Jon K. Lærdahl,
Structural Bioinformatics

Genome browsers

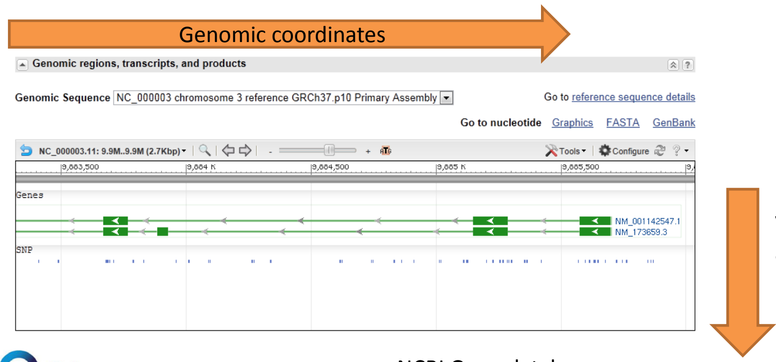
Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

Genome browsers

- Graphical interface for genomic data
- Shows information from biological databases mapped onto genomic sequence



Genomic coordinates

Genomic regions, transcripts, and products

Genomic Sequence NC_000003 chromosome 3 reference GRCh37.p10 Primary Assembly

Go to reference sequence details

Go to nucleotide Graphics FASTA GenBank

Various annotations = “tracks”

NCBI Gene database

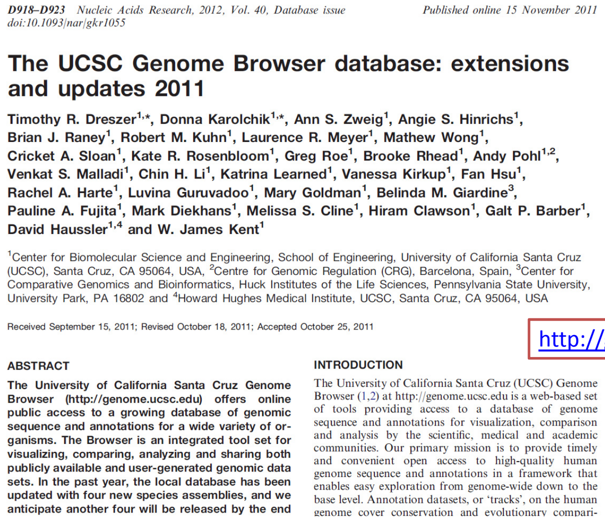
Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

UCSC Genome Browser

- Developed and maintained at the University of California, Santa Cruz (UCSC)
- Interactive website
- Access to genome sequence data from
  - Human genome
    - Latest assembly (GRCh37), but also earlier versions
  - Mouse, rat, and approx. 40 other mammals
  - Chicken, turkey, reptiles, frogs, and fish
  - Insects, nematodes, *S. cerevisiae* and more

UCSC Genome Browser



D918–D923 Nucleic Acids Research, 2012, Vol. 40, Database issue *Published online 15 November 2011*
doi:10.1093/nar/gkr1055

The UCSC Genome Browser database: extensions and updates 2011

Timothy R. Dreszer[1,*], Donna Karolchik[1,*], Ann S. Zweig[1], Angie S. Hinrichs[1], Brian J. Raney[1], Robert M. Kuhn[1], Laurence R. Meyer[1], Mathew Wong[1], Cricket A. Sloan[1], Kate R. Rosenbloom[1], Greg Roe[1], Brooke Rhead[1], Andy Pohl[1,2], Venkat S. Malladi[1], Chin H. Li[1], Katrina Learned[1], Vanessa Kirkup[1], Fan Hsu[1], Rachel A. Harte[1], Luvina Guruvadoo[1], Mary Goldman[1], Belinda M. Giardine[3], Pauline A. Fujita[1], Mark Diekhans[1], Melissa S. Cline[1], Hiram Clawson[1], Galt P. Barber[1], David Haussler[1,4] and W. James Kent[1]

[1]Center for Biomolecular Science and Engineering, School of Engineering, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA, [2]Centre for Genomic Regulation (CRG), Barcelona, Spain, [3]Center for Comparative Genomics and Bioinformatics, Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802 and [4]Howard Hughes Medical Institute, UCSC, Santa Cruz, CA 95064, USA

Received September 15, 2011; Revised October 18, 2011; Accepted October 25, 2011

ABSTRACT

The University of California Santa Cruz Genome Browser (http://genome.ucsc.edu) offers online public access to a growing database of genomic sequence and annotations for a wide variety of organisms. The Browser is an integrated tool set for visualizing, comparing, analyzing and sharing both publicly available and user-generated genomic data sets. In the past year, the local database has been updated with four new species assemblies, and we anticipate another four will be released by the end

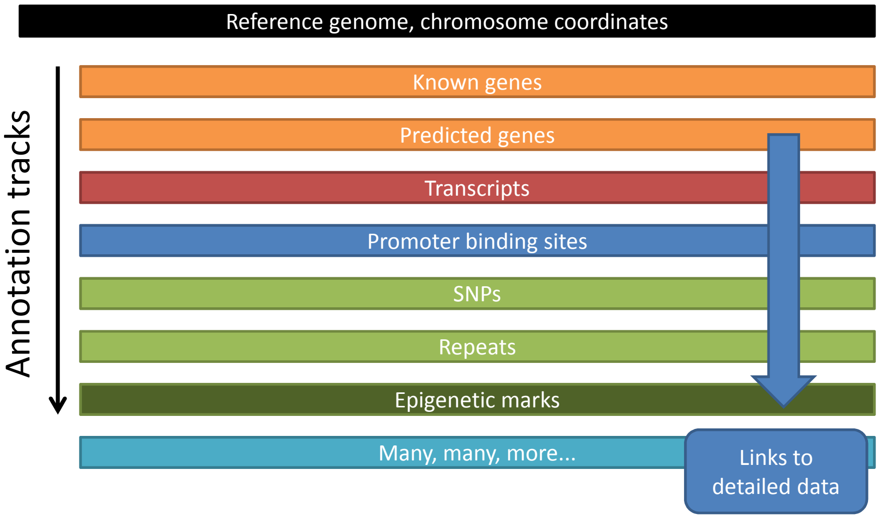
INTRODUCTION

The University of California Santa Cruz (UCSC) Genome Browser (1,2) at http://genome.ucsc.edu is a web-based set of tools providing access to a database of genome sequence and annotations for visualization, comparison and analysis by the scientific, medical and academic communities. Our primary mission is to provide timely and convenient open access to high-quality human genome sequence and annotations in a framework that enables easy exploration from genome-wide down to the base level. Annotation datasets, or 'tracks', on the human genome cover conservation and evolutionary compari-

http://genome.ucsc.edu

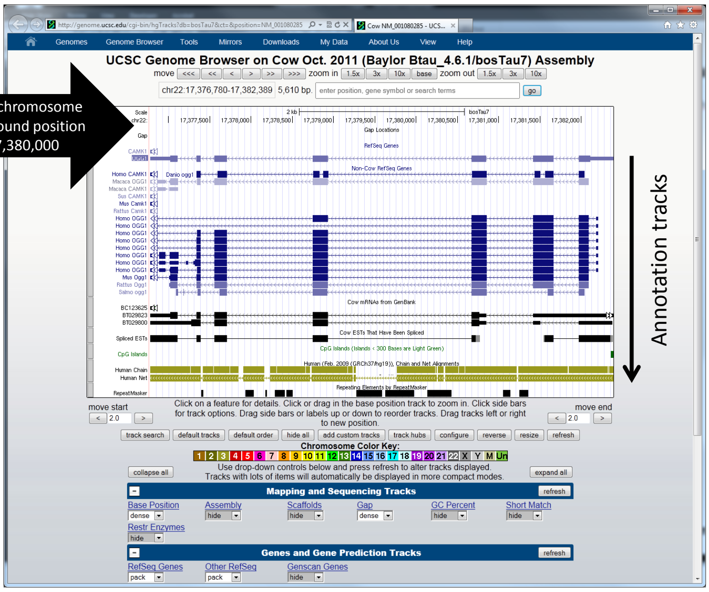
Jon K. Lærdahl,
Structural Bioinformatics

UCSC Genome Browser



Jon K. Lærdahl,
Structural Bioinformatics

UCSC Genome Browser



Cow, chromosome 22, around position 17,380,000

Annotation tracks

http://genome.ucsc.edu/cgi-bin/hgTracks?db=bosTau7&ct=&position=NM_001080285

Cow NM_001080285 - UCS...

UCSC Genome Browser on Cow Oct. 2011 (Baylor Btau_4.6.1/bosTau7) Assembly

chr22:17,376,780-17,382,389 5,610 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Chromosome Color Key:

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

Base Position: dense
Assembly: hide
Scaffolds: hide
Gap: dense
GC Percent: hide
Short Match: hide
Restr Enzymes: hide

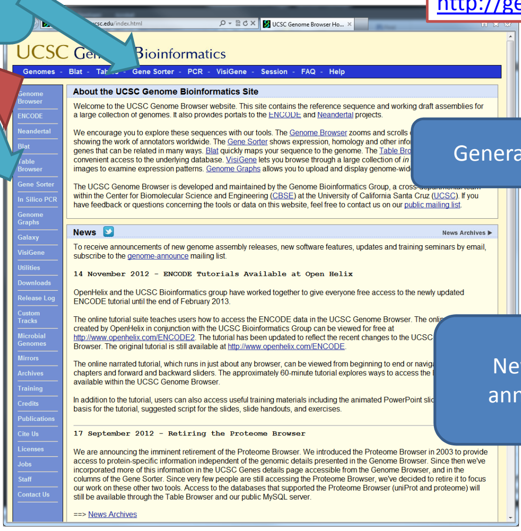
Genes and Gene Prediction Tracks

RefSeq Genes: pack
Other RefSeq: pack
Genscan Genes: hide

UCSC Genome Browser

Jon K. Lærdahl,
Structural Bioinformatics

http://genome.ucsc.edu



About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other info genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Bro convenient access to the underlying database. VisiGene lets you browse through a large collection of *in* images to examine expression patterns. Genome Graphs allows you to upload and display genome-wid

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

News News Archives ▶

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

`14 November 2012 - ENCODE Tutorials Available at Open Helix`

OpenHelix and the UCSC Bioinformatics group have worked together to give everyone free access to the newly updated ENCODE tutorial until the end of February 2013.

The online tutorial suite teaches users how to access the ENCODE data in the UCSC Genome Browser. The onli created by OpenHelix in conjunction with the UCSC Bioinformatics Group can be viewed for free at http://www.openhelix.com/ENCODE2. The tutorial has been updated to reflect the recent changes to the UCSC Browser. The original tutorial is still available at http://www.openhelix.com/ENCODE.

The online narrated tutorial, which runs in just about any browser, can be viewed from beginning to end or naviga chapters and forward and backward sliders. The approximately 60-minute tutorial explores ways to access the available within the UCSC Genome Browser.

In addition to the tutorial, users can also access useful training materials including the animated PowerPoint slic basis for the tutorial, suggested script for the slides, slide handouts, and exercises.

`17 September 2012 - Retiring the Proteome Browser`

We are announcing the imminent retirement of the Proteome Browser. We introduced the Proteome Browser in 2003 to provide access to protein-specific information independent of the genomic details presented in the Genome Browser. Since then we've incorporated more of this information in the UCSC Genes details page accessible from the Genome Browser, and in the columns of the Gene Sorter. Since very few people are still accessing the Proteome Browser, we've decided to retire it to focus our work on these other two tools. Access to the databases that supported the Proteome Browser (uniProt and proteome) will still be available through the Table Browser and our public MySQL server.

==> News Archives

Access to the databases and tools

Start here

General information

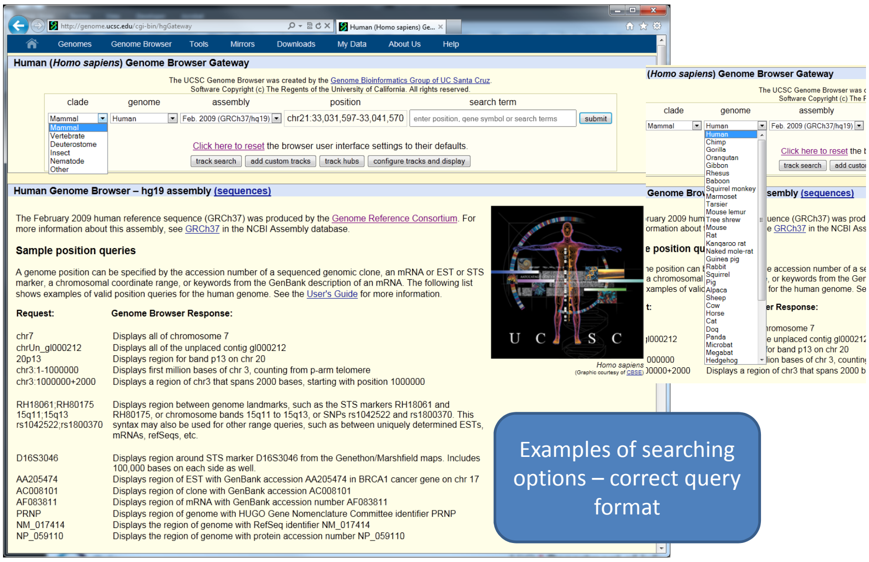
News, updates, announcements

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

UCSC Genome Browser

Jon K. Lærdahl,
Structural Bioinformatics



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position	search term	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	enter position, gene symbol or search terms	submit

Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks and display

Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175 15q11;15q13 rs1042522;rs1800370	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110

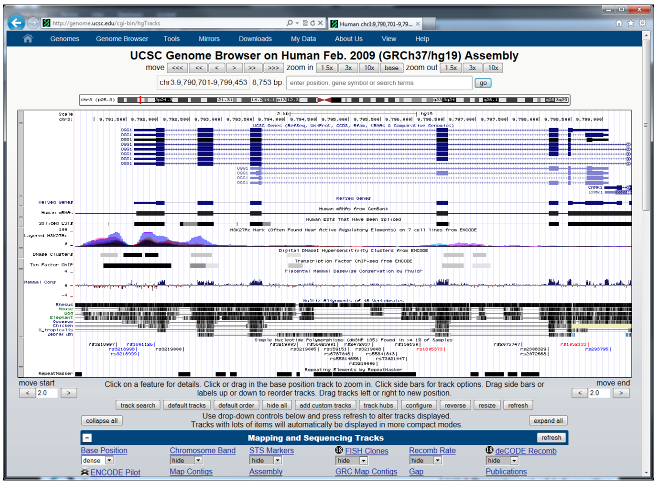
Examples of searching options – correct query format

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

UCSC Genome Browser brief demo



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr3:9,790,701-9,799,453 8,753 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

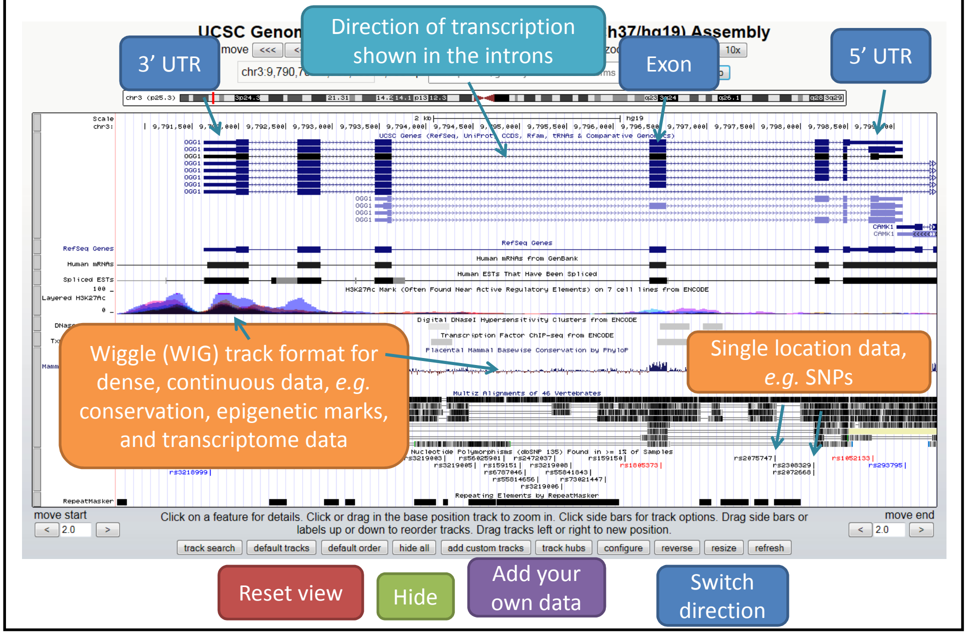
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

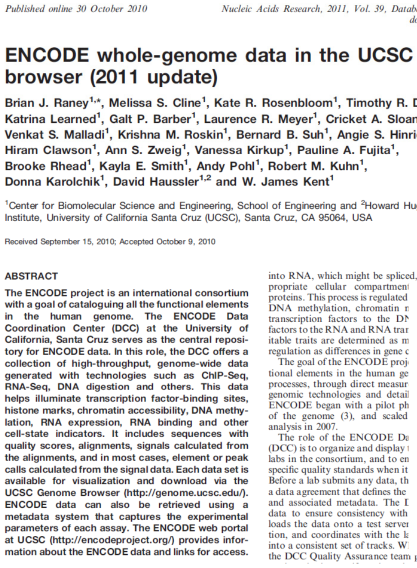
Different kinds of data



Direction of transcription shown in the introns
3' UTR
Exon
5' UTR
Wiggle (WIG) track format for dense, continuous data, *e.g.* conservation, epigenetic marks, and transcriptome data
Single location data, *e.g.* SNPs
Reset view
Hide
Add your own data
Switch direction
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
OGG1
CAMK1
RefSeq Genes
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP 135) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
move start
move end
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

ENCODE data in UCSC



Published online 30 October 2010 Nucleic Acids Research, 2011, Vol. 39, Database issue D871–D875
doi:10.1093/nar/gkq1017

ENCODE whole-genome data in the UCSC genome browser (2011 update)

Brian J. Raney[1,*], Melissa S. Cline[1], Kate R. Rosenbloom[1], Timothy R. Dreszer[1], Katrina Learned[1], Galt P. Barber[1], Laurence R. Meyer[1], Cricket A. Sloan[1], Venkat S. Malladi[1], Krishna M. Roskin[1], Bernard B. Suh[1], Angie S. Hinrichs[1], Hiram Clawson[1], Ann S. Zweig[1], Vanessa Kirkup[1], Pauline A. Fujita[1], Brooke Rhead[1], Kayla E. Smith[1], Andy Pohl[1], Robert M. Kuhn[1], Donna Karolchik[1], David Haussler[1,2] and W. James Kent[1]

[1]Center for Biomolecular Science and Engineering, School of Engineering and [2]Howard Hughes Medical Institute, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA

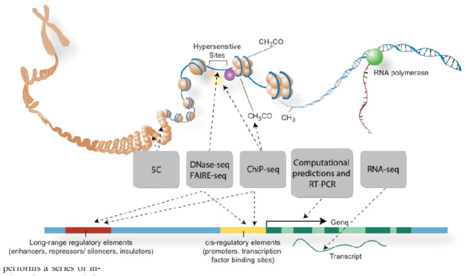
Received September 15, 2010; Accepted October 9, 2010

ABSTRACT

The ENCODE project is an international consortium with a goal of cataloguing all the functional elements in the human genome. The ENCODE Data Coordination Center (DCC) at the University of California, Santa Cruz serves as the central repository for ENCODE data. In this role, the DCC offers a collection of high-throughput, genome-wide data generated with technologies such as ChIP-Seq, RNA-Seq, DNA digestion and others. This data helps illuminate transcription factor-binding sites, histone marks, chromatin accessibility, DNA methylation, RNA expression, RNA binding and other cell-state indicators. It includes sequences with quality scores, alignments, signals calculated from the alignments, and in most cases, element or peak calls calculated from the signal data. Each data set is available for visualization and download via the UCSC Genome Browser (http://genome.ucsc.edu/). ENCODE data can also be retrieved using a metadata system that captures the experimental parameters of each assay. The ENCODE web portal at UCSC (http://encodeproject.org/) provides information about the ENCODE data and links for access.

into RNA, which might be spliced, transported to an ap-
propriate cellular compartment
proteins. This process is regulated
DNA methylation, chromatin n
transcription factors to the DN
factors to the RNA and RNA tran
itable traits are determined as m
regulation as differences in gene c
The goal of the ENCODE proj
tional elements in the human ge
processes, through direct measur
genomic technologies and detail
ENCODE began with a pilot ph
of the genome (3), and scaled
analysis in 2007.
The role of the ENCODE Da
(DCC) is to organize and display t
labs in the consortium, and to en
specific quality standards when it
Before a lab submits any data, th
a data agreement that defines the
and associated metadata. The D
data to ensure consistency with
loads the data onto a test server
tion, and coordinates with the la
into a consistent set of tracks. W
the DCC Quality Assurance team performs a series of in-

http://genome.ucsc.edu/ENCODE/aboutScaleup.html

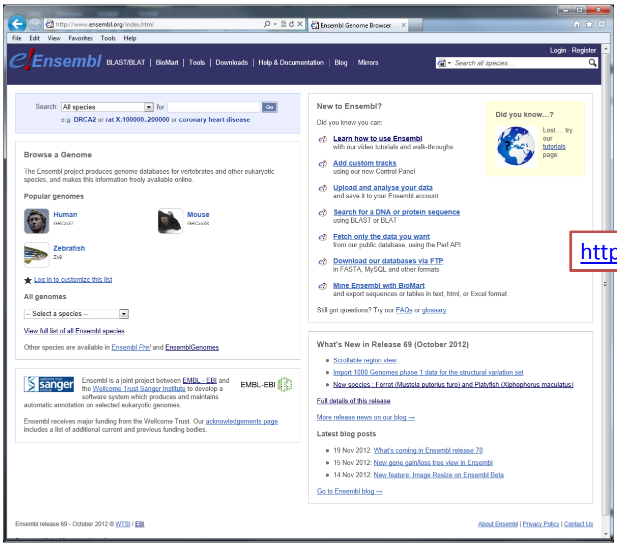

Hypersensitive Sites
CH3CO
RNA polymerase
CH3CO
CH3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

Ensembl Genome Browser

- Joint project between EMBL-EBI and the Wellcome Trust Sanger Institute
- Central resource for studying genomes of vertebrates
  - Mainly chordates, but some few extra (*e.g. C. elegans and S. cerevisiae*)
  - Updated several times a year with new genome assemblies and new species
  - Annotations of genomes (*e.g.* genes and their splice variant, SNPs) added by the Ensembl pipeline
  - Automatic gene prediction (with or without experimental evidence) & some curator input

Jon K. Lærdahl,
Structural Bioinformatics

Ensembl Genome Browser


http://www.ensembl.org/index.html
Ensembl Genome Browser
File Edit View Favorites Tools Help
e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login · Register
Search all species...

Search: All species for Go
e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh37
Mouse GRCm38
Zebrafish Zv9

★ Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

New to Ensembl?

Did you know you can:

- **Learn how to use Ensembl** with our video tutorials and walk-throughs
- **Add custom tracks** using our new Control Panel
- **Upload and analyse your data** and save it to your Ensembl account
- **Search for a DNA or protein sequence** using BLAST or BLAT
- **Fetch only the data you want** from our public database, using the Perl API
- **Download our databases via FTP** in FASTA, MySQL and other formats
- **Mine Ensembl with BioMart** and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs or glossary

Did you know...?

Lost ... try our tutorials page.

What's New in Release 69 (October 2012)

- Scrollable region view
- Import 1000 Genomes phase 1 data for the structural variation set
- New species : Ferret (Mustela putorius furo) and Platyfish (Xiphophorus maculatus)

Full details of this release

More release news on our blog →

Latest blog posts

- 19 Nov 2012: What's coming in Ensembl release 70
- 15 Nov 2012: New gene gain/loss tree view in Ensembl
- 14 Nov 2012: New feature: Image Resize on Ensembl Beta

Go to Ensembl blog →

Ensembl release 69 - October 2012 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

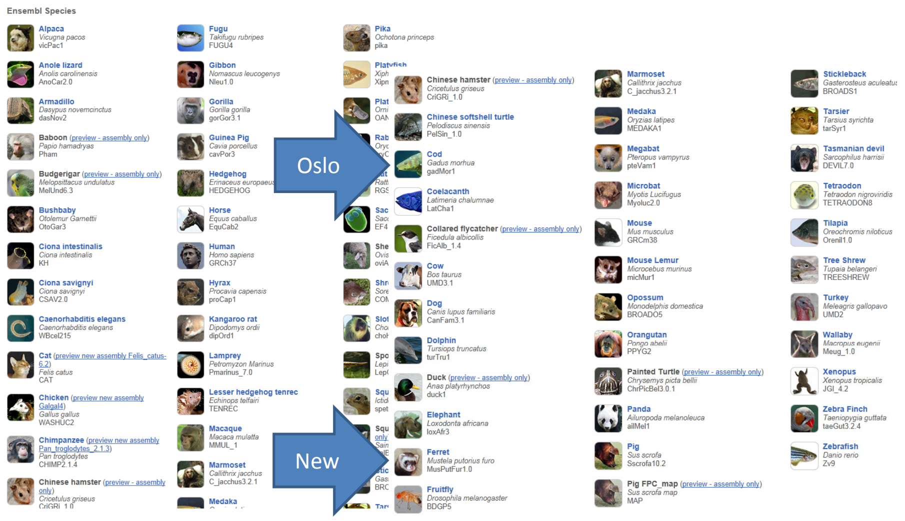
http://www.ensembl.org

Oslo universitetssykehus

UiO Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

Ensembl Genome Browser



Ensembl Species

Alpaca
Vicugna pacos
vicPac1

Anole lizard
Anolis carolinensis
AnoCar2.0

Armadillo
Dasypus novemcinctus
dasNov2

Baboon (preview - assembly only)
Papio hamadryas
Pham

Budgerigar (preview - assembly only)
Melopsittacus undulatus
MelUnd6.3

Bushbaby
Otolemur Garnettii
OtoGar3

Ciona intestinalis
Ciona intestinalis
KH

Ciona savignyi
Ciona savignyi
CSAV2.0

Caenorhabditis elegans
Caenorhabditis elegans
WBcel215

Cat (preview new assembly Felis_catus-6.2)
Felis catus
CAT

Chicken (preview new assembly Galgal4)
Gallus gallus
WASHUC2

Chimpanzee (preview new assembly Pan_troglodytes_2.1.3)
Pan troglodytes
CHIMP2.1.4

Chinese hamster (preview - assembly only)
Cricetulus griseus
CriGRi_1.0

Fugu
Takifugu rubripes
FUGU4

Gibbon
Nomascus leucogenys
Nleu1.0

Gorilla
Gorilla gorilla
gorGor3.1

Guinea Pig
Cavia porcellus
cavPor3

Hedgehog
Erinaceus europaeus
HEDGEHOG

Horse
Equus caballus
EquCab2

Human
Homo sapiens
GRCh37

Hyrax
Procavia capensis
proCap1

Kangaroo rat
Dipodomys ordii
dipOrd1

Lamprey
Petromyzon Marinus
Pmarinus_7.0

Lesser hedgehog tenrec
Echinops telfairi
TENREC

Macaque
Macaca mulatta
MMUL_1

Marmoset
Callithrix jacchus
C_jacchus3.2.1

Medaka

Pika
Ochotona princeps
pika

Platyfish
Xiph
Xipn

Plat
Orni
OAN

Rab
Oryc

Rat
Ratt
RGS

Sac
Sacc
EF4

She
Ovis
oviA

Shr
Sore
COM

Slot
Cho
chol

Spo
Lepi
Lep

Squ
Ictid
spet

Squ
only
Sain

Stic
Gas
BRC

Tar

Chinese hamster (preview - assembly only)
Cricetulus griseus
CriGRi_1.0

Chinese softshell turtle
Pelodiscus sinensis
PelSin_1.0

Cod
Gadus morhua
gadMor1

Coelacanth
Latimeria chalumnae
LatCha1

Collared flycatcher (preview - assembly only)
Ficedula albicollis
FicAlb_1.4

Cow
Bos taurus
UMD3.1

Dog
Canis lupus familiaris
CanFam3.1

Dolphin
Tursiops truncatus
turTru1

Duck (preview - assembly only)
Anas platyrhynchos
duck1

Elephant
Loxodonta africana
loxAfr3

Ferret
Mustela putorius furo
MusPutFur1.0

Fruitfly
Drosophila melanogaster
BDGP5

Marmoset
Callithrix jacchus
C_jacchus3.2.1

Medaka
Oryzias latipes
MEDAKA1

Megabat
Pteropus vampyrus
pteVam1

Microbat
Myotis Lucifugus
Myoluc2.0

Mouse
Mus musculus
GRCm38

Mouse Lemur
Microcebus murinus
micMur1

Opossum
Monodelphis domestica
BROADO5

Orangutan
Pongo abelii
PPYG2

Painted Turtle (preview - assembly only)
Chrysemys picta bellii
ChrPicBel3.0.1

Panda
Ailuropoda melanoleuca
ailMel1

Pig
Sus scrofa
Sscrofa10.2

Pig FPC_map (preview - assembly only)
Sus scrofa map
MAP

Stickleback
Gasterosteus aculeatus
BROADS1

Tarsier
Tarsius syrichta
tarSyr1

Tasmanian devil
Sarcophilus harrisii
DEVIL7.0

Tetraodon
Tetraodon nigroviridis
TETRAODON8

Tilapia
Oreochromis niloticus
Orenil1.0

Tree Shrew
Tupaia belangeri
TREESHREW

Turkey
Meleagris gallopavo
UMD2

Wallaby
Macropus eugenii
Meug_1.0

Xenopus
Xenopus tropicalis
JGI_4.2

Zebra Finch
Taeniopygia guttata
taeGut3.2.4

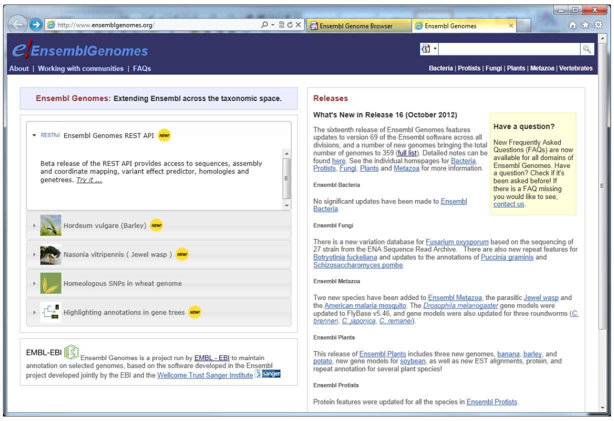
Zebrafish
Danio rerio
Zv9

Oslo universitetssykehus

UiO Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

EnsemblGenomes


http://www.ensemblgenomes.org/
Ensembl Genome Browser
Ensembl Genomes
EnsemblGenomes
About | Working with communities | FAQs
Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates

Ensembl Genomes: Extending Ensembl across the taxonomic space.

RESTful Ensembl Genomes REST API NEW!

Beta release of the REST API provides access to sequences, assembly and coordinate mapping, variant effect predictor, homologies and genetrees. *Try it ...*

Hordeum vulgare (Barley) NEW!

Nasonia vitripennis (Jewel wasp) NEW!

Homeologous SNPs in wheat genome

Highlighting annotations in gene trees NEW!

EMBL-EBI Ensembl Genomes is a project run by EMBL - EBI to maintain annotation on selected genomes, based on the software developed in the Ensembl project developed jointly by the EBI and the Wellcome Trust Sanger Institute sanger

Releases

What's New in Release 16 (October 2012)

The sixteenth release of Ensembl Genomes features updates to version 69 of the Ensembl software across all divisions, and a number of new genomes bringing the total number of genomes to 359 (full list). Detailed notes can be found here. See the individual homepages for Bacteria, Protists, Fungi, Plants and Metazoa for more information.

Have a question?

New Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing you would like to see, contact us.

Ensembl Bacteria

No significant updates have been made to Ensembl Bacteria.

Ensembl Fungi

There is a new variation database for Fusarium oxysporum based on the sequencing of 27 strain from the ENA Sequence Read Archive. There are also new repeat features for Botryotinia fuckeliana and updates to the annotations of Puccinia graminis and Schizosaccharomyces pombe.

Ensembl Metazoa

Two new species have been added to Ensembl Metazoa, the parasitic Jewel wasp and the American malaria mosquito. The *Drosophila melanogaster* gene models were updated to FlyBase v5.46, and gene models were also updated for three roundworms (*C. brenneri*, *C. japonica*, *C. remanei*).

Ensembl Plants

This release of Ensembl Plants includes three new genomes, banana, barley, and potato, new gene models for soybean, as well as new EST alignments, protein, and repeat annotation for several plant species!

Ensembl Protists

Protein features were updated for all the species in Ensembl Protists.

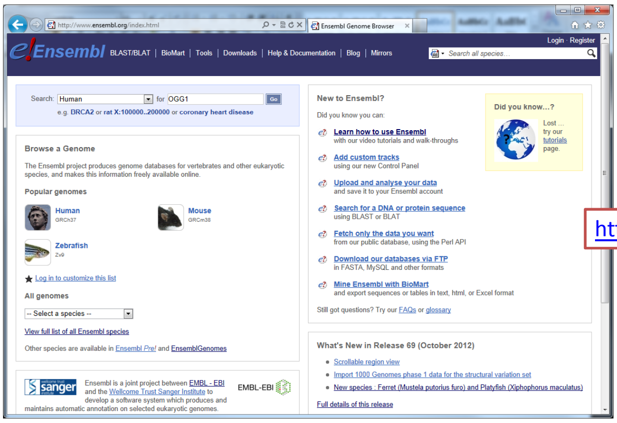
- Bacteria, protists, fungi, plants and other metazoa (359 genomes)

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

Ensembl Genome Browser


http://www.ensembl.org/index.html
Ensembl Genome Browser
Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login · Register
Search all species...

Search: Human for OGG1 Go

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh37

Mouse GRCm38

Zebrafish Zv9

★ Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

sanger Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. EMBL-EBI

New to Ensembl?

Did you know you can:

Learn how to use Ensembl
with our video tutorials and walk-throughs

Add custom tracks
using our new Control Panel

Upload and analyse your data
and save it to your Ensembl account

Search for a DNA or protein sequence
using BLAST or BLAT

Fetch only the data you want
from our public database, using the Perl API

Download our databases via FTP
in FASTA, MySQL and other formats

Mine Ensembl with BioMart
and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs or glossary

Did you know...?

Lost ... try our tutorials page.

What's New in Release 69 (October 2012)

- Scrollable region view
- Import 1000 Genomes phase 1 data for the structural variation set
- New species : Ferret (Mustela putorius furo) and Platyfish (Xiphophorus maculatus)

Full details of this release

http://www.ensembl.org

Explore!

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

Now something different!

Galaxy

Not a genome browser!

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

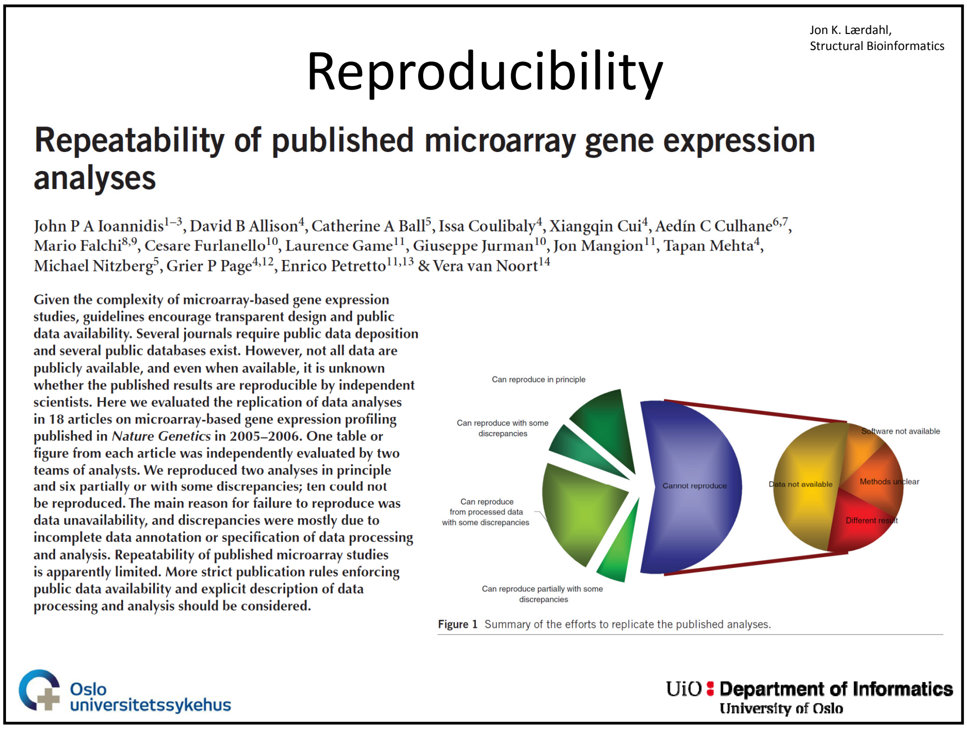
Jon K. Lærdahl,
Structural Bioinformatics

Galaxy

- Galaxy is a platform (open, web-based) for computational medical projects and bioinformatics
  - Accessible: Not necessary to know programming, Unix, or how to install programs
  - Reproducible: You can build and store complete workflows, pipelines, and the full computational analysis
  - Transparent: Users can publish and share whole worksflows
- A bioinformatics workflow management system

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo



Jon K. Lærdahl,
Structural Bioinformatics

Reproducibility

Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

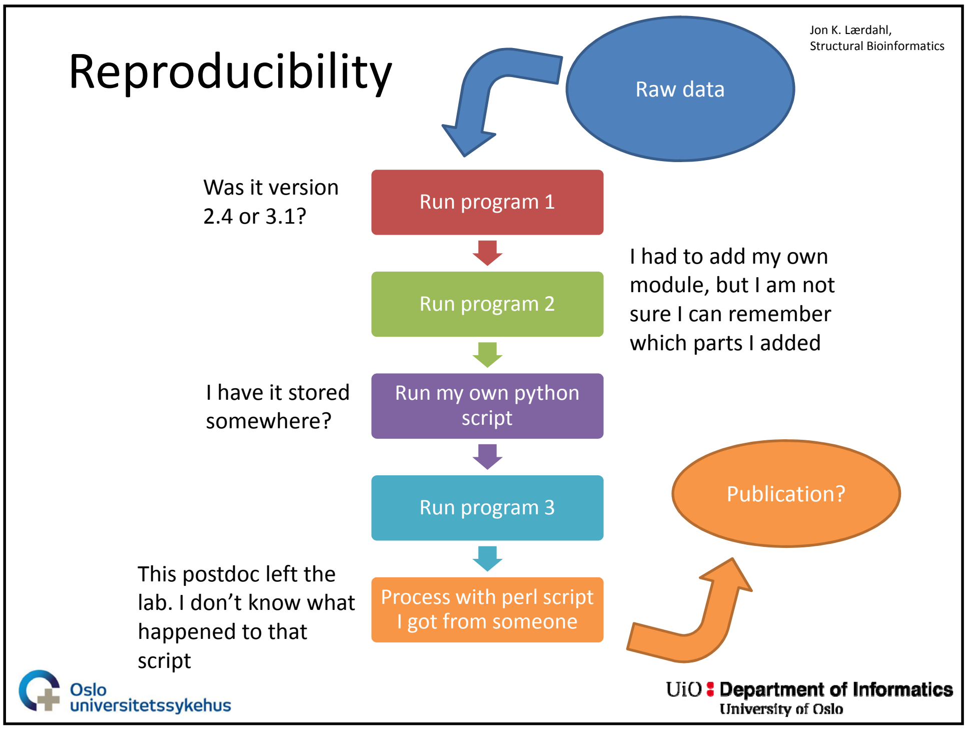
Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data are publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis. Repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered.

Can reproduce in principle
Can reproduce with some discrepancies
Can reproduce from processed data with some discrepancies
Can reproduce partially with some discrepancies
Cannot reproduce
Data not available
Software not available
Methods unclear
Different result

Figure 1 Summary of the efforts to replicate the published analyses.

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo



Jon K. Lærdahl,
Structural Bioinformatics

Reproducibility

Was it version 2.4 or 3.1?

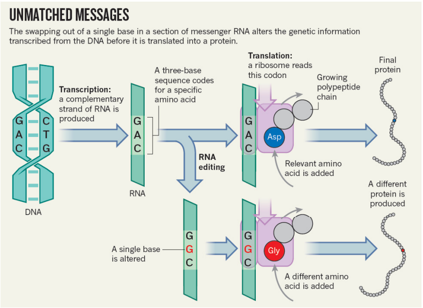
I had to add my own module, but I am not sure I can remember which parts I added

I have it stored somewhere?

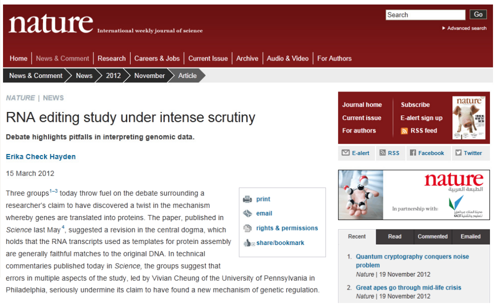
This postdoc left the lab. I don't know what happened to that script

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo


UNMATCHED MESSAGES
The swapping out of a single base in a section of messenger RNA alters the genetic information transcribed from the DNA before it is translated into a protein.
Transcription: a complementary strand of RNA is produced
A three-base sequence codes for a specific amino acid
Translation: a ribosome reads this codon
Growing polypeptide chain
Final protein
G A C
C T G
DNA
G A C
RNA
RNA editing
Asp
Relevant amino acid is added
A different protein is produced
A single base is altered
G G C
Gly
A different amino acid is added

4 3 2 | NATURE | VOL 473 | 26 MAY 2011



NATURE | NEWS

RNA editing study under intense scrutiny

Debate highlights pitfalls in interpreting genomic data.

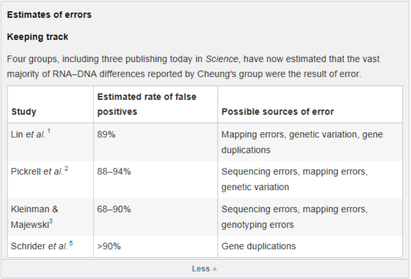
Erika Check Hayden

15 March 2012

Three groups[1–3] today throw fuel on the debate surrounding a researcher's claim to have discovered a twist in the mechanism whereby genes are translated into proteins. The paper, published in *Science* last May[4], suggested a revision in the central dogma, which holds that the RNA transcripts used as templates for protein assembly are generally faithful matches to the original DNA. In technical commentaries published today in *Science*, the groups suggest that errors in multiple aspects of the study, led by Vivian Cheung of the University of Pennsylvania in Philadelphia, seriously undermine its claim to have found a new mechanism of genetic regulation.

1. Quantum cryptography conquers noise problem
Nature | 19 November 2012
2. Great apes go through mid-life crisis
Nature | 19 November 2012

Li, M. *et al.* Science **33**, 53 (2011), July 2011:
- >10,000 sites in the human genome where an RNA sequence did not match the sequence of the DNA
- Evidence for a new mechanism of RNA editing?
- Central dogma is wrong in 10,000 places?



Estimates of errors

Keeping track

Four groups, including three publishing today in *Science*, have now estimated that the vast majority of RNA–DNA differences reported by Cheung's group were the result of error.

Study	Estimated rate of false positives	Possible sources of error
Lin *et al.*[1]	89%	Mapping errors, genetic variation, gene duplications
Pickrell *et al.*[2]	88–94%	Sequencing errors, mapping errors, genetic variation
Kleinman & Majewski[3]	68–90%	Sequencing errors, mapping errors, genotyping errors
Schrider *et al.*[5]	>90%	Gene duplications

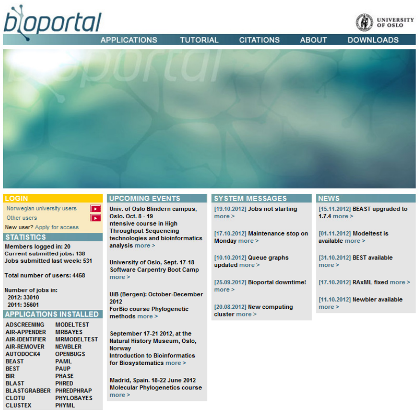
Less ▲

Galaxy

- Galaxy is a platform (open, web-based) for computational medical projects and bioinformatics
  - Accessible: Not necessary to know programming, Unix, or how to install programs
  - ***Reproducible: You can build and store complete workflows, pipelines, and the full computational analysis***
  - ***Transparent: Users can publish and share whole worksflows***
- By the way, Galaxy is written in Python...
- Developed by the labs of Anton Nekrutenko (Penn State University) and James Taylor (Emory University)

Galaxy

Jon K. Lærdahl,
Structural Bioinformatics



LOGIN

Norwegian university users

Other users

New user? Apply for access

STATISTICS

Members logged in: 20
Current submitted jobs: 138
Jobs submitted last week: 531

Total number of users: 4458

Number of jobs in:
2012: 33010
2011: 35601

APPLICATIONS INSTALLED

ADSCREENING
AIR-APPENDER
AIR-IDENTIFIER
AIR-REMOVER
AUTODOCK4
BEAST
BEST
BIR
BLAST
BLASTGRABBER
CLOTU
CLUSTEX
MODELTEST
MRBAYES
MRMODELTEST
NEWBLER
OPENBUGS
PAML
PAUP
PHASE
PHRED
PHREDPHRAP
PHYLOBAYES
PHYML

UPCOMING EVENTS

Univ. of Oslo Blindern campus, Oslo. Oct. 8 - 19
ntensive course in High Throughput Sequencing technologies and bioinformatics analysis more >

University of Oslo, Sept. 17-18
Software Carpentry Boot Camp more >

UiB (Bergen): October-December 2012
ForBio course Phylogenetic methods more >

September 17-21 2012, at the Natural History Museum, Oslo, Norway
Introduction to Bioinformatics for Biosystematics more >

Madrid, Spain. 18-22 June 2012
Molecular Phylogenetics course more >

SYSTEM MESSAGES

[19.10.2012] Jobs not starting more >

[17.10.2012] Maintenance stop on Monday more >

[10.10.2012] Queue graphs updated more >

[25.09.2012] Bioportal downtime! more >

[20.08.2012] New computing cluster more >

NEWS

[15.11.2012] BEAST upgraded to 1.7.4 more >

[01.11.2012] Modeltest is available more >

[31.10.2012] BEST available more >

[17.10.2012] RAxML fixed more >

[11.10.2012] Newbler available more >

- Bioportal -> Galaxy
- Galaxy at UiO submitting jobs to the Abel cluster

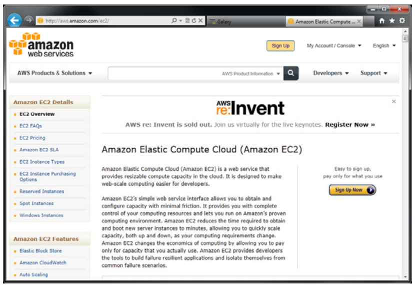
Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Galaxy

Jon K. Lærdahl,
Structural Bioinformatics

- Can be run on a free public server at Penn State
- You can install Galaxy on your own server or computer cluster (soon on Abel)
- You can run Galaxy in the cloud



AWS re: Invent is sold out. Join us virtually for the live keynotes. Register Now »

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

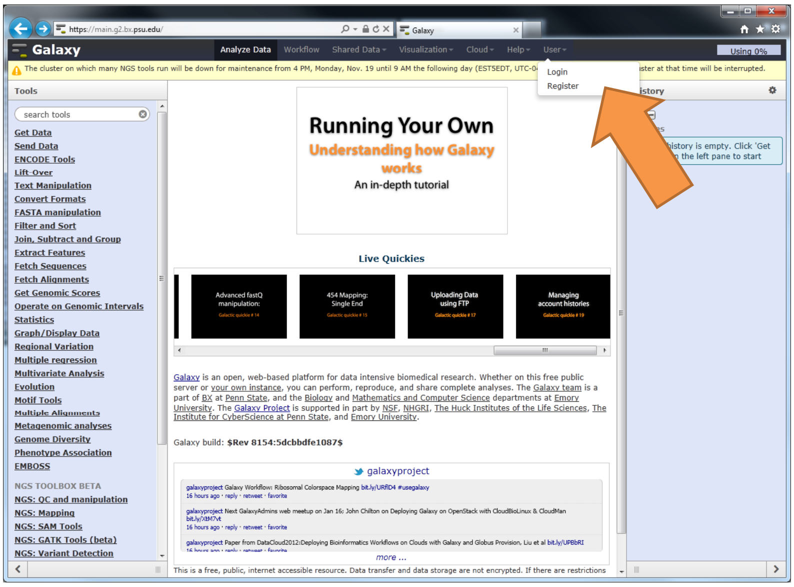
Easy to sign up, pay only for what you use

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Free public server:

Jon K. Lærdahl,
Structural Bioinformatics



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-04 …ster at that time will be interrupted.

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
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
Genome Diversity
Phenotype Association
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Variant Detection

…history is empty. Click 'Get … the left pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: **$Rev 8154:5dcbbdfe1087$**

galaxyproject Galaxy Workflow: Ribosomal Colorspace Mapping bit.ly/URflD4 #usegalaxy
16 hours ago · reply · retweet · favorite

galaxyproject Next GalaxyAdmins web meetup on Jan 16; John Chilton on Deploying Galaxy on OpenStack with CloudBioLinux & CloudMan bit.ly/XtM7vt
16 hours ago · reply · retweet · favorite

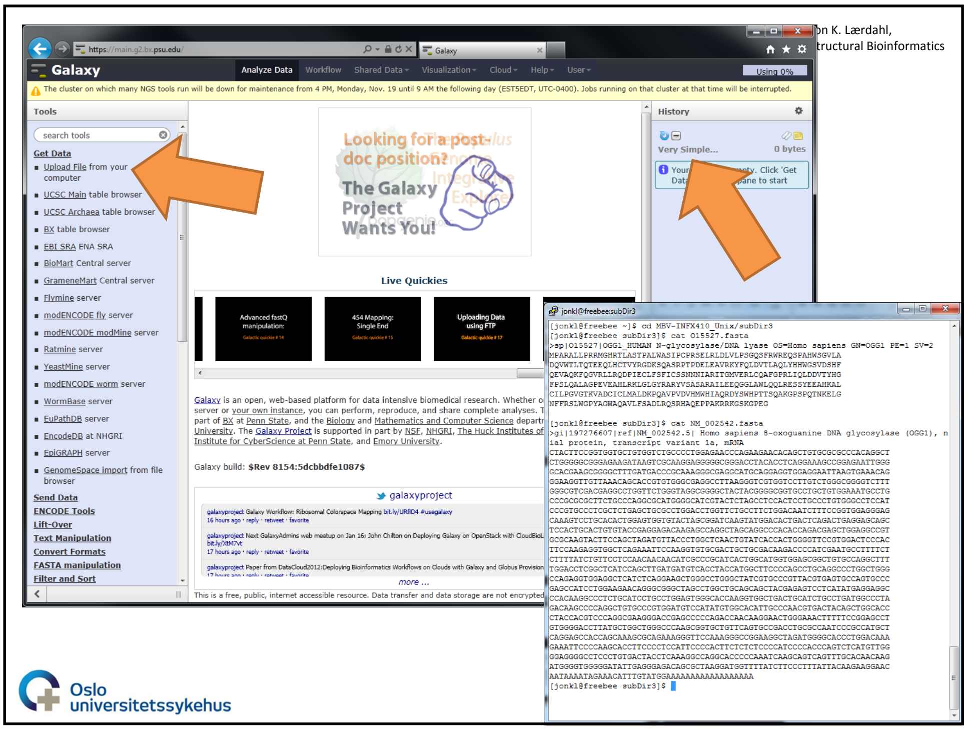
galaxyproject Paper from DataCloud2012:Deploying Bioinformatics Workflows on Clouds with Galaxy and Globus Provision, Liu et al bit.ly/UPBbRI
16 hours ago · reply · retweet · favorite

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Get Data
- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- GenomeSpace import from file browser

Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort

Your … Data … ty. Click 'Get … pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. Whether o… server or your own instance, you can perform, reproduce, and share complete analyses. T… part of BX at Penn State, and the Biology and Mathematics and Computer Science depart… University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of… Institute for CyberScience at Penn State, and Emory University.

Galaxy build: **$Rev 8154:5dcbbdfe1087$**

galaxyproject Galaxy Workflow: Ribosomal Colorspace Mapping bit.ly/URflD4 #usegalaxy
16 hours ago · reply · retweet · favorite

galaxyproject Next GalaxyAdmins web meetup on Jan 16; John Chilton on Deploying Galaxy on OpenStack with CloudBioL…
bit.ly/XtM7vt
17 hours ago · reply · retweet · favorite

galaxyproject Paper from DataCloud2012:Deploying Bioinformatics Workflows on Clouds with Galaxy and Globus Provision…

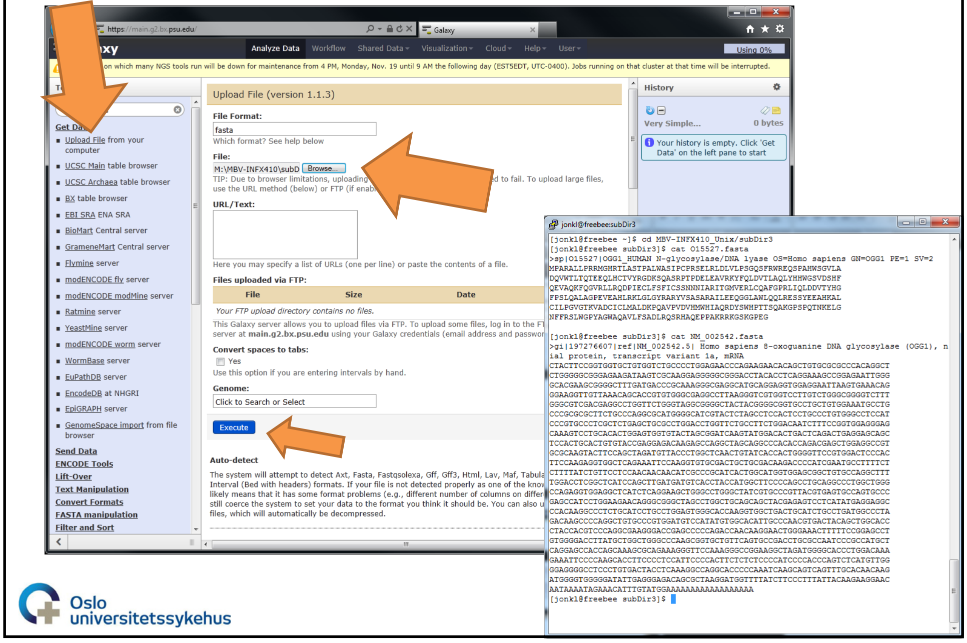
This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted…

```
[jonkl@freebee ~]$ cd MBV-INFX410_Unix/subDir3
[jonkl@freebee subDir3]$ cat O15527.fasta
>sp|O15527|OGG1_HUMAN N-glycosylase/DNA lyase OS=Homo sapiens GN=OGG1 PE=1 SV=2
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG

[jonkl@freebee subDir3]$ cat NM_002542.fasta
>gi|197276607|ref|NM_002542.5| Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), n
ial protein, transcript variant 1a, mRNA
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGAACCCAGAAGAACACAGCTGTGCGCGCCCACAGGCT
CTGGGGGCGGGAGAAGATAAGTCGCAAGGAGGGGGCGGGACCTACACCTCAGGAAAGCCGGAGAATTGGG
GCACGAAGCGGGGCTTTGATGACCCGCAAAGGGCGAGGCATGCAGGAGGTGGAGGAATTAAGTGAAACAG
GGAAGGTTGTTAAACAGCACCGTGTGGGCGAGGCCTTAAGGGTCGTGGTCCTTGTCTGGGCGGGGTCTTT
GGGCGTCGACGAGGCCTGGTTCTGGGTAGGCGGGGCTACTACGGGGCGGTGCCTGCTGTGGAAATGCCTG
CCCGCGCGCTTCTGCCCAGGCGCATGGGGCATCGTACTCTAGCCTCCACTCCTGCCCTGTGGGCCTCCAT
CCCGTGCCCTCGCTCTGAGCTGCGCCTGGACCTGGTTCTGCCTTCTGGACAATCTTTCCGGTGGAGGGAG
CAAAGTCCTGCACACTGGAGTGGTGTACTAGCGGATCAAGTATGGACACTGACTCAGACTGAGGAGCAGC
TCCACTGCACTGTGTACCGAGGAGACAAGAGCCAGGCTAGCAGGCCCACACCAGACGAGCTGGAGGCCGT
GCGCAAGTACTTCCAGCTAGATGTTACCCTGGCTCAACTGTATCACCACTGGGGTTCCGTGGACTCCCAC
TTCCAAGAGGTGGCTCAGAAATTCCAAGGTGTGCGACTGCTGCGACAAGACCCCATCGAATGCCTTTTCT
CTTTTATCTGTTCCTCCAACAACAACATCGCCCGCATCACTGGCATGGTGGAGCGGCTGTGCCAGGCTTT
TGGACCTCGGCTCATCCAGCTTGATGATGTCACCTACCATGGCTTCCCCAGCCTGCAGGCCCTGGCTGGG
CCAGAGGTGGAGGCTCATCTCAGGAAGCTGGGCCTGGGCTATCGTGCCCGTTACGTGAGTGCCAGTGCCC
GAGCCATCCTGGAAGAACAGGGCGGGCTAGCCTGGCTGCAGCAGCTACGAGAGTCCTCATATGAGGAGGC
CCACAAGGCCCTCTGCATCCTGCCTGGAGTGGGCACCAAGGTGGCTGACTGCATCTGCCTGATGGCCCTA
GACAAGCCCCAGGCTGTGCCCGTGGATGTCCATATGTGGCACATTGCCCAACGTGACTACAGCTGGCACC
CTACCACGTCCCAGGCGAAGGGACCGAGCCCCCAGACCAACAAGGAACTGGGAAACTTTTTCCGGAGCCT
GTGGGGACCTTATGCTGGCTGGGCCCAAGCGGTGCTGTTCAGTGCCGACCTGCGCCAATCCCGCCATGCT
CAGGAGCCACCAGCAAAGCGCAGAAAGGGTTCCAAAGGGCCGGAAGGCTAGATGGGGCACCCTGGACAAA
GAAATTCCCCAAGCACCTTCCCCTCCATTCCCCACTTCTCTCTCCCCATCCCCACCCAGTCTCATGTTGG
GGAGGGGCCTCCCTGTGACTACCTCAAAGGCCAGGCACCCCCAAATCAAGCAGTCAGTTTGCACAACAAG
ATGGGGTGGGGGATATTGAGGGAGACAGCGCTAAGGATGGTTTTATCTTCCCTTTATTACAAGAAGGAAC
AATAAAATAGAAACATTTGTATGGAAAAAAAAAAAAAAAAAA
[jonkl@freebee subDir3]$
```

Oslo universitetssykehus

Jon K. Lærdahl,
Structural Bioinformatics



on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Get Da

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- GenomeSpace import from file browser

Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort

Upload File (version 1.1.3)

File Format:
fasta
Which format? See help below

File:
M:\MBV-INFX410\subD Browse...
TIP: Due to browser limitations, uploading … d to fail. To upload large files, use the URL method (below) or FTP (if enab

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the F server at **main.g2.bx.psu.edu** using your Galaxy credentials (email address and passwor

Convert spaces to tabs:
☐ Yes
Use this option if you are entering intervals by hand.

Genome:
Click to Search or Select

Execute

Auto-detect

The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabula Interval (Bed with headers) formats. If your file is not detected properly as one of the know likely means that it has some format problems (e.g., different number of columns on differ still coerce the system to set your data to the format you think it should be. You can also u files, which will automatically be decompressed.

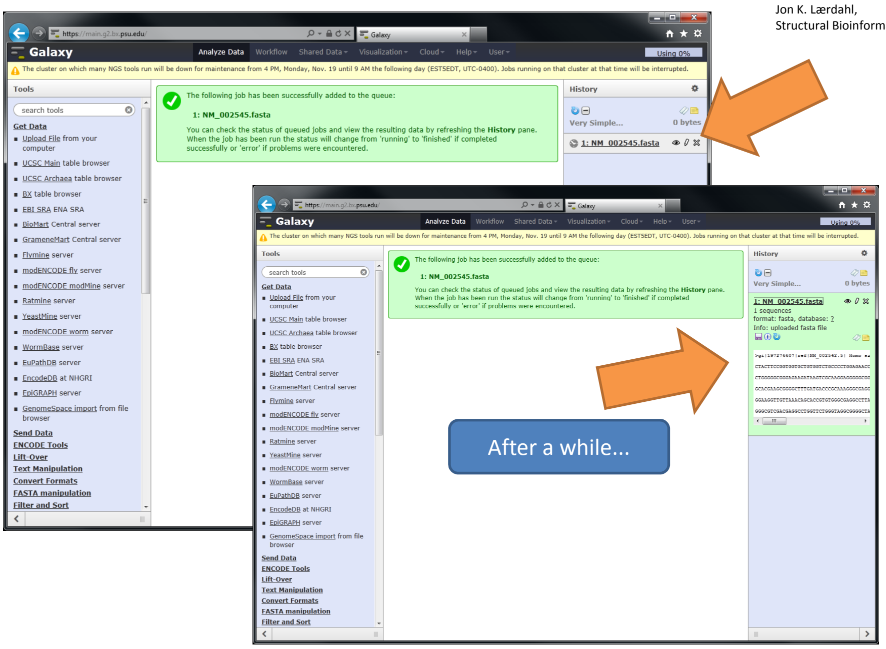
Your history is empty. Click 'Get Data' on the left pane to start

jonkl@freebee:subDir3

```
[jonkl@freebee ~]$ cd MBV-INFX410_Unix/subDir3
[jonkl@freebee subDir3]$ cat O15527.fasta
>sp|O15527|OGG1_HUMAN N-glycosylase/DNA lyase OS=Homo sapiens GN=OGG1 PE=1 SV=2
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG

[jonkl@freebee subDir3]$ cat NM_002542.fasta
>gi|197276607|ref|NM_002542.5| Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), n
ial protein, transcript variant 1a, mRNA
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGAACCCAGAAGAACACAGCTGTGCGCGCCCACAGGCT
CTGGGGGCGGGAGAAGATAAGTCGCAAGGAGGGGGCGGGACCTACACCTCAGGAAAGCCGGAGAATTGGG
GCACGAAGCGGGGCTTTGATGACCCGCAAAGGGCGAGGCATGCAGGAGGTGGAGGAATTAAGTGAAACAG
GGAAGGTTGTTAAACAGCACCGTGTGGGCGAGGCCTTAAGGGTCGTGGTCCTTGTCTGGGCGGGGTCTTT
GGGCGTCGACGAGGCCTGGTTCTGGGTAGGCGGGGCTACTACGGGGCGGTGCCTGCTGTGGAAATGCCTG
CCCGCGCGCTTCTGCCCAGGCGCATGGGGCATCGTACTCTAGCCTCCACTCCTGCCCTGTGGGCCTCCAT
CCCGTGCCCTCGCTCTGAGCTGCGCCTGGACCTGGTTCTGCCTTCTGGACAATCTTTCCGGTGGAGGGAG
CAAAGTCCTGCACACTGGAGTGGTGTACTAGCGGATCAAGTATGGACACTGACTCAGACTGAGGAGCAGC
TCCACTGCACTGTGTACCGAGGAGACAAGAGCCAGGCTAGCAGGCCCACACCAGACGAGCTGGAGGCCGT
GCGCAAGTACTTCCAGCTAGATGTTACCCTGGCTCAACTGTATCACCACTGGGGTTCCGTGGACTCCCAC
TTCCAAGAGGTGGCTCAGAAATTCCAAGGTGTGCGACTGCTGCGACAAGACCCCATCGAATGCCTTTTCT
CTTTTATCTGTTCCTCCAACAACAACATCGCCCGCATCACTGGCATGGTGGAGCGGCTGTGCCAGGCTTT
TGGACCTCGGCTCATCCAGCTTGATGATGTCACCTACCATGGCTTCCCCAGCCTGCAGGCCCTGGCTGGG
CCAGAGGTGGAGGCTCATCTCAGGAAGCTGGGCCTGGGCTATCGTGCCCGTTACGTGAGTGCCAGTGCCC
GAGCCATCCTGGAAGAACAGGGCGGGCTAGCCTGGCTGCAGCAGCTACGAGAGTCCTCATATGAGGAGGC
CCACAAGGCCCTCTGCATCCTGCCTGGAGTGGGCACCAAGGTGGCTGACTGCATCTGCCTGATGGCCCTA
GACAAGCCCCAGGCTGTGCCCGTGGATGTCCATATGTGGCACATTGCCCAACGTGACTACAGCTGGCACC
CTACCACGTCCCAGGCGAAGGGACCGAGCCCCCAGACCAACAAGGAACTGGGAAACTTTTTCCGGAGCCT
GTGGGGACCTTATGCTGGCTGGGCCCAAGCGGTGCTGTTCAGTGCCGACCTGCGCCAATCCCGCCATGCT
CAGGAGCCACCAGCAAAGCGCAGAAAGGGTTCCAAAGGGCCGGAAGGCTAGATGGGGCACCCTGGACAAA
GAAATTCCCCAAGCACCTTCCCCTCCATTCCCCACTTCTCTCTCCCCATCCCCACCCAGTCTCATGTTGG
GGAGGGGCCTCCCTGTGACTACCTCAAAGGCCAGGCACCCCCAAATCAAGCAGTCAGTTTGCACAACAAG
ATGGGGTGGGGGATATTGAGGGAGACAGCGCTAAGGATGGTTTTATCTTCCCTTTATTACAAGAAGGAAC
AATAAAATAGAAACATTTGTATGGAAAAAAAAAAAAAAAAAAA
[jonkl@freebee subDir3]$
```

Jon K. Lærdahl,
Structural Bioinformatics



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- GenomeSpace import from file browser

Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort

The following job has been successfully added to the queue:

1: NM_002545.fasta

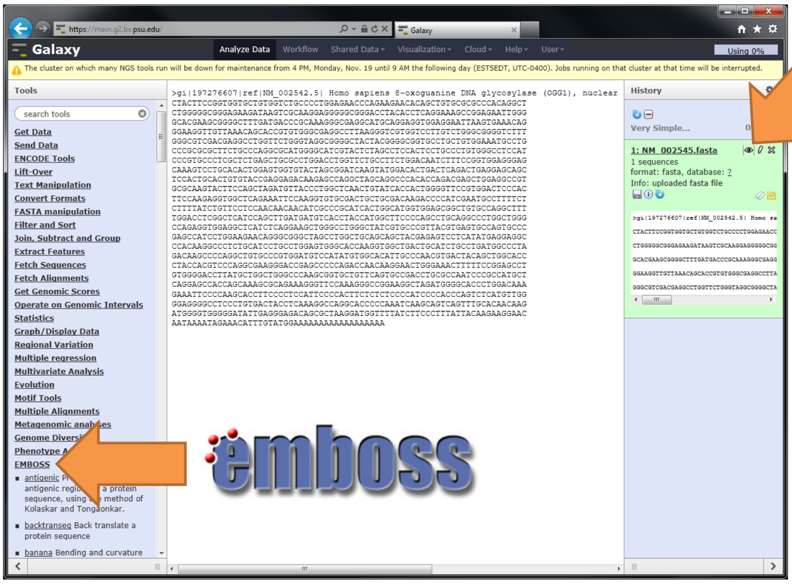
You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

1: NM_002545.fasta

After a while...

1: NM_002545.fasta
1 sequences
format: fasta, database: ?
Info: uploaded fasta file

```
>gi|197276607|ref|NM_002542.5| Homo sa
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGAACC
CTGGGGGCGGGAGAAGATAAGTCGCAAGGAGGGGGCGG
GCACGAAGCGGGGCTTTGATGACCCGCAAAGGGCGAGG
GGAAGGTTGTTAAACAGCACCGTGTGGGCGAGGCCTTA
GGGCGTCGACGAGGCCTGGTTCTGGGTAGGCGGGGCTA
```

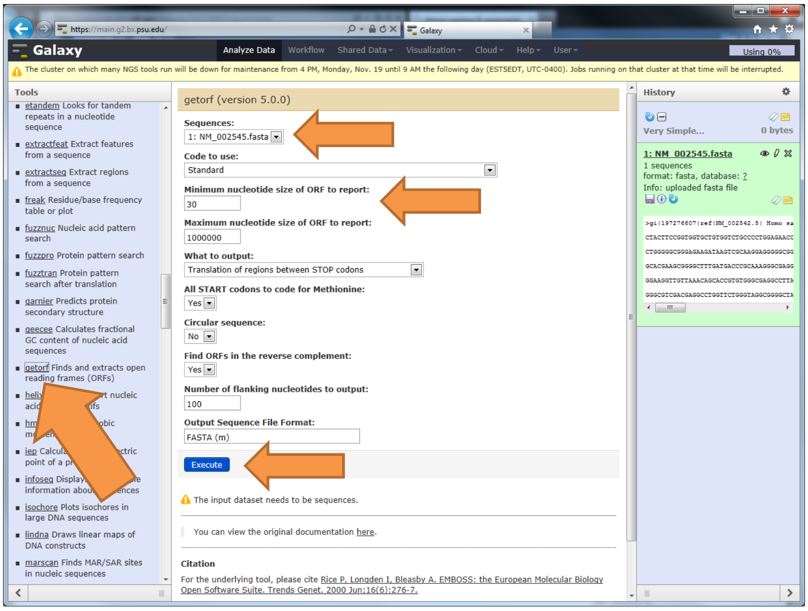


The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- 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
- Genome Diversi
- Phenotype A
- EMBOSS
  - antigenic Pr antigenic regio a protein sequence, using e method of Kolaskar and Tongaonkar.
  - backtranseq Back translate a protein sequence
  - banana Bending and curvature

```
>gi|197276607|ref|NM_002542.5| Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGAACCCAGAAGAACACAGCTGTGCGCGCCCACAGGCT
CTGGGGGCGGGAGAAGATAAGTCGCAAGGAGGGGGCGGGACCTACACCTCAGGAAAGCCGGAGAATTGGG
GCACGAAGCGGGGCTTTGATGACCCGCAAAGGGCGAGGCATGCAGGAGGTGGAGGAATTAAGTGAAACAG
GGAAGGTTGTTAAACAGCACCGTGTGGGCGAGGCCTTAAGGGTCGTGGTCCTTGTCTGGGCGGGGTCTTT
GGGCGTCGACGAGGCCTGGTTCTGGGTAGGCGGGGCTACTACGGGGCGGTGCCTGCTGTGGAAATGCCTG
CCCGCGCGCTTCTGCCCAGGCGCATGGGGCATCGTACTCTAGCCTCCACTCCTGCCCTGTGGGCCTCCAT
CCCGTGCCCTCGCTCTGAGCTGCGCCTGGACCTGGTTCTGCCTTCTGGACAATCTTTCCGGTGGAGGGAG
CAAAGTCCTGCACACTGGAGTGGTGTACTAGCGGATCAAGTATGGACACTGACTCAGACTGAGGAGCAGC
TCCACTGCACTGTGTACCGAGGAGACAAGAGCCAGGCTAGCAGGCCCACACCAGACGAGCTGGAGGCCGT
GCGCAAGTACTTCCAGCTAGATGTTACCCTGGCTCAACTGTATCACCACTGGGGTTCCGTGGACTCCCAC
TTCCAAGAGGTGGCTCAGAAATTCCAAGGTGTGCGACTGCTGCGACAAGACCCCATCGAATGCCTTTTCT
CTTTTATCTGTTCCTCCAACAACAACATCGCCCGCATCACTGGCATGGTGGAGCGGCTGTGCCAGGCTTT
TGGACCTCGGCTCATCCAGCTTGATGATGTCACCTACCATGGCTTCCCCAGCCTGCAGGCCCTGGCTGGG
CCAGAGGTGGAGGCTCATCTCAGGAAGCTGGGCCTGGGCTATCGTGCCCGTTACGTGAGTGCCAGTGCCC
GAGCCATCCTGGAAGAACAGGGCGGGCTAGCCTGGCTGCAGCAGCTACGAGAGTCCTCATATGAGGAGGC
CCACAAGGCCCTCTGCATCCTGCCTGGAGTGGGCACCAAGGTGGCTGACTGCATCTGCCTGATGGCCCTA
GACAAGCCCCAGGCTGTGCCCGTGGATGTCCATATGTGGCACATTGCCCAACGTGACTACAGCTGGCACC
CTACCACGTCCCAGGCGAAGGGACCGAGCCCCCAGACCAACAAGGAACTGGGAAACTTTTTCCGGAGCCT
GTGGGGACCTTATGCTGGCTGGGCCCAAGCGGTGCTGTTCAGTGCCGACCTGCGCCAATCCCGCCATGCT
CAGGAGCCACCAGCAAAGCGCAGAAAGGGTTCCAAAGGGCCGGAAGGCTAGATGGGGCACCCTGGACAAA
GAAATTCCCCAAGCACCTTCCCCTCCATTCCCCACTTCTCTCTCCCCATCCCCACCCAGTCTCATGTTGG
GGAGGGGCCTCCCTGTGACTACCTCAAAGGCCAGGCACCCCCAAATCAAGCAGTCAGTTTGCACAACAAG
ATGGGGTGGGGGATATTGAGGGAGACAGCGCTAAGGATGGTTTTATCTTCCCTTTATTACAAGAAGGAAC
AATAAAATAGAAACATTTGTATGGAAAAAAAAAAAAAAAAAAA
```

1: NM_002545.fasta
1 sequences
format: fasta, database: ?
Info: uploaded fasta file



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

- etandem Looks for tandem repeats in a nucleotide sequence
- extractfeat Extract features from a sequence
- extractseq Extract regions from a sequence
- freak Residue/base frequency table or plot
- fuzznuc Nucleic acid pattern search
- fuzzpro Protein pattern search
- fuzztran Protein pattern search after translation
- garnier Predicts protein secondary structure
- geecee Calculates fractional GC content of nucleic acid sequences
- getorf Finds and extracts open reading frames (ORFs)
- helix t nucleic acid ifs
- hm obic m
- iep Calcula ctric point of a pr
- infoseq Display e information abou ences
- isochore Plots isochores in large DNA sequences
- lindna Draws linear maps of DNA constructs
- marscan Finds MAR/SAR sites in nucleic sequences

getorf (version 5.0.0)

Sequences: 1: NM_002545.fasta

Code to use: Standard

Minimum nucleotide size of ORF to report: 30

Maximum nucleotide size of ORF to report: 1000000

What to output: Translation of regions between STOP codons

All START codons to code for Methionine: Yes

Circular sequence: No

Find ORFs in the reverse complement: Yes

Number of flanking nucleotides to output: 100

Output Sequence File Format: FASTA (m)

Execute

The input dataset needs to be sequences.

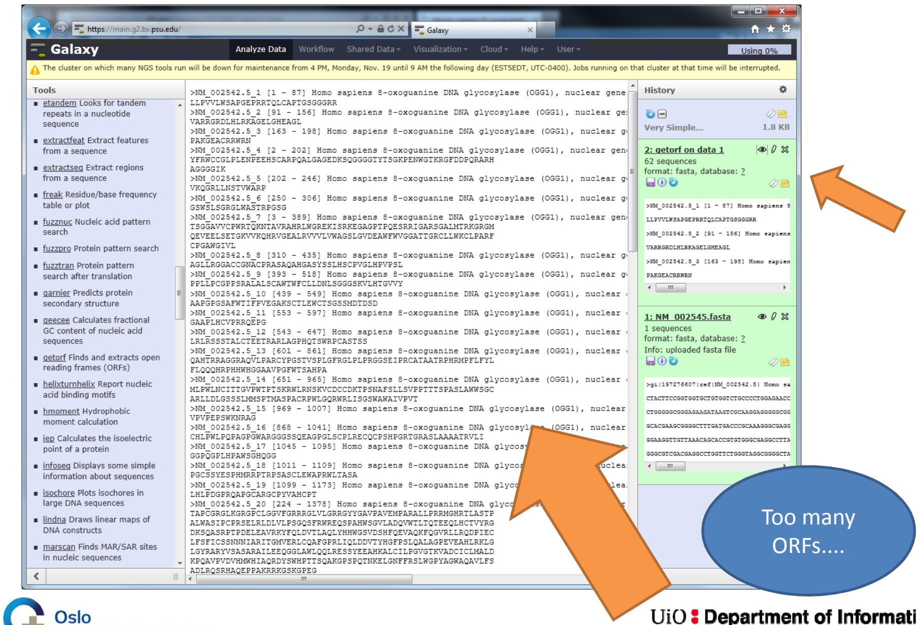
You can view the original documentation here.

Citation

For the underlying tool, please cite Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276-7.

1: NM_002545.fasta
1 sequences
format: fasta, database: ?
Info: uploaded fasta file

Jon K. Lærdahl,
Structural Bioinformatics



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Tools

- etandem Looks for tandem repeats in a nucleotide sequence
- extractfeat Extract features from a sequence
- extractseq Extract regions from a sequence
- freak Residue/base frequency table or plot
- fuzznuc Nucleic acid pattern search
- fuzzpro Protein pattern search
- fuzztran Protein pattern search after translation
- garnier Predicts protein secondary structure
- geecee Calculates fractional GC content of nucleic acid sequences
- getorf Finds and extracts open reading frames (ORFs)
- helixturnhelix Report nucleic acid binding motifs
- hmoment Hydrophobic moment calculation
- iep Calculates the isoelectric point of a protein
- infoseq Displays some simple information about sequences
- isochore Plots isochores in large DNA sequences
- lindna Draws linear maps of DNA constructs
- marscan Finds MAR/SAR sites in nucleic sequences

```
>NM_002542.5_1 [1 - 87] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear gene
LLPVVLWSAPGEPRRTQLCAPTGSGGGRR
>NM_002542.5_2 [91 - 156] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear ge
VARRGRDLHLRKAGELGHEAGL
>NM_002542.5_3 [163 - 198] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear g
PAKGEACRRWRN
>NM_002542.5_4 [2 - 202] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear gen
YFRWCCGLPLENPEEHSCARPQALGAGEDKSQGGGGTYTSGKPENWGTKRGFDDPQRARH
AGGGGIK
>NM_002542.5_5 [202 - 246] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear g
VKQGRLLNSTVWARP
>NM_002542.5_6 [250 - 306] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear g
GSWSLSGRGLWASTRPGSG
>NM_002542.5_7 [3 - 389] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear gen
TSGGAVVCPWRTQKNTAVRAHRLWGREKISRKEGAGPTPQESRRIGARSGALMTRKGRGM
QEVEELSETGKVVKQHRVGEALRVVVLVWAGSLGVDEAWFWVGGATTGRCLLWKCLPARF
CPGAWGIVL
>NM_002542.5_8 [310 - 435] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear g
AGLLRGGACCGNACPRASAQAHGASYSSLHSCPVGLHPVPSL
>NM_002542.5_9 [393 - 518] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear g
PPLLPCGPPSRALALSCAWTWFCLLDNLSGGGSKVLHTGVVY
>NM_002542.5_10 [439 - 549] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
AAPGPGSAFWTIFPVEGAKSCTLEWCTSGSSMDTDSD
>NM_002542.5_11 [553 - 597] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
GAAPLHCVPRRQEPG
>NM_002542.5_12 [543 - 647] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
LRLRSSSTALCTEETRARLAGPHQTSWRPCASTSS
>NM_002542.5_13 [601 - 861] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
QAHTRRAGGRAQVLPARCYPGSTVSPLGFRGLPLPRGGSEIPRCATAATRPHRMPFLFYL
FLQQQHRPHHWHGGAAVPGFWTSAHPA
>NM_002542.5_14 [651 - 965] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
MLPWLNCITTGVPWTPTSKRWLRNSKVCDCCDKTPSNAFSLLSVPPTTTSPASLAWWSGC
ARLLDLGSSSLMMSPTMASPACRPWLGQRWRLISGSWAWAIVPVT
>NM_002542.5_15 [969 - 1007] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
VPVPEPSWKNRAG
>NM_002542.5_16 [868 - 1041] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear
CHLPWLPQPAGPGWARGGGSSQEAGPGLSCPLRECQCPSHPGRTGRASLAAAATRVLI
>NM_002542.5_17 [1045 - 1095] Homo sapiens 8-oxoguanine DNA glycos
GGPQGPLHPAWSGHQGG
>NM_002542.5_18 [1011 - 1109] Homo sapiens 8-oxoguanine DNA glycos                 nuclea
PGCSSYESPHMRRPTRPSASCLEWAPRWLTASA
>NM_002542.5_19 [1099 - 1173] Homo sapiens 8-oxoguanine DNA glyco                   clea
LHLPDGPRQAPGCARGCPYVAHCPT
>NM_002542.5_20 [224 - 1378] Homo sapiens 8-oxoguanine DNA glyco
TAPCGRGLKGRGPCLGGVFGRRRGLVLGRRGYYGAVPAVEMPARALLPRRMGHRTLASTP
ALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQTEEQLHCTVYRG
DKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIEC
LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLG
LGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALD
KPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFS
ADLRQSRHAQEPPAKRRKGSKGPEG
```

History

Very Simple... 1.8 KB

2: getorf on data 1
62 sequences
format: fasta, database: ?

```
>NM_002542.5_1 [1 - 87] Homo sapiens 8
LLPVVLWSAPGEPRRTQLCAPTGSGGGRR
>NM_002542.5_2 [91 - 156] Homo sapiens
VARRGRDLHLRKAGELGHEAGL
>NM_002542.5_3 [163 - 198] Homo sapien
PAKGEACRRWRN
```

1: NM_002545.fasta
1 sequences
format: fasta, database: ?
Info: uploaded fasta file

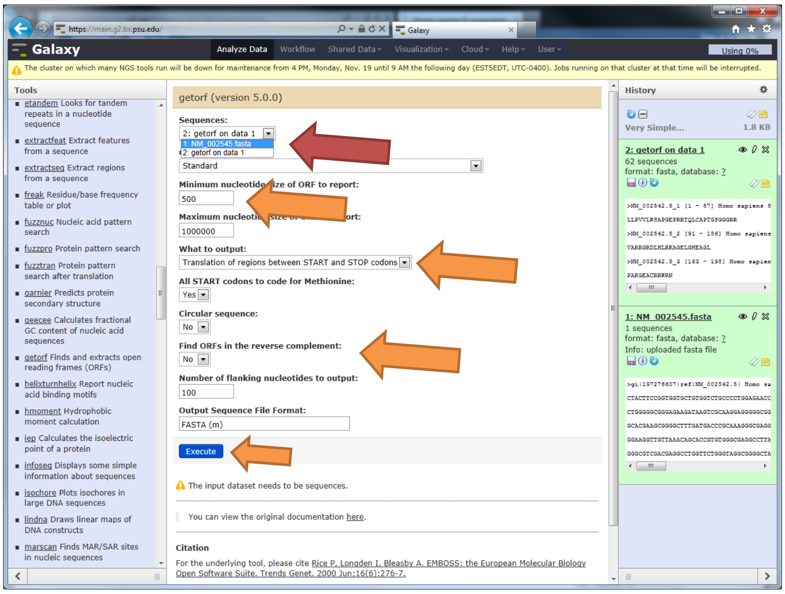
```
>gi|197276607|ref|NM_002542.5| Homo sa
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGAACC
CTGGGGGCGGGAGAAGATAAGTCGCAAGGAGGGGGCGG
GCACGAAGCGGGGCTTTGATGACCCGCAAAGGGCGAGG
GGAAGGTTGTTAAACAGCACCGTGTGGGCGAGGCCTTA
GGGCGTCGACGAGGCCTGGTTCTGGGTAGGCGGGGCTA
```

Too many ORFs....

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Tools

- etandem Looks for tandem repeats in a nucleotide sequence
- extractfeat Extract features from a sequence
- extractseq Extract regions from a sequence
- freak Residue/base frequency table or plot
- fuzznuc Nucleic acid pattern search
- fuzzpro Protein pattern search
- fuzztran Protein pattern search after translation
- garnier Predicts protein secondary structure
- geecee Calculates fractional GC content of nucleic acid sequences
- getorf Finds and extracts open reading frames (ORFs)
- helixturnhelix Report nucleic acid binding motifs
- hmoment Hydrophobic moment calculation
- iep Calculates the isoelectric point of a protein
- infoseq Displays some simple information about sequences
- isochore Plots isochores in large DNA sequences
- lindna Draws linear maps of DNA constructs
- marscan Finds MAR/SAR sites in nucleic sequences

getorf (version 5.0.0)

Sequences: 2: getorf on data 1
1: NM_002545.fasta
2: getorf on data 1

Standard

Minimum nucleotide size of ORF to report: 500

Maximum nucleotide size of ORF to report: 1000000

What to output: Translation of regions between START and STOP codons

All START codons to code for Methionine: Yes

Circular sequence: No

Find ORFs in the reverse complement: No

Number of flanking nucleotides to output: 100

Output Sequence File Format: FASTA (m)

Execute

The input dataset needs to be sequences.

You can view the original documentation here.

Citation

For the underlying tool, please cite Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276-7.

History

Very Simple... 1.8 KB

2: getorf on data 1
62 sequences
format: fasta, database: ?

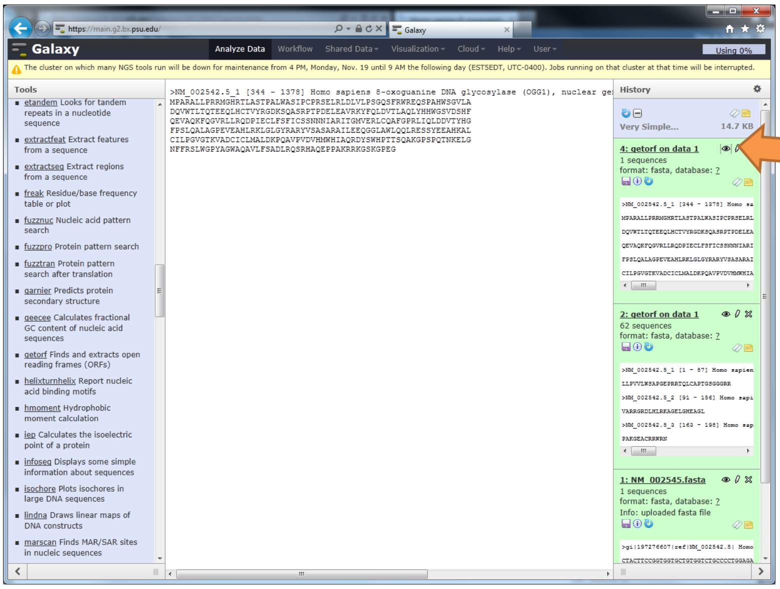
```
>NM_002542.5_1 [1 - 87] Homo sapiens 8
LLPVVLWSAPGEPRRTQLCAPTGSGGGRR
>NM_002542.5_2 [91 - 156] Homo sapiens
VARRGRDLHLRKAGELGHEAGL
>NM_002542.5_3 [163 - 198] Homo sapien
PAKGEACRRWRN
```

1: NM_002545.fasta
1 sequences
format: fasta, database: ?
Info: uploaded fasta file

```
>gi|197276607|ref|NM_002542.5| Homo sa
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGAACC
CTGGGGGCGGGAGAAGATAAGTCGCAAGGAGGGGGCGG
GCACGAAGCGGGGCTTTGATGACCCGCAAAGGGCGAGG
GGAAGGTTGTTAAACAGCACCGTGTGGGCGAGGCCTTA
GGGCGTCGACGAGGCCTGGTTCTGGGTAGGCGGGGCTA
```

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Tools

- etandem Looks for tandem repeats in a nucleotide sequence
- extractfeat Extract features from a sequence
- extractseq Extract regions from a sequence
- freak Residue/base frequency table or plot
- fuzznuc Nucleic acid pattern search
- fuzzpro Protein pattern search
- fuzztran Protein pattern search after translation
- garnier Predicts protein secondary structure
- geecee Calculates fractional GC content of nucleic acid sequences
- getorf Finds and extracts open reading frames (ORFs)
- helixturnhelix Report nucleic acid binding motifs
- hmoment Hydrophobic moment calculation
- iep Calculates the isoelectric point of a protein
- infoseq Displays some simple information about sequences
- isochore Plots isochores in large DNA sequences
- lindna Draws linear maps of DNA constructs
- marscan Finds MAR/SAR sites in nucleic sequences

```
>NM_002542.5_1 [344 - 1378] Homo sapiens 8-oxoguanine DNA glycosylase (OGG1), nuclear ge
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
```

History

Very Simple... 14.7 KB

4: getorf on data 1
1 sequences
format: fasta, database: ?

```
>NM_002542.5_1 [344 - 1378] Homo sa
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEA
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARI
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAI
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIA
```

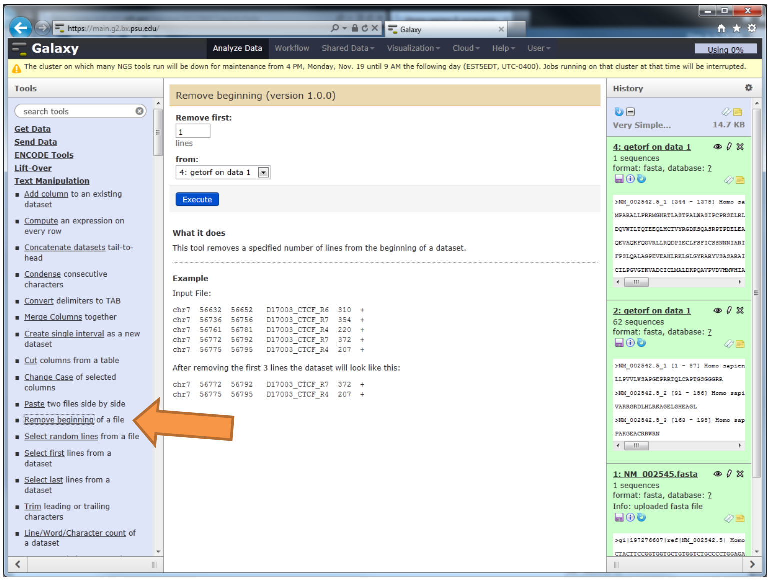
2: getorf on data 1
62 sequences
format: fasta, database: ?

```
>NM_002542.5_1 [1 - 87] Homo sapien
LLPVVLWSAPGEPRRTQLCAPTGSGGGRR
>NM_002542.5_2 [91 - 156] Homo sapi
VARRGRDLHLRKAGELGHEAGL
>NM_002542.5_3 [163 - 198] Homo sap
PAKGEACRRWRN
```

1: NM_002545.fasta
1 sequences
format: fasta, database: ?
Info: uploaded fasta file

```
>gi|197276607|ref|NM_002542.5| Homo
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGA
```

Where is 3?
I did a mistake and had to remove it



The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Tools

search tools

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Condense consecutive characters
- Convert delimiters to TAB
- Merge Columns together
- Create single interval as a new dataset
- Cut columns from a table
- Change Case of selected columns
- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset

Remove beginning (version 1.0.0)

Remove first:
1
lines

from:
4: getorf on data 1

Execute

What it does

This tool removes a specified number of lines from the beginning of a dataset.

Example

Input File:

```
chr7  56632  56652   D17003_CTCF_R6  310  +
chr7  56736  56756   D17003_CTCF_R7  354  +
chr7  56761  56781   D17003_CTCF_R4  220  +
chr7  56772  56792   D17003_CTCF_R7  372  +
chr7  56775  56795   D17003_CTCF_R4  207  +
```

After removing the first 3 lines the dataset will look like this:

```
chr7  56772  56792   D17003_CTCF_R7  372  +
chr7  56775  56795   D17003_CTCF_R4  207  +
```

History

Very Simple... 14.7 KB

4: getorf on data 1
1 sequences
format: fasta, database: ?

```
>NM_002542.5_1 [344 - 1378] Homo sa
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEA
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARI
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAI
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIA
```

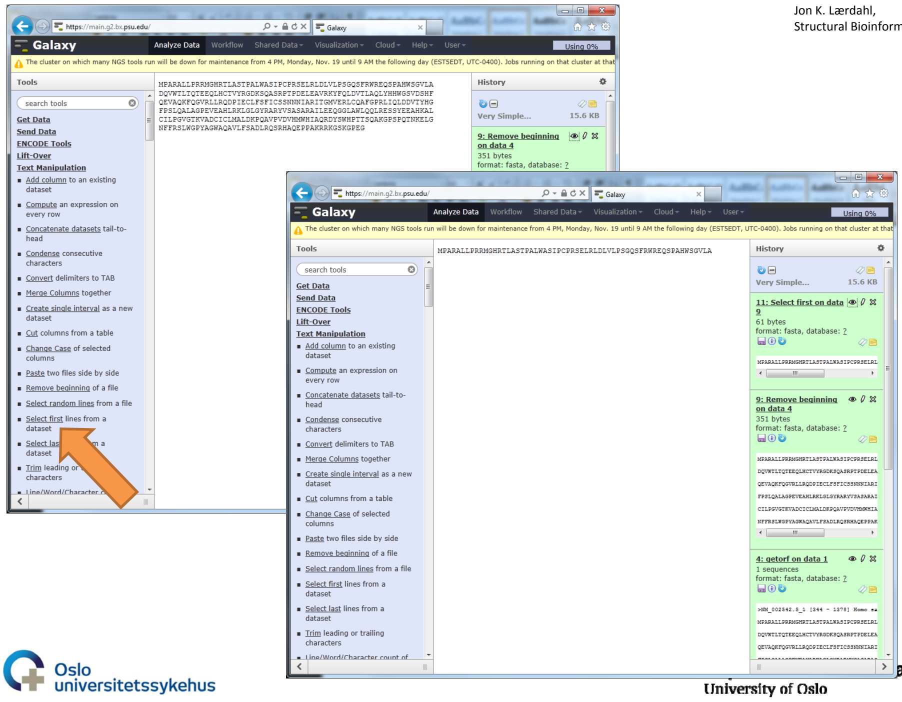
2: getorf on data 1
62 sequences
format: fasta, database: ?

```
>NM_002542.5_1 [1 - 87] Homo sapien
LLPVVLWSAPGEPRRTQLCAPTGSGGGRR
>NM_002542.5_2 [91 - 156] Homo sapi
VARRGRDLHLRKAGELGHEAGL
>NM_002542.5_3 [163 - 198] Homo sap
PAKGEACRRWRN
```

1: NM_002545.fasta
1 sequences
format: fasta, database: ?
Info: uploaded fasta file

```
>gi|197276607|ref|NM_002542.5| Homo
CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGA
```

Jon K. Lærdahl,
Structural Bioinformatics


https://main.g2.bx.psu.edu/
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Add column to an existing dataset
Compute an expression on every row
Concatenate datasets tail-to-head
Condense consecutive characters
Convert delimiters to TAB
Merge Columns together
Create single interval as a new dataset
Cut columns from a table
Change Case of selected columns
Paste two files side by side
Remove beginning of a file
Select random lines from a file
Select first lines from a dataset
Select last lines from a dataset
Trim leading or trailing characters
Line/Word/Character count of
History
Very Simple...
15.6 KB

The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
```

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
```

11: Select first on data 9
61 bytes
format: fasta, database: ?

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL
```

9: Remove beginning on data 4
351 bytes
format: fasta, database: ?

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEA
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARI
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAI
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIA
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAK
```

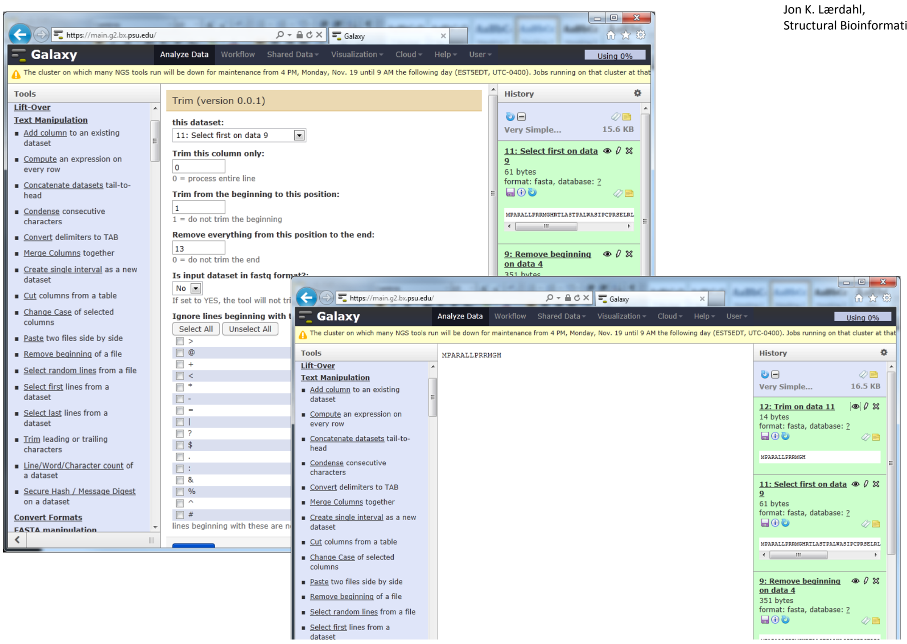
4: getorf on data 1
1 sequences
format: fasta, database: ?

```
>NM_002542.5_1 [344 - 1378] Homo sa
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEA
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARI
```

Oslo universitetssykehus

atics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics


https://main.g2.bx.psu.edu/
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
Lift-Over
Text Manipulation
Add column to an existing dataset
Compute an expression on every row
Concatenate datasets tail-to-head
Condense consecutive characters
Convert delimiters to TAB
Merge Columns together
Create single interval as a new dataset
Cut columns from a table
Change Case of selected columns
Paste two files side by side
Remove beginning of a file
Select random lines from a file
Select first lines from a dataset
Select last lines from a dataset
Trim leading or trailing characters
Line/Word/Character count of a dataset
Secure Hash / Message Digest on a dataset
Convert Formats
FASTA manipulation
History
Very Simple...
15.6 KB
16.5 KB

The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that

Trim (version 0.0.1)

this dataset: 11: Select first on data 9

Trim this column only: 0
0 = process entire line

Trim from the beginning to this position: 1
1 = do not trim the beginning

Remove everything from this position to the end: 13
0 = do not trim the end

Is input dataset in fastq format?: No
If set to YES, the tool will not tri

Ignore lines beginning with t
Select All Unselect All
□ >
□ @
□ +
□ <
□ *
□ -
□ =
□ |
□ ?
□ $
□ .
□ :
□ &
□ %
□ ^
□ #
lines beginning with these are n

```
MPARALLPRRMGH
```

12: Trim on data 11
14 bytes
format: fasta, database: ?

```
MPARALLPRRMGH
```

11: Select first on data 9
61 bytes
format: fasta, database: ?

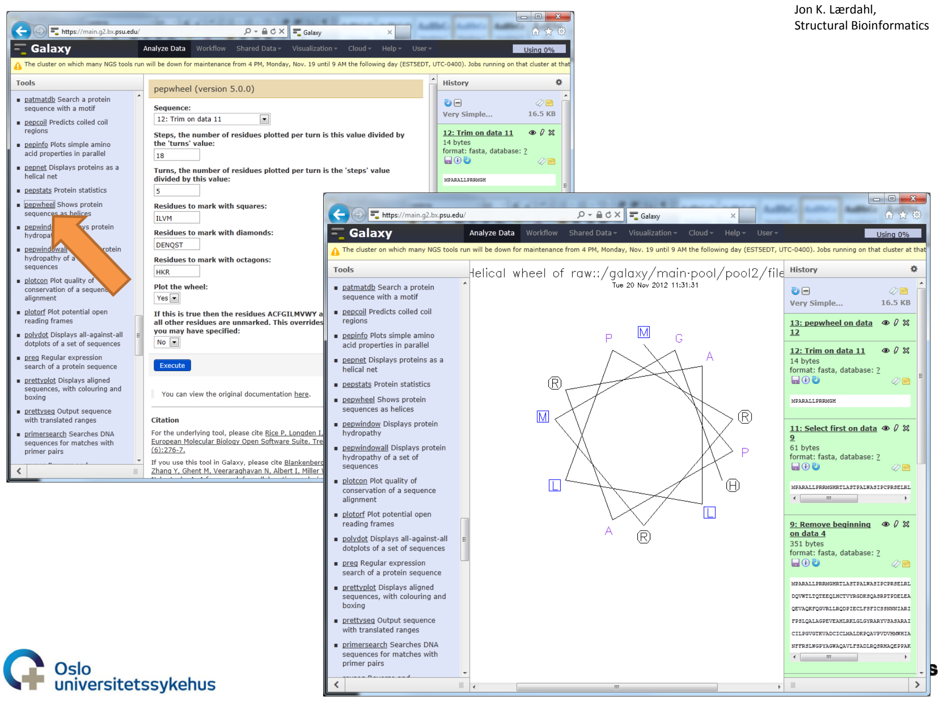
```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL
```

9: Remove beginning on data 4
351 bytes
format: fasta, database: ?

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics



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

The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that

- patmatdb Search a protein sequence with a motif
- pepcoil Predicts coiled coil regions
- pepinfo Plots simple amino acid properties in parallel
- pepnet Displays proteins as a helical net
- pepstats Protein statistics
- pepwheel Shows protein sequences as helices
- pepwindow Displays protein hydropathy
- pepwindowall Displays protein hydropathy of a set of sequences
- plotcon Plot quality of conservation of a sequence alignment
- plotorf Plot potential open reading frames
- polydot Displays all-against-all dotplots of a set of sequences
- preg Regular expression search of a protein sequence
- prettyplot Displays aligned sequences, with colouring and boxing
- prettyseq Output sequence with translated ranges
- primersearch Searches DNA sequences for matches with primer pairs

pepwheel (version 5.0.0)

Sequence: 12: Trim on data 11

Steps, the number of residues plotted per turn is this value divided by the 'turns' value: 18

Turns, the number of residues plotted per turn is the 'steps' value divided by this value: 5

Residues to mark with squares: ILVM

Residues to mark with diamonds: DENQST

Residues to mark with octagons: HKR

Plot the wheel: Yes

If this is true then the residues ACFGILMVWY a… all other residues are unmarked. This overrides… you may have specified: No

You can view the original documentation here.

Citation

For the underlying tool, please cite Rice P, Longden I, … European Molecular Biology Open Software Suite. Tre… (6):276-7.

If you use this tool in Galaxy, please cite Blankenberg… Zhang Y, Ghent M, Veeraraghavan N, Albert I, Miller …

Helical wheel of raw::/galaxy/main-pool/pool2/file
Tue 20 Nov 2012 11:31:31

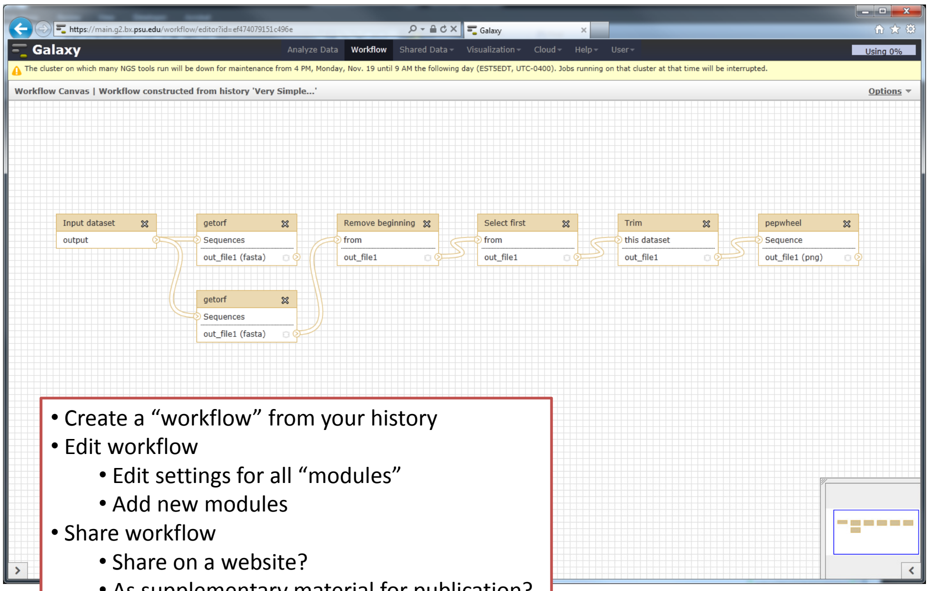
Very Simple... 16.5 KB

13: pepwheel on data 12

12: Trim on data 11
14 bytes
format: fasta, database: ?
MPARALLPRRMGH

11: Select first on data 9
61 bytes
format: fasta, database: ?
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL

9: Remove beginning on data 4
351 bytes
format: fasta, database: ?
MPARALLPRRMGHRTLASTPALWASIPCPRSELRL
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEA
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARI
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAI
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIA
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAK

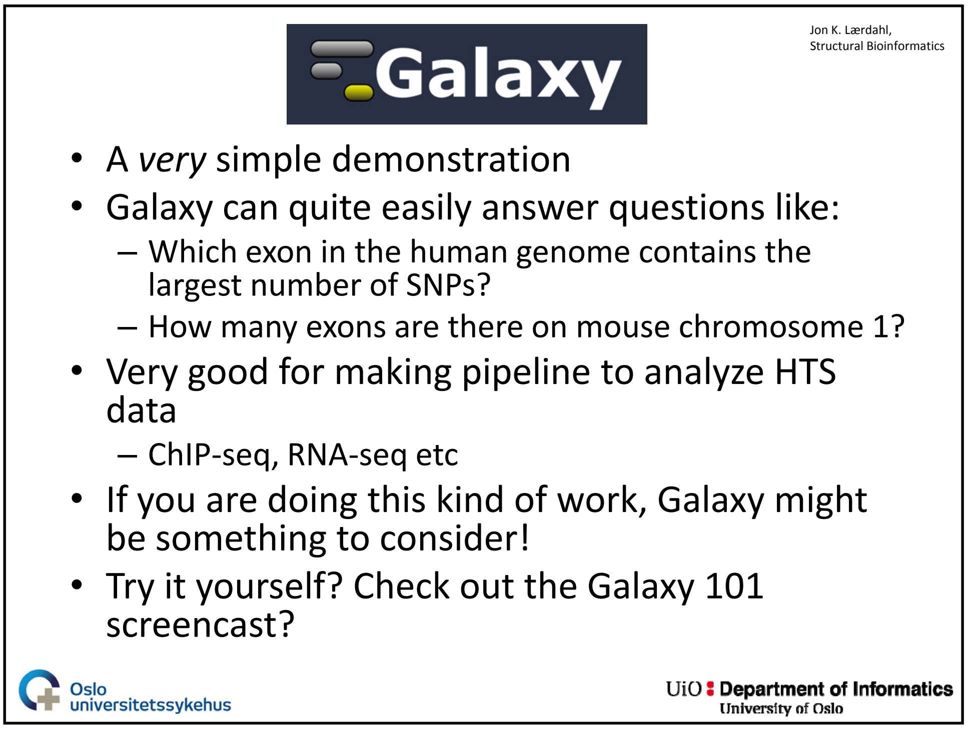


https://main.g2.bx.psu.edu/workflow/editor?id=ef474079151c496e

The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST5EDT, UTC-0400). Jobs running on that cluster at that time will be interrupted.

Workflow Canvas | Workflow constructed from history 'Very Simple...'

- Create a “workflow” from your history
- Edit workflow
  - Edit settings for all “modules”
  - Add new modules
- Share workflow
  - Share on a website?
  - As supplementary material for publication?



Galaxy

Jon K. Lærdahl,
Structural Bioinformatics

- A *very* simple demonstration
- Galaxy can quite easily answer questions like:
  - Which exon in the human genome contains the largest number of SNPs?
  - How many exons are there on mouse chromosome 1?
- Very good for making pipeline to analyze HTS data
  - ChIP-seq, RNA-seq etc
- If you are doing this kind of work, Galaxy might be something to consider!
- Try it yourself? Check out the Galaxy 101 screencast?

Oslo
universitetssykehus

UiO : Department of Informatics
University of Oslo