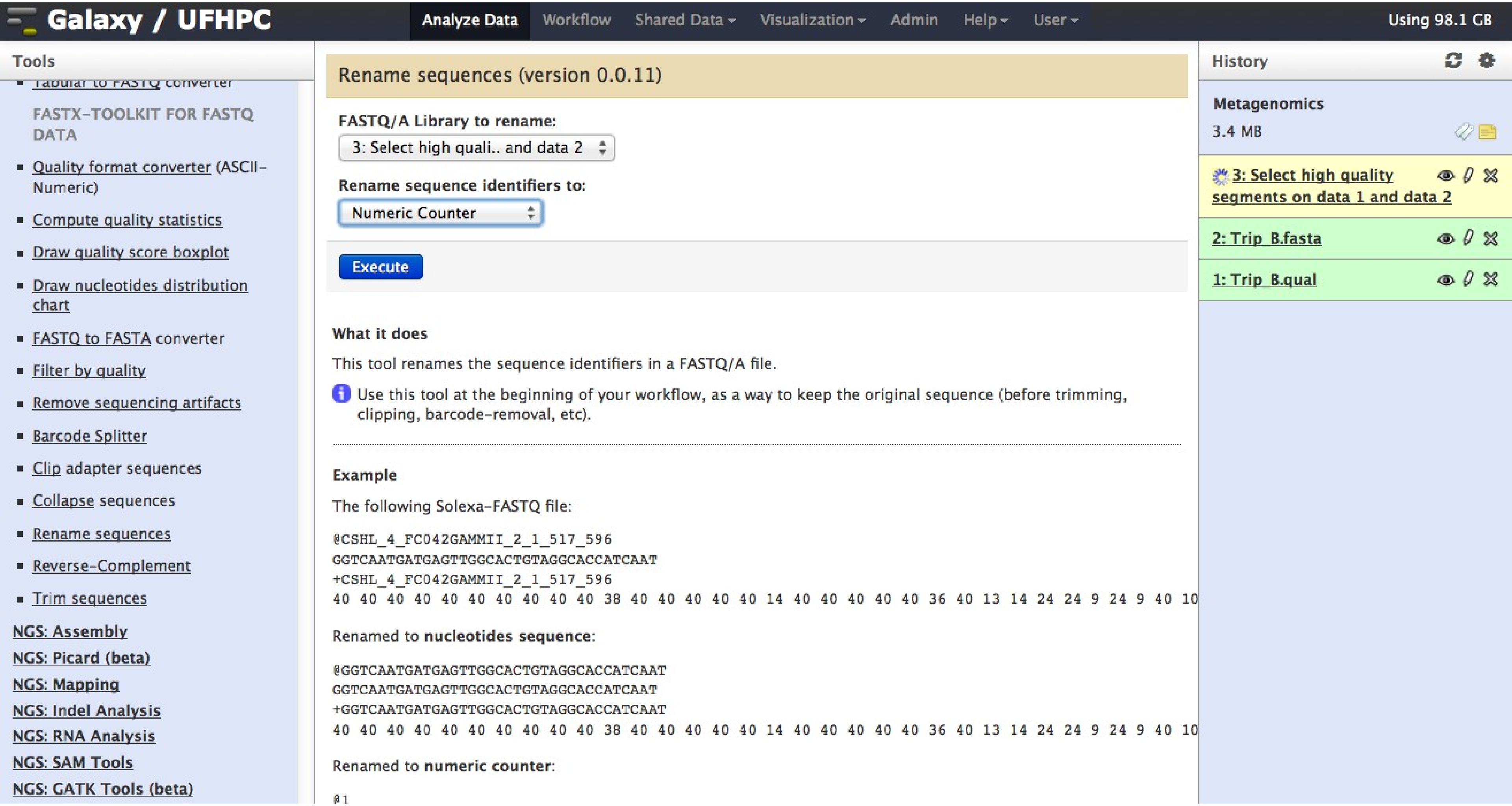
Select technology: Roche (454) or ABI SOLiD

Low quality bases in homopolymers: DO NOT trigger splitting

if set to 'DO NOT trigger splitting' the program will not count low quality bases that are within or adjacent to homonucleotide runs. This will significantly reduce fragmentation of 454 data

Execute

⚠ To use this tool, your dataset needs to be in the *Quality Score* format. Click the pencil icon next to your dataset



Rename sequences (version 0.0.11)

FASTQ/A Library to rename: 3: Select high quali.. and data 2

Rename sequence identifiers to: Numeric Counter

Execute

What it does

This tool renames the sequence identifiers in a FASTQ/A file.

Use this tool at the beginning of your workflow, as a way to keep the original sequence (before trimming, clipping, barcode-removal, etc).

Example

The following Solexa-FASTQ file:

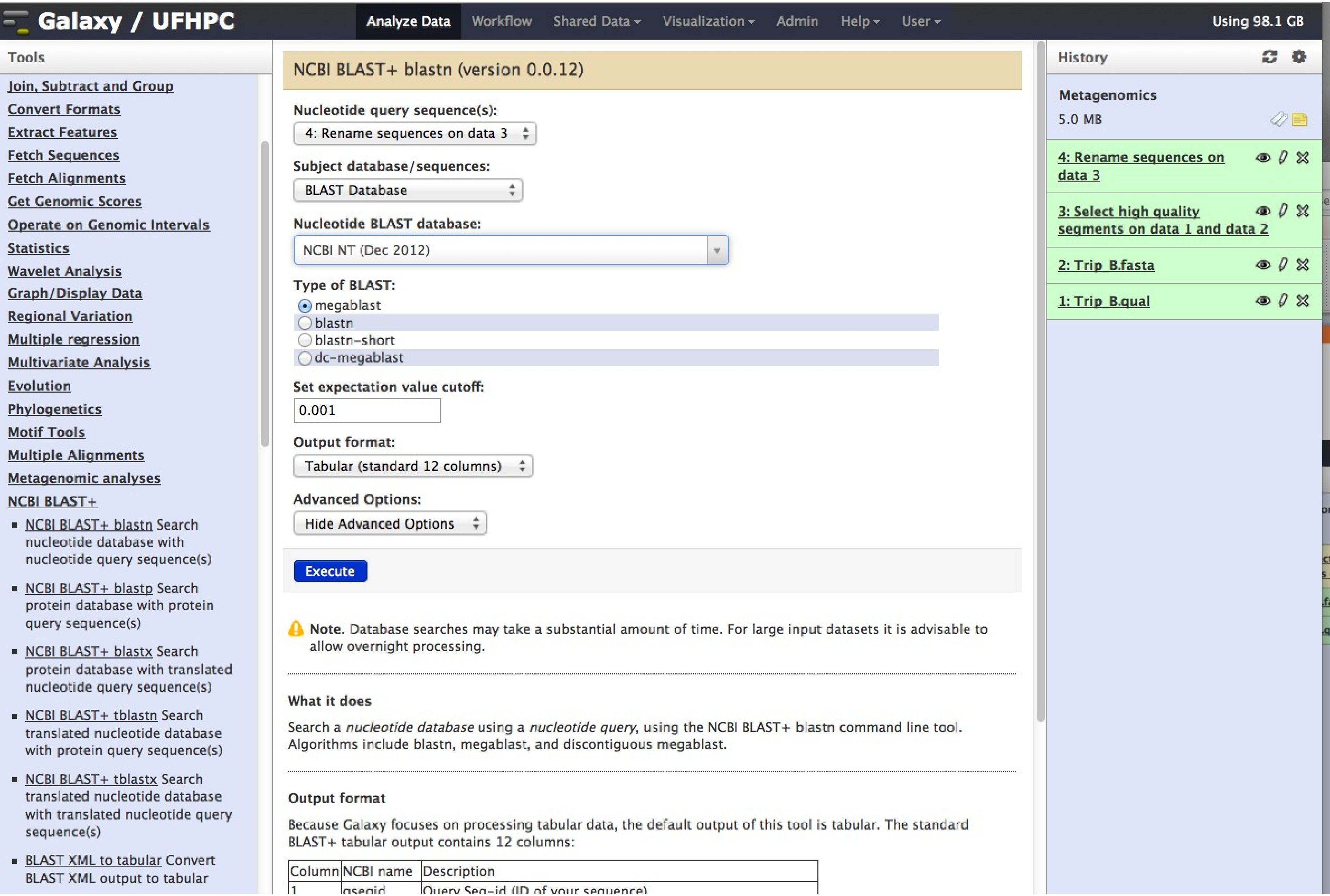
```
@CSHL_4_FC042GAMMII_2_1_517_596
GGTCAATGATGAGTTGGCACTGTAGGCACCATCAAT
+CSHL_4_FC042GAMMII_2_1_517_596
40 40 40 40 40 40 40 40 40 40 38 40 40 40 40 40 14 40 40 40 40 40 36 40 13 14 24 24 9 24 9 40 10
```

Renamed to **nucleotides sequence**:

```
@GGTCAATGATGAGTTGGCACTGTAGGCACCATCAAT
GGTCAATGATGAGTTGGCACTGTAGGCACCATCAAT
+GGTCAATGATGAGTTGGCACTGTAGGCACCATCAAT
40 40 40 40 40 40 40 40 40 40 38 40 40 40 40 40 14 40 40 40 40 40 36 40 13 14 24 24 9 24 9 40 10
```

Renamed to **numeric counter**:

```
@1
```



- NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)
- NCBI BLAST+ blastp Search protein database with protein query sequence(s)
- NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
- NCBI BLAST+ tblastn Search translated nucleotide database with protein query sequence(s)
- NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)
- BLAST XML to tabular Convert BLAST XML output to tabular

NCBI BLAST+ blastn (version 0.0.12)

Nucleotide query sequence(s): 4: Rename sequences on data 3

Subject database/sequences: BLAST Database

Nucleotide BLAST database: NCBI NT (Dec 2012)

Type of BLAST:

● megablast
○ blastn
○ blastn-short
○ dc-megablast

Set expectation value cutoff: 0.001

Output format: Tabular (standard 12 columns)

Advanced Options: Hide Advanced Options

Execute

Note. Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.

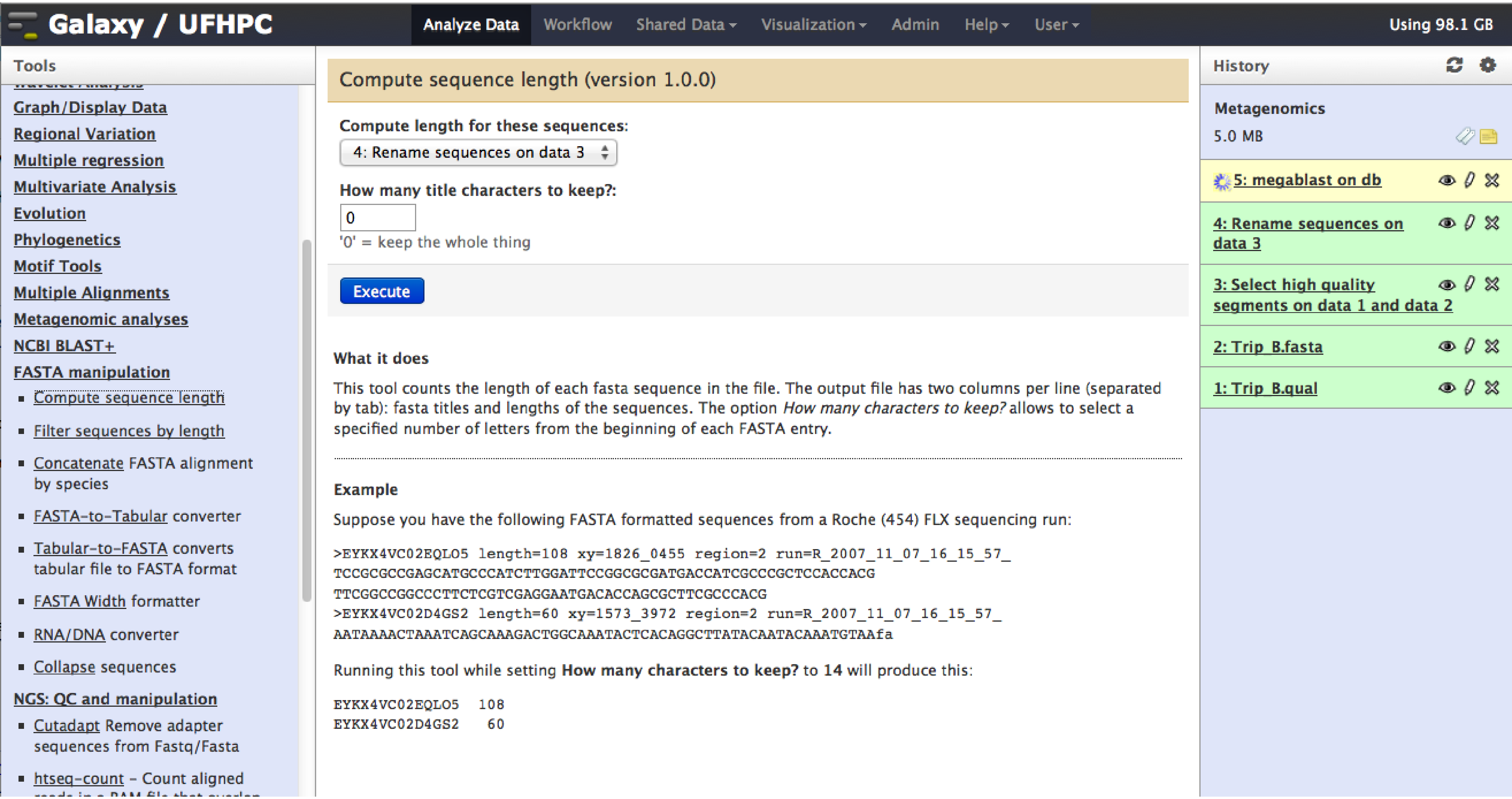
What it does

Search a *nucleotide database* using a *nucleotide query*, using the NCBI BLAST+ blastn command line tool. Algorithms include blastn, megablast, and discontiguous megablast.

Output format

Because Galaxy focuses on processing tabular data, the default output of this tool is tabular. The standard BLAST+ tabular output contains 12 columns:

Column	NCBI name	Description
1	qseqid	Query Seq-id (ID of your sequence)



Compute sequence length (version 1.0.0)

Compute length for these sequences: 4: Rename sequences on data 3

How many title characters to keep?: 0

'0' = keep the whole thing

Execute

What it does

This tool counts the length of each fasta sequence in the file. The output file has two columns per line (separated by tab): fasta titles and lengths of the sequences. The option *How many characters to keep?* allows to select a specified number of letters from the beginning of each FASTA entry.

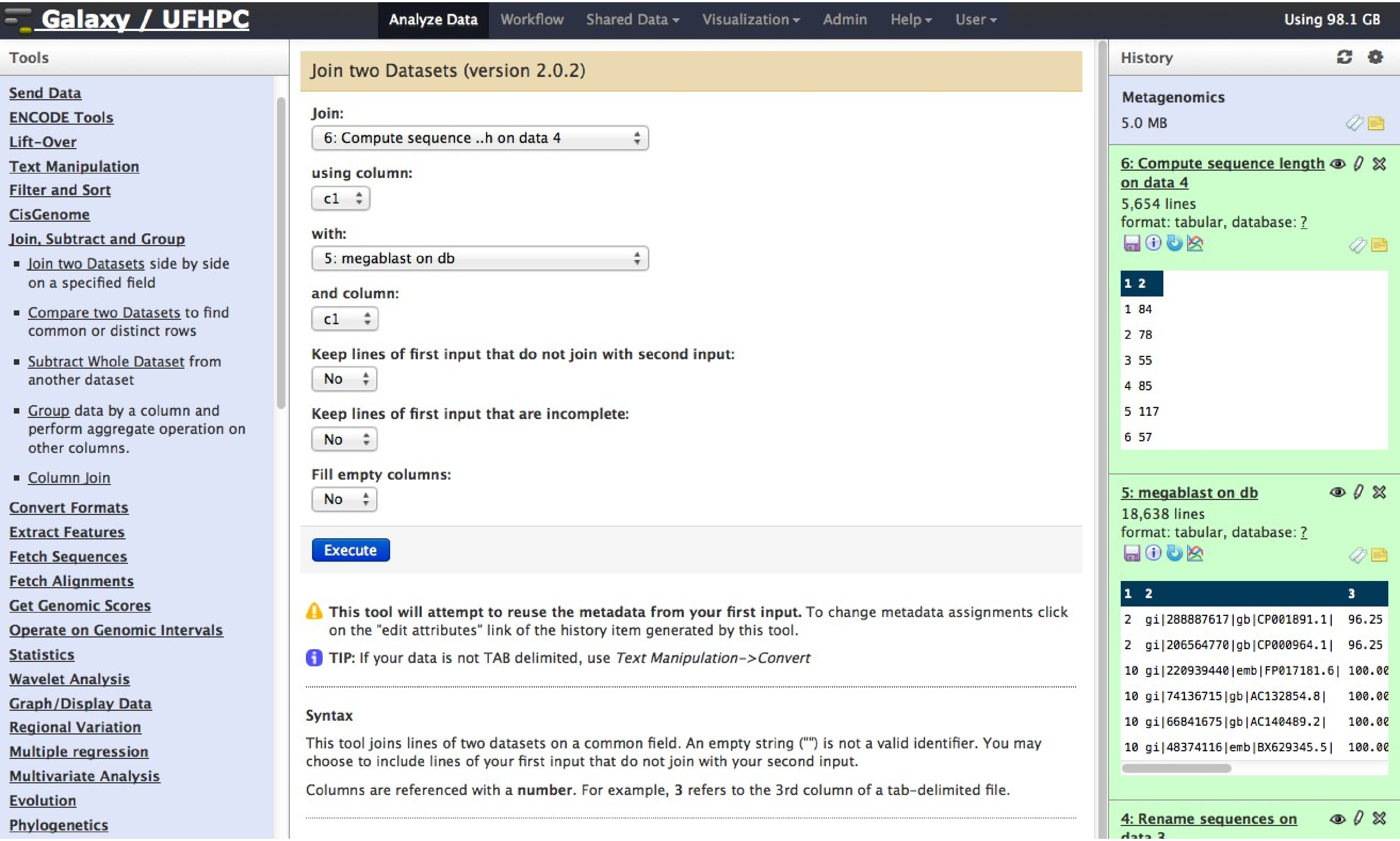
Example

Suppose you have the following FASTA formatted sequences from a Roche (454) FLX sequencing run:

```
>EYKX4VC02EQLO5 length=108 xy=1826_0455 region=2 run=R_2007_11_07_16_15_57_
TCCGCGCCGAGCATGCCCATCTTGGATTCCGGCGCGATGACCATCGCCCGCTCCACCACG
TTCGGCCGGCCCTTCTCGTCGAGGAATGACACCAGCGCTTCGCCCACG
>EYKX4VC02D4GS2 length=60 xy=1573_3972 region=2 run=R_2007_11_07_16_15_57_
AATAAAACTAAATCAGCAAAGACTGGCAAATACTCACAGGCTTATACAATACAAATGTAAfa
```

Running this tool while setting **How many characters to keep?** to **14** will produce this:

```
EYKX4VC02EQLO5  108
EYKX4VC02D4GS2   60
```



Join two Datasets (version 2.0.2)

Join: 6: Compute sequence ..h on data 4

using column: c1

with: 5: megablast on db

and column: c1

Keep lines of first input that do not join with second input: No

Keep lines of first input that are incomplete: No

Fill empty columns: No

This tool will attempt to reuse the metadata from your first input. To change metadata assignments click on the "edit attributes" link of the history item generated by this tool.

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax

This tool joins lines of two datasets on a common field. An empty string ("") is not a valid identifier. You may choose to include lines of your first input that do not join with your second input.

Columns are referenced with a **number**. For example, **3** refers to the 3rd column of a tab-delimited file.

Metagenomics
5.0 MB

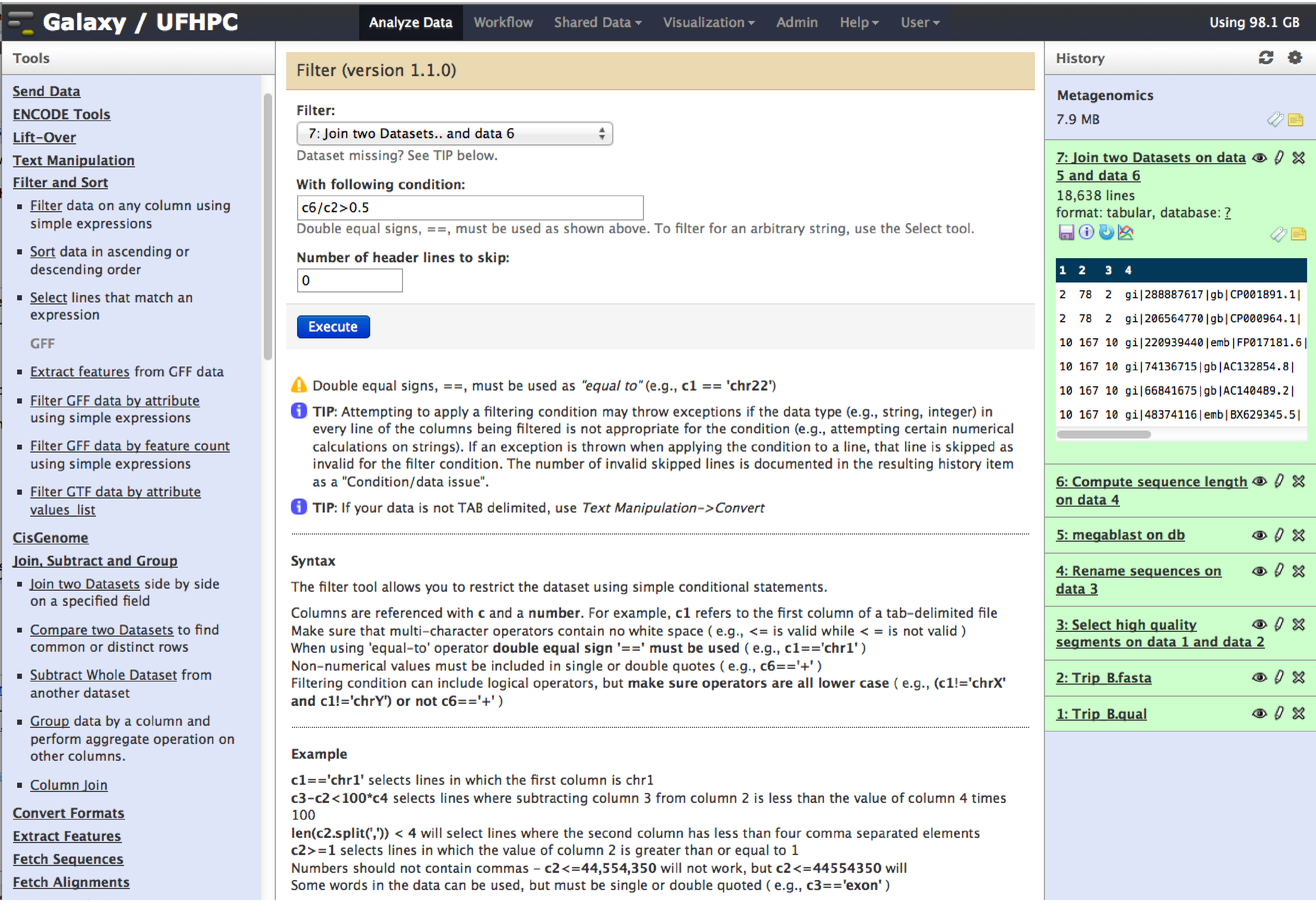
6: Compute sequence length on data 4
5,654 lines
format: tabular, database: ?

1	2
1	84
2	78
3	55
4	85
5	117
6	57

5: megablast on db
18,638 lines
format: tabular, database: ?

1	2	3
2	gi|288887617|gb|CP001891.1|	96.25
2	gi|206564770|gb|CP000964.1|	96.25
10	gi|220939440|emb|FP017181.6|	100.00
10	gi|74136715|gb|AC132854.8|	100.00
10	gi|66841675|gb|AC140489.2|	100.00
10	gi|48374116|emb|BX629345.5|	100.00

4: Rename sequences on data 3



Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression

GFF

- Extract features from GFF data
- Filter GFF data by attribute using simple expressions
- Filter GFF data by feature count using simple expressions
- Filter GTF data by attribute values_list

CisGenome
Join, Subtract and Group

- Join two Datasets side by side on a specified field
- Compare two Datasets to find common or distinct rows
- Subtract Whole Dataset from another dataset
- Group data by a column and perform aggregate operation on other columns.
- Column Join

Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments

Filter (version 1.1.0)

Filter:
7: Join two Datasets.. and data 6
Dataset missing? See TIP below.

With following condition:
c6/c2>0.5
Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip:
0

⚠ Double equal signs, ==, must be used as *"equal to"* (e.g., **c1 == 'chr22'**)

ⓘ **TIP**: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

ⓘ **TIP**: If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of a tab-delimited file
Make sure that multi-character operators contain no white space (e.g., **<=** is valid while **< =** is not valid)
When using 'equal-to' operator **double equal sign '==' must be used** (e.g., **c1=='chr1'**)
Non-numerical values must be included in single or double quotes (e.g., **c6=='+'**)
Filtering condition can include logical operators, but **make sure operators are all lower case** (e.g., **(c1!='chrX' and c1!='chrY') or not c6=='+'**)

Example

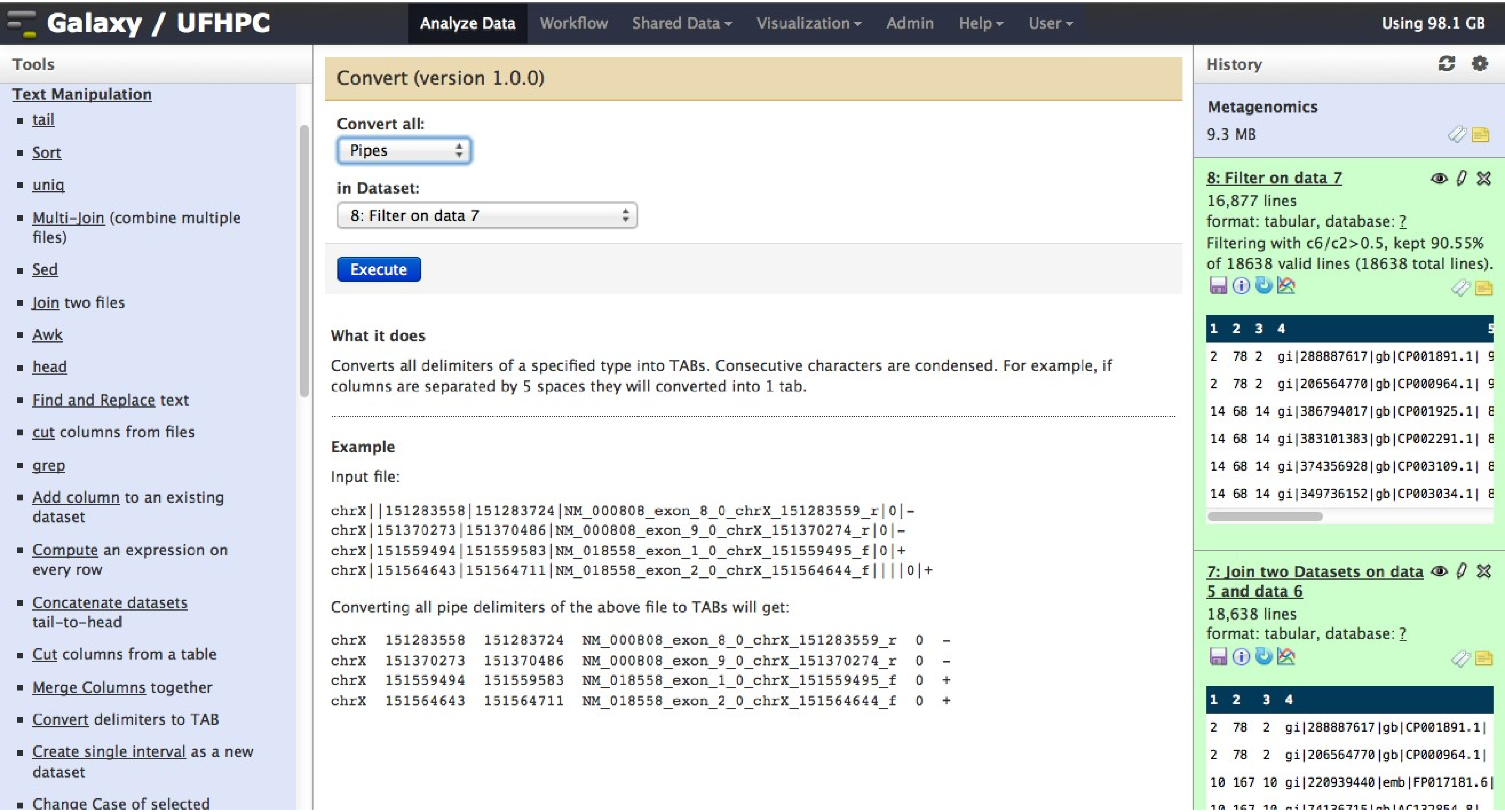
c1=='chr1' selects lines in which the first column is chr1
c3-c2<100*c4 selects lines where subtracting column 3 from column 2 is less than the value of column 4 times 100
len(c2.split(',')) < 4 will select lines where the second column has less than four comma separated elements
c2>=1 selects lines in which the value of column 2 is greater than or equal to 1
Numbers should not contain commas - **c2<=44,554,350** will not work, but **c2<=44554350** will
Some words in the data can be used, but must be single or double quoted (e.g., **c3=='exon'**)

Metagenomics
7.9 MB

7: Join two Datasets on data 5 and data 6
18,638 lines
format: tabular, database: ?

```
1  2   3  4
2  78  2  gi|288887617|gb|CP001891.1|
2  78  2  gi|206564770|gb|CP000964.1|
10 167 10 gi|220939440|emb|FP017181.6|
10 167 10 gi|74136715|gb|AC132854.8|
10 167 10 gi|66841675|gb|AC140489.2|
10 167 10 gi|48374116|emb|BX629345.5|
```

6: Compute sequence length on data 4
5: megablast on db
4: Rename sequences on data 3
3: Select high quality segments on data 1 and data 2
2: Trip_B.fasta
1: Trip_B.qual



Convert (version 1.0.0)

Convert all: Pipes

in Dataset: 8: Filter on data 7

What it does

Converts all delimiters of a specified type into TABs. Consecutive characters are condensed. For example, if columns are separated by 5 spaces they will converted into 1 tab.

Example

Input file:

```
chrX||151283558|151283724|NM_000808_exon_8_0_chrX_151283559_r|0|-
chrX|151370273|151370486|NM_000808_exon_9_0_chrX_151370274_r|0|-
chrX|151559494|151559583|NM_018558_exon_1_0_chrX_151559495_f|0|+
chrX|151564643|151564711|NM_018558_exon_2_0_chrX_151564644_f||||0|+
```

Converting all pipe delimiters of the above file to TABs will get:

```
chrX  151283558  151283724  NM_000808_exon_8_0_chrX_151283559_r  0  -
chrX  151370273  151370486  NM_000808_exon_9_0_chrX_151370274_r  0  -
chrX  151559494  151559583  NM_018558_exon_1_0_chrX_151559495_f  0  +
chrX  151564643  151564711  NM_018558_exon_2_0_chrX_151564644_f  0  +
```

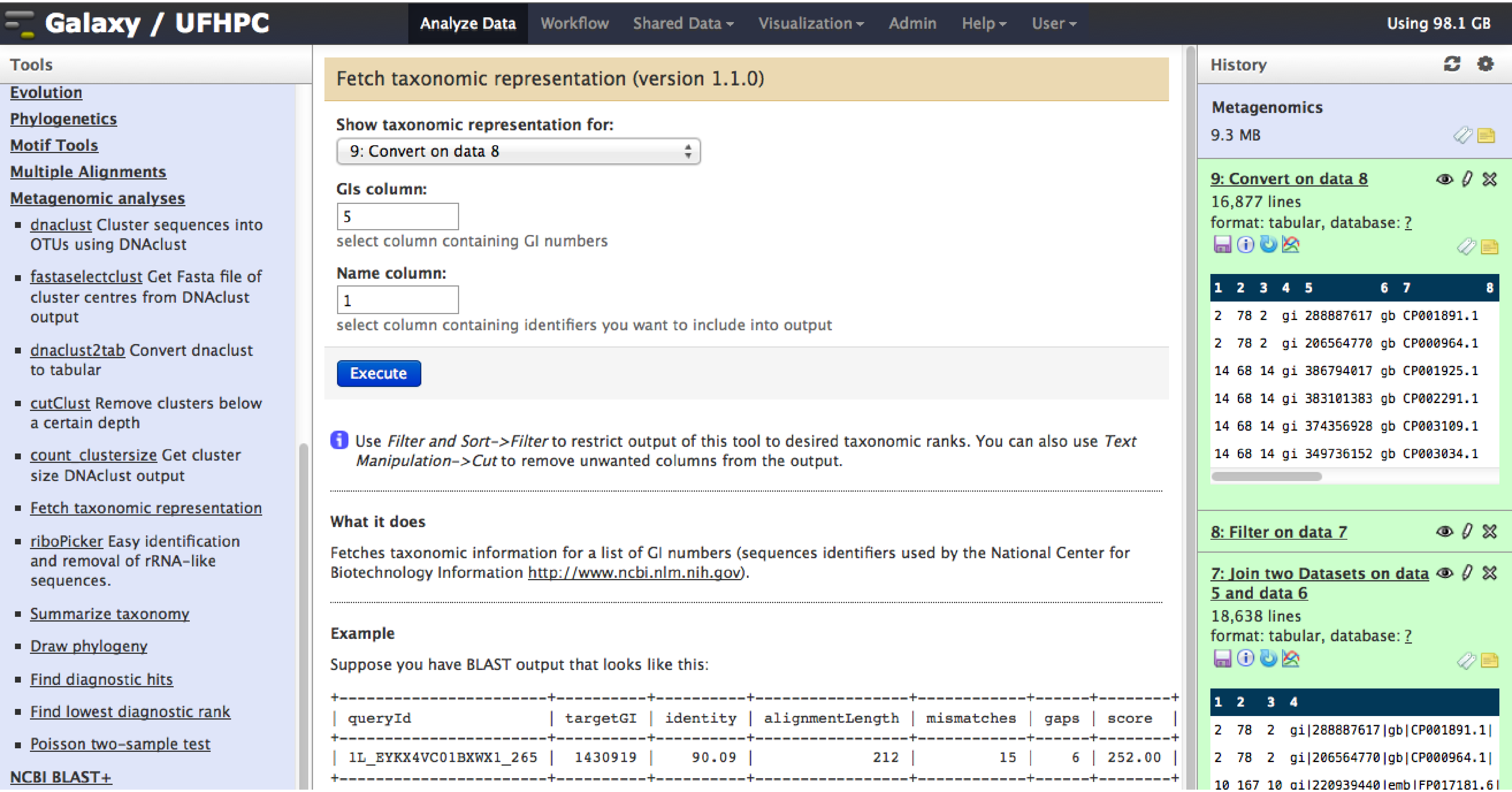
Metagenomics
9.3 MB

8: Filter on data 7
16,877 lines
format: tabular, database: ?
Filtering with c6/c2>0.5, kept 90.55% of 18638 valid lines (18638 total lines).

```
1  2  3  4                                   5
2  78 2  gi|288887617|gb|CP001891.1| 9
2  78 2  gi|206564770|gb|CP000964.1| 9
14 68 14 gi|386794017|gb|CP001925.1| 8
14 68 14 gi|383101383|gb|CP002291.1| 8
14 68 14 gi|374356928|gb|CP003109.1| 8
14 68 14 gi|349736152|gb|CP003034.1| 8
```

7: Join two Datasets on data 5 and data 6
18,638 lines
format: tabular, database: ?

```
1  2   3  4
2  78  2  gi|288887617|gb|CP001891.1|
2  78  2  gi|206564770|gb|CP000964.1|
10 167 10 gi|220939440|emb|FP017181.6|
```


Galaxy / UFHPC
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 98.1 GB
Tools
Evolution
Phylogenetics
Motif Tools
Multiple Alignments
Metagenomic analyses
dnaclust Cluster sequences into OTUs using DNAclust
fastaselectclust Get Fasta file of cluster centres from DNAclust output
dnaclust2tab Convert dnaclust to tabular
cutClust Remove clusters below a certain depth
count_clustersize Get cluster size DNAclust output
Fetch taxonomic representation
riboPicker Easy identification and removal of rRNA-like sequences.
Summarize taxonomy
Draw phylogeny
Find diagnostic hits
Find lowest diagnostic rank
Poisson two-sample test
NCBI BLAST+
History

Fetch taxonomic representation (version 1.1.0)

Show taxonomic representation for: 9: Convert on data 8

GIs column: 5

select column containing GI numbers

Name column: 1

select column containing identifiers you want to include into output

Execute

Use *Filter and Sort->Filter* to restrict output of this tool to desired taxonomic ranks. You can also use *Text Manipulation->Cut* to remove unwanted columns from the output.

What it does

Fetches taxonomic information for a list of GI numbers (sequences identifiers used by the National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov).

Example

Suppose you have BLAST output that looks like this:

```
+-----------------------+----------+----------+-----------------+------------+------+--------+
| queryId               | targetGI | identity | alignmentLength | mismatches | gaps | score  |
+-----------------------+----------+----------+-----------------+------------+------+--------+
| 1L_EYKX4VC01BXWX1_265 |  1430919 |    90.09 |             212 |         15 |    6 | 252.00 |
+-----------------------+----------+----------+-----------------+------------+------+--------+
```

Metagenomics

9.3 MB

9: Convert on data 8

16,877 lines

format: tabular, database: ?

```
1  2  3  4  5          6  7            8
2  78 2  gi 288887617 gb CP001891.1
2  78 2  gi 206564770 gb CP000964.1
14 68 14 gi 386794017 gb CP001925.1
14 68 14 gi 383101383 gb CP002291.1
14 68 14 gi 374356928 gb CP003109.1
14 68 14 gi 349736152 gb CP003034.1
```

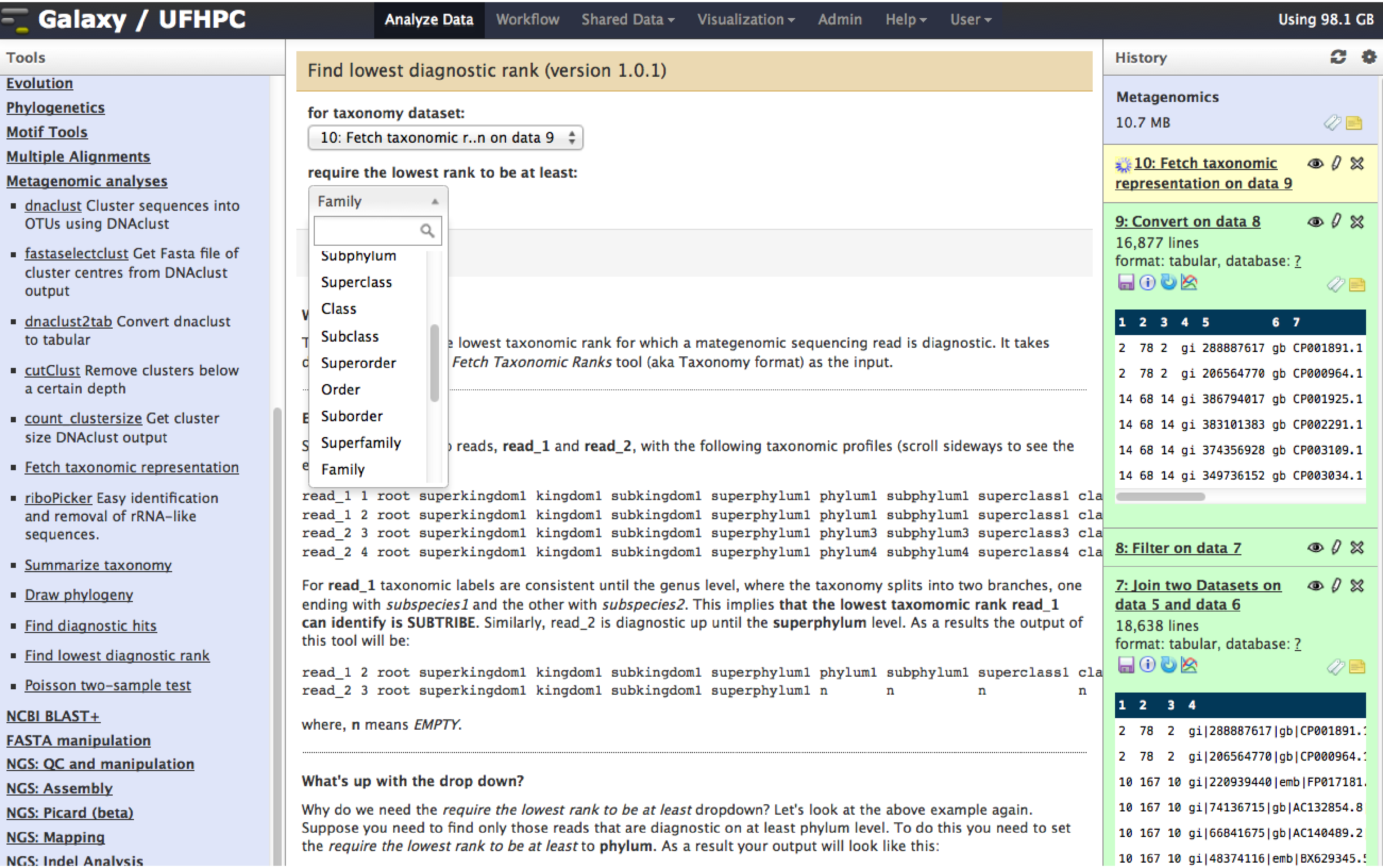
8: Filter on data 7

7: Join two Datasets on data 5 and data 6

18,638 lines

format: tabular, database: ?

```
1  2   3  4
2  78  2  gi|288887617|gb|CP001891.1|
2  78  2  gi|206564770|gb|CP000964.1|
10 167 10 gi|220939440|emb|FP017181.6|
```



Tools

- Evolution
- Phylogenetics
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
  - dnaclust Cluster sequences into OTUs using DNAclust
  - fastaselectclust Get Fasta file of cluster centres from DNAclust output
  - dnaclust2tab Convert dnaclust to tabular
  - cutClust Remove clusters below a certain depth
  - count_clustersize Get cluster size DNAclust output
  - Fetch taxonomic representation
  - riboPicker Easy identification and removal of rRNA-like sequences.
  - Summarize taxonomy
  - Draw phylogeny
  - Find diagnostic hits
  - Find lowest diagnostic rank
  - Poisson two-sample test
- NCBI BLAST+
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Assembly
- NGS: Picard (beta)
- NGS: Mapping
- NGS: Indel Analysis

Find lowest diagnostic rank (version 1.0.1)

for taxonomy dataset:

10: Fetch taxonomic r..n on data 9

require the lowest rank to be at least:

Family

Subphylum
Superclass
Class
Subclass
Superorder
Order
Suborder
Superfamily
Family

W

T lowest taxonomic rank for which a mategenomic sequencing read is diagnostic. It takes c *Fetch Taxonomic Ranks* tool (aka Taxonomy format) as the input.

E

S reads, **read_1** and **read_2**, with the following taxonomic profiles (scroll sideways to see the e

```
read_1 1 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum1 subphylum1 superclass1 cla
read_1 2 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum1 subphylum1 superclass1 cla
read_2 3 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum3 subphylum3 superclass3 cla
read_2 4 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum4 subphylum4 superclass4 cla
```

For **read_1** taxonomic labels are consistent until the genus level, where the taxonomy splits into two branches, one ending with *subspecies1* and the other with *subspecies2*. This implies **that the lowest taxomomic rank read_1 can identify is SUBTRIBE**. Similarly, read_2 is diagnostic up until the **superphylum** level. As a results the output of this tool will be:

```
read_1 2 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum1 subphylum1 superclass1 cla
read_2 3 root superkingdom1 kingdom1 subkingdom1 superphylum1 n       n          n           n
```

where, **n** means *EMPTY*.

What's up with the drop down?

Why do we need the *require the lowest rank to be at least* dropdown? Let's look at the above example again. Suppose you need to find only those reads that are diagnostic on at least phylum level. To do this you need to set the *require the lowest rank to be at least* to **phylum**. As a result your output will look like this:

History

Metagenomics
10.7 MB

10: Fetch taxonomic representation on data 9

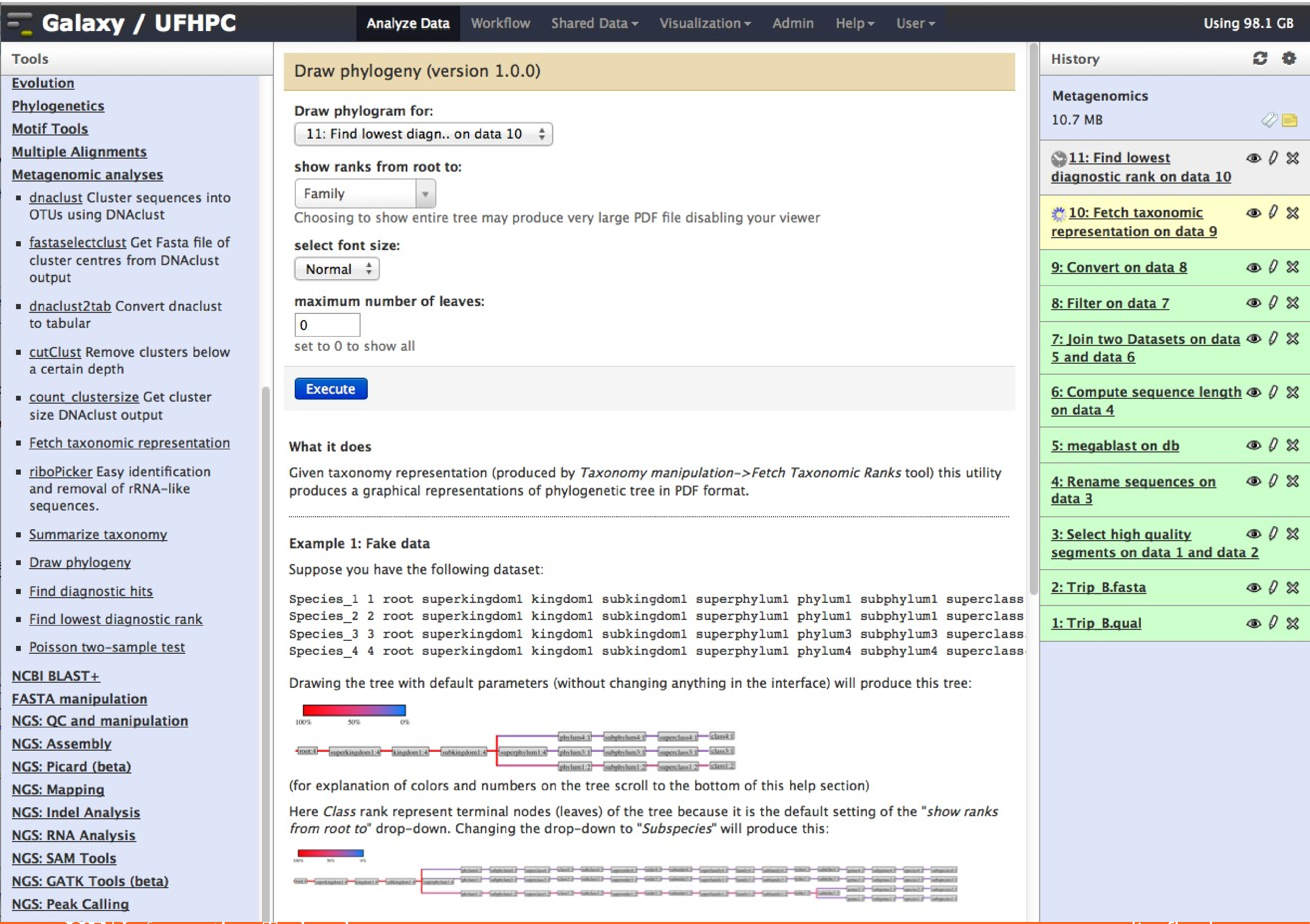
9: Convert on data 8
16,877 lines
format: tabular, database: ?

```
1  2  3  4  5          6  7
2  78 2  gi 288887617 gb CP001891.1
2  78 2  gi 206564770 gb CP000964.1
14 68 14 gi 386794017 gb CP001925.1
14 68 14 gi 383101383 gb CP002291.1
14 68 14 gi 374356928 gb CP003109.1
14 68 14 gi 349736152 gb CP003034.1
```

8: Filter on data 7

7: Join two Datasets on data 5 and data 6
18,638 lines
format: tabular, database: ?

```
1  2   3  4
2  78  2  gi|288887617|gb|CP001891.1
2  78  2  gi|206564770|gb|CP000964.1
10 167 10 gi|220939440|emb|FP017181.
10 167 10 gi|74136715|gb|AC132854.8
10 167 10 gi|66841675|gb|AC140489.2
10 167 10 gi|48374116|emb|BX629345.5
```



Tools

- Evolution
- Phylogenetics
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
  - dnaclust Cluster sequences into OTUs using DNAclust
  - fastaselectclust Get Fasta file of cluster centres from DNAclust output
  - dnaclust2tab Convert dnaclust to tabular
  - cutClust Remove clusters below a certain depth
  - count_clustersize Get cluster size DNAclust output
  - Fetch taxonomic representation
  - riboPicker Easy identification and removal of rRNA-like sequences.
  - Summarize taxonomy
  - Draw phylogeny
  - Find diagnostic hits
  - Find lowest diagnostic rank
  - Poisson two-sample test
- NCBI BLAST+
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Assembly
- NGS: Picard (beta)
- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- NGS: Peak Calling

Draw phylogeny (version 1.0.0)

Draw phylogram for:
11: Find lowest diagn.. on data 10

show ranks from root to:
Family
Choosing to show entire tree may produce very large PDF file disabling your viewer

select font size:
Normal

maximum number of leaves:
0
set to 0 to show all

What it does

Given taxonomy representation (produced by *Taxonomy manipulation->Fetch Taxonomic Ranks* tool) this utility produces a graphical representations of phylogenetic tree in PDF format.

Example 1: Fake data

Suppose you have the following dataset:

```
Species_1 1 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum1 subphylum1 superclass
Species_2 2 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum1 subphylum1 superclass
Species_3 3 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum3 subphylum3 superclass
Species_4 4 root superkingdom1 kingdom1 subkingdom1 superphylum1 phylum4 subphylum4 superclass
```

Drawing the tree with default parameters (without changing anything in the interface) will produce this tree:

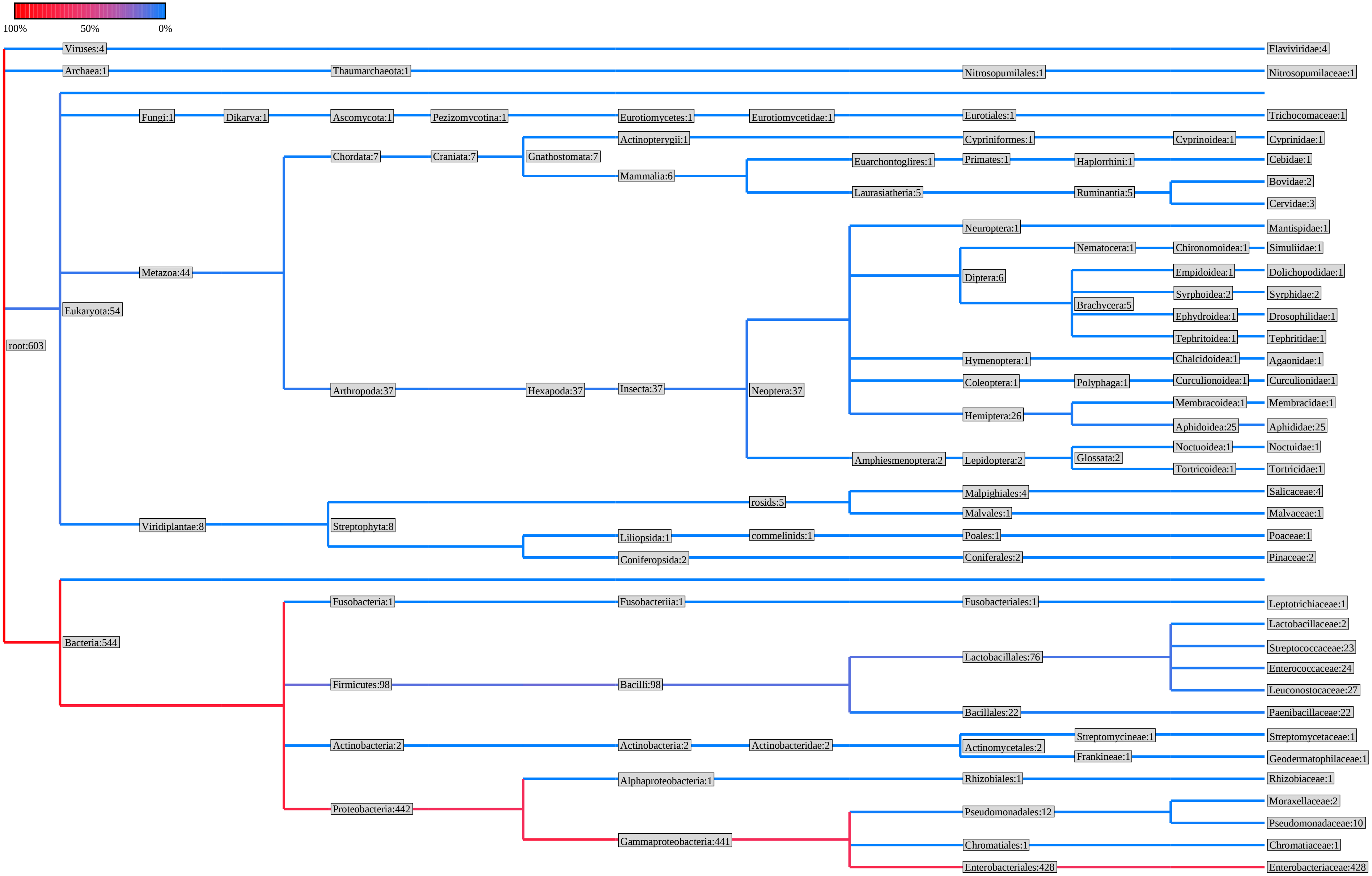
(for explanation of colors and numbers on the tree scroll to the bottom of this help section)

Here *Class* rank represent terminal nodes (leaves) of the tree because it is the default setting of the "*show ranks from root to*" drop-down. Changing the drop-down to "*Subspecies*" will produce this:

History

Metagenomics
10.7 MB

- 11: Find lowest diagnostic rank on data 10
- 10: Fetch taxonomic representation on data 9
- 9: Convert on data 8
- 8: Filter on data 7
- 7: Join two Datasets on data 5 and data 6
- 6: Compute sequence length on data 4
- 5: megablast on db
- 4: Rename sequences on data 3
- 3: Select high quality segments on data 1 and data 2
- 2: Trip_B.fasta
- 1: Trip_B.qual



100% 50% 0%

root:603

Viruses:4 — Flaviviridae:4

Archaea:1 — Thaumarchaeota:1 — Nitrosopumilales:1 — Nitrosopumilaceae:1

Eukaryota:54

Fungi:1 — Dikarya:1 — Ascomycota:1 — Pezizomycotina:1 — Eurotiomycetes:1 — Eurotiomycetidae:1 — Eurotiales:1 — Trichocomaceae:1

Metazoa:44

Chordata:7 — Craniata:7 — Gnathostomata:7

Actinopterygii:1 — Cypriniformes:1 — Cyprinoidea:1 — Cyprinidae:1

Mammalia:6

Euarchontoglires:1 — Primates:1 — Haplorrhini:1 — Cebidae:1

Laurasiatheria:5 — Ruminantia:5 — Bovidae:2, Cervidae:3

Arthropoda:37 — Hexapoda:37 — Insecta:37 — Neoptera:37

Neuroptera:1 — Mantispidae:1

Diptera:6

Nematocera:1 — Chironomoidea:1 — Simuliidae:1

Brachycera:5 — Empidoidea:1 — Dolichopodidae:1; Syrphoidea:2 — Syrphidae:2; Ephydroidea:1 — Drosophilidae:1; Tephritoidea:1 — Tephritidae:1

Hymenoptera:1 — Chalcidoidea:1 — Agaonidae:1

Coleoptera:1 — Polyphaga:1 — Curculionoidea:1 — Curculionidae:1

Hemiptera:26 — Membracoidea:1 — Membracidae:1; Aphidoidea:25 — Aphididae:25

Amphiesmenoptera:2 — Lepidoptera:2 — Glossata:2 — Noctuoidea:1 — Noctuidae:1; Tortricoidea:1 — Tortricidae:1

Viridiplantae:8 — Streptophyta:8

rosids:5 — Malpighiales:4 — Salicaceae:4; Malvales:1 — Malvaceae:1

Liliopsida:1 — commelinids:1 — Poales:1 — Poaceae:1

Coniferopsida:2 — Coniferales:2 — Pinaceae:2

Bacteria:544

Fusobacteria:1 — Fusobacteriia:1 — Fusobacteriales:1 — Leptotrichiaceae:1

Firmicutes:98 — Bacilli:98

Lactobacillales:76 — Lactobacillaceae:2, Streptococcaceae:23, Enterococcaceae:24, Leuconostocaceae:27

Bacillales:22 — Paenibacillaceae:22

Actinobacteria:2 — Actinobacteria:2 — Actinobacteridae:2 — Actinomycetales:2 — Streptomycineae:1 — Streptomycetaceae:1; Frankineae:1 — Geodermatophilaceae:1

Proteobacteria:442

Alphaproteobacteria:1 — Rhizobiales:1 — Rhizobiaceae:1

Gammaproteobacteria:441

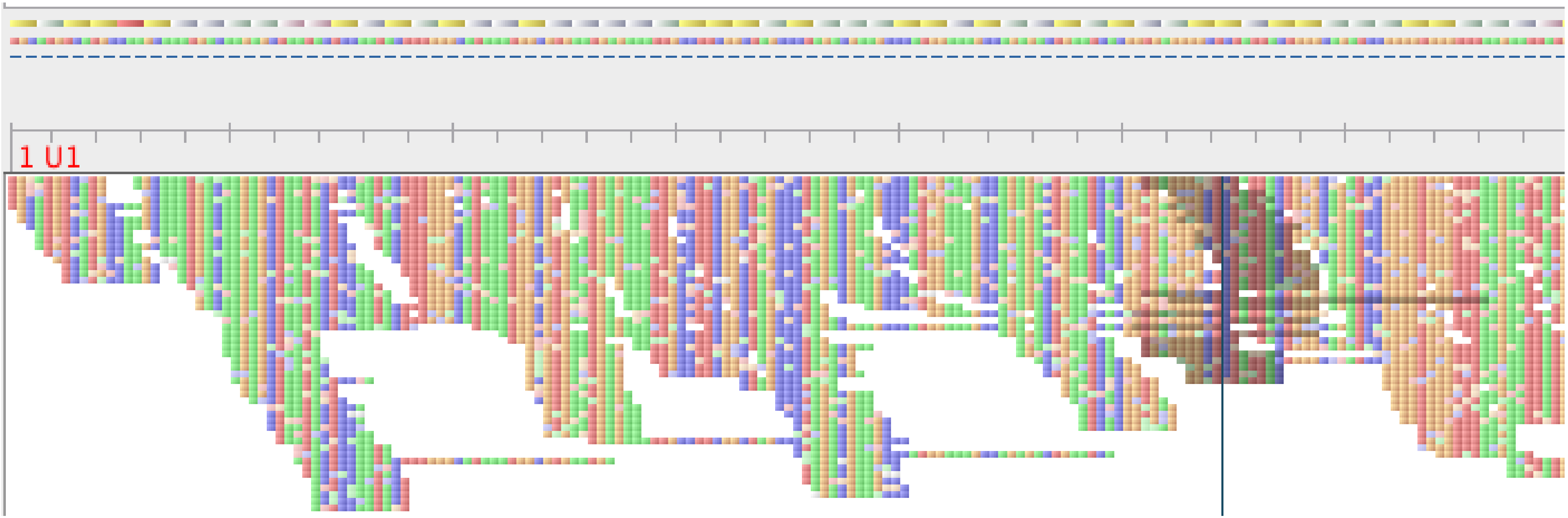
Pseudomonadales:12 — Moraxellaceae:2, Pseudomonadaceae:10

Chromatiales:1 — Chromatiaceae:1

Enterobacteriales:428 — Enterobacteriaceae:428

Reference-based mapping

- Map NGS reads onto a reference genome
  - Identify SNPs
  - RNA-seq
  - ChIP-seq
  - Etc.

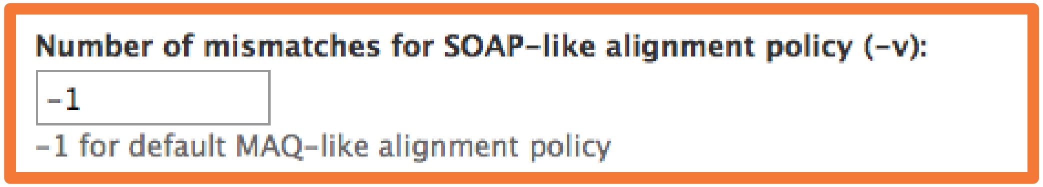


Bowtie (Langmead *et al*. 2009)

- Pre-built reference genome index
  - Burrows-Wheeler transform
  - Index needs to be computed prior to mapping
    - Either build your own: bowtie-build
    - Or ask for index to be installed for you
- Important parameters
  - -v vs. –n
    - Two mapping modes

Bowtie (Langmead *et al.* 2009)

- Mapping mode
  - -v: map reads that have less than *v* mismatches
    - Ignores quality scores
    - -v can be 0-3



Number of mismatches for SOAP-like alignment policy (-v):

-1

-1 for default MAQ-like alignment policy

```
Reference  ATGCGTAGTACGTCAACGTGTCACGTGACAGACAGT
     Read     CGAAGTACGACAACGGGTCAC
```

If number of mismatches <= v, read maps

Bowtie (Langmead *et al*. 2009)

- Mapping mode
  - -n: map using quality scores
    - -n: Mismatches in seed (0-3), ignores quality
    - -l: seed length (default 28bp)
    - -e: max quality score of mismatches across read (default 70)
      - Quality scores range from 0-40

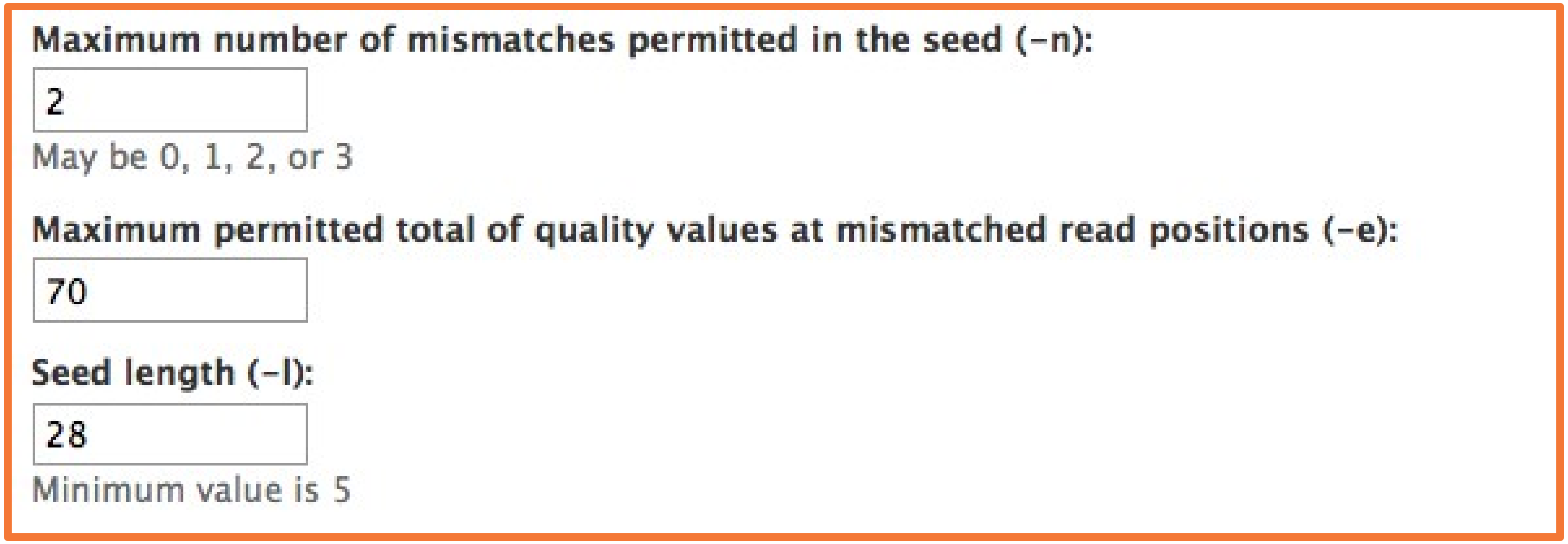
```
Reference  ATGCGTAGTACGTCAACGTGTCACGTGACAGACAGT
     Read     CGAAGTACGACAACGGGTCAC
```

Seed: -l 7
-n 1

If sum of quality scores on the mismatches is <=e, read maps here, otherwise not

Bowtie (Langmead *et al.* 2009)

- Mapping mode
  - -n: map using quality scores
    - -n: Mismatches in seed (0-3), ignores quality
    - -l: seed length (default 28bp)
    - -e: max quality score of mismatches across read (default 70)



Maximum number of mismatches permitted in the seed (-n):

2

May be 0, 1, 2, or 3

Maximum permitted total of quality values at mismatched read positions (-e):

70

Seed length (-l):

28

Minimum value is 5

Bowtie (Langmead *et al*. 2009)

- Dealing with multiple mappings
  - -k: report up to *k* good alignments per read (1)
  - -a: report all alignments for a read (slow!)
  - -m: don’t report if more than m alignments exist
  - -M: like –m, but report 1 random alignment
  - --best: guarantees alignment is in best stratum
  - --strata: don’t report suboptimal strata

Bowtie (Langmead *et al*. 2009)

- Keeping unmapped/mapped reads
  - --un <filename> unmapped reads
  - --al <filename> mapped reads
  - Can be helpful for downstream analyses

- Use –S for SAM output
  - Most likely will process output using SAM anyway

- -p: Bowtie is threaded, can run using multiple cores on ***one*** node
  - E.g.: nodes=1:ppn=8

Bowtie2 (Langmead & Salzberg 2012)

- Adds gapped read alignment (indels)
- Faster than Bowtie for reads longer than 50bp
- Supports local alignment
  - Can trim ends that don’t map
- Can map reads over Ns in reference
- No colorspace option

Bowtie2 (Langmead & Salzberg 2012)

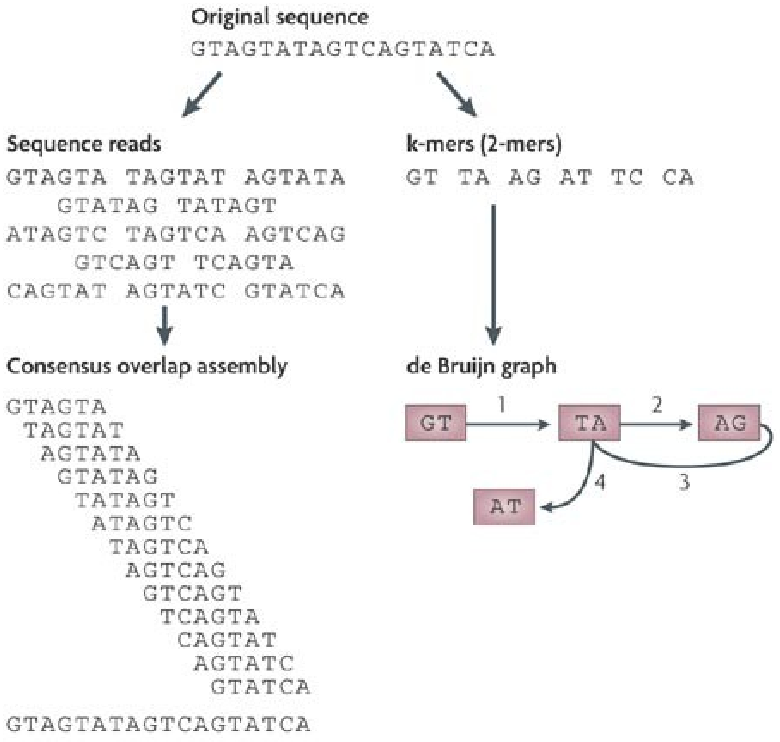
- Presets for both global and local
  - --very-fast(-local)
  - --fast(-local)
  - **--sensitive(-local) Defaults**
  - --very-sensitive(-local)

Other mapping applications

- BWA
- Lastz
- Maq
    - Bowtie is generally faster
- Mosaik
    - Handles gapped alignments relative to reference
- PerM
- SRMA

de Novo Assembly

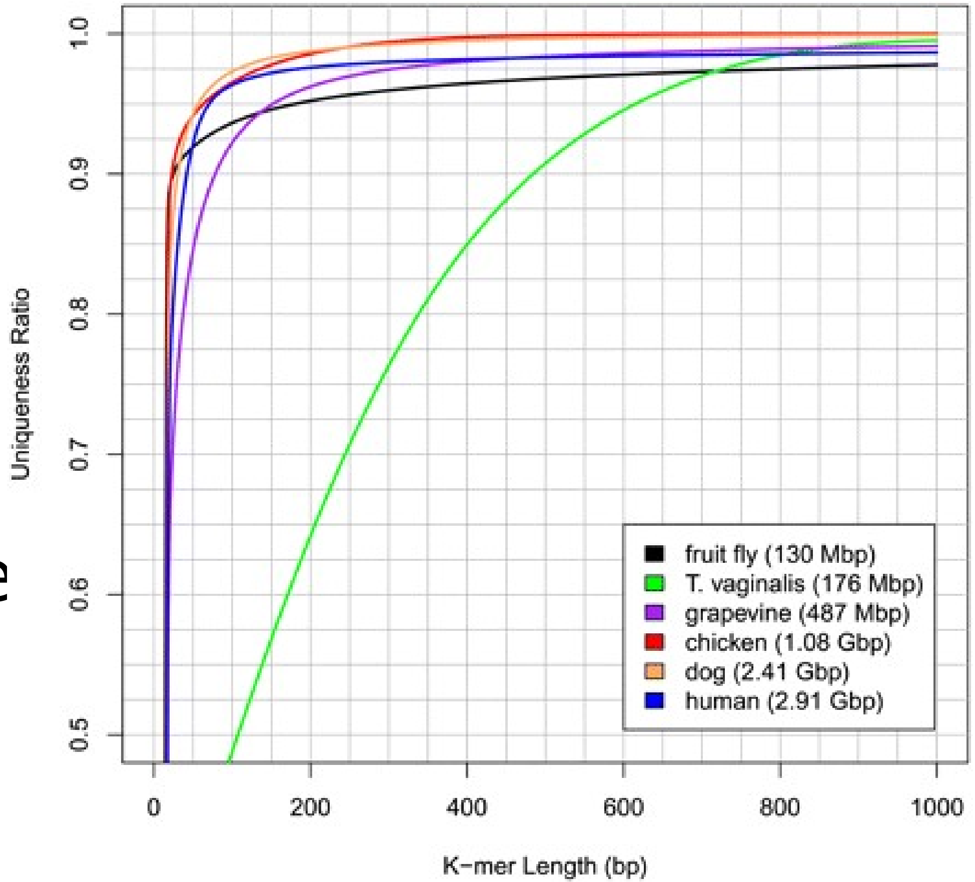
- No reference genome
- Assemble contigs from reads
  - Assemble scaffolds using paired-end data
- Most short-read assemblers are de Buijn graph-based



Nature Reviews | Microbiology

kmers

- A kmer is a sequence of length *k*
  - Longer kmer
    - More unique
    - Fewer reads/kmer
  - Shorter kmer
    - Less unique
    - More reads/kmer



Schatz *et al*, 2010

- The kmer you use does matter!
  - Try different kmers

Velvet (Zerbino & Birney 2008)

- Two stages
  - velveth
    - Creates the hash table of kmers
  - velvetg
    - Uses the de Bruijn graph to create contigs & scaffolds
- kmer is critical
  - Default maximum value is 31
  - If you need longer kmer, let us know

Velvet (Zerbino & Birney 2008)

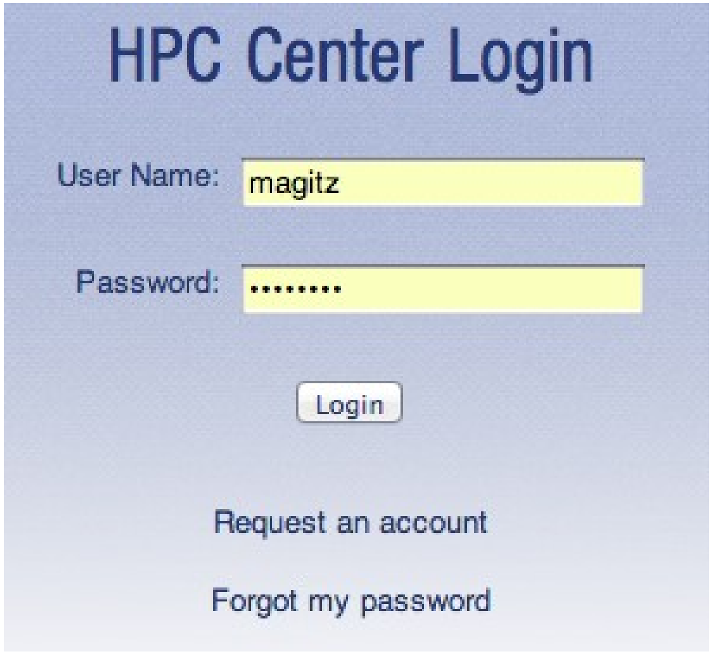
- Can use multiple types of sequencing inputs
  - Short, long
  - Paired, single
  - Different insert sizes
  - Reference
- A mix of library types is typically needed for de novo genome assembly
- Many helpful scripts distributed with Velvet
  - VelvetOptimiser—helps pick best kmer

Other de novo assembly applications

- Abyss
- ALLPATHS-LG
  - Has very specific requirements for library types and coverage
- Metavelvet
  - Modified version of Velvet for metagenomics
- Newbler
  - Provided by Roche (454), but can use Illumina data
- SOAPdenovo
- For RNA-seq
  - Oases (builds on after Velvet)
  - SOAPdenovo-TRANS
  - Trinity

Galaxy demo

http://galaxy.hpc.ufl.edu



HPC Center Login

User Name: magitz

Password: •••••••••

Questions?

Thank you!