Bio-Linux:

A FLOSS (Free/Libre Open Source Software) platform for genomic data analysis

Tony Travis

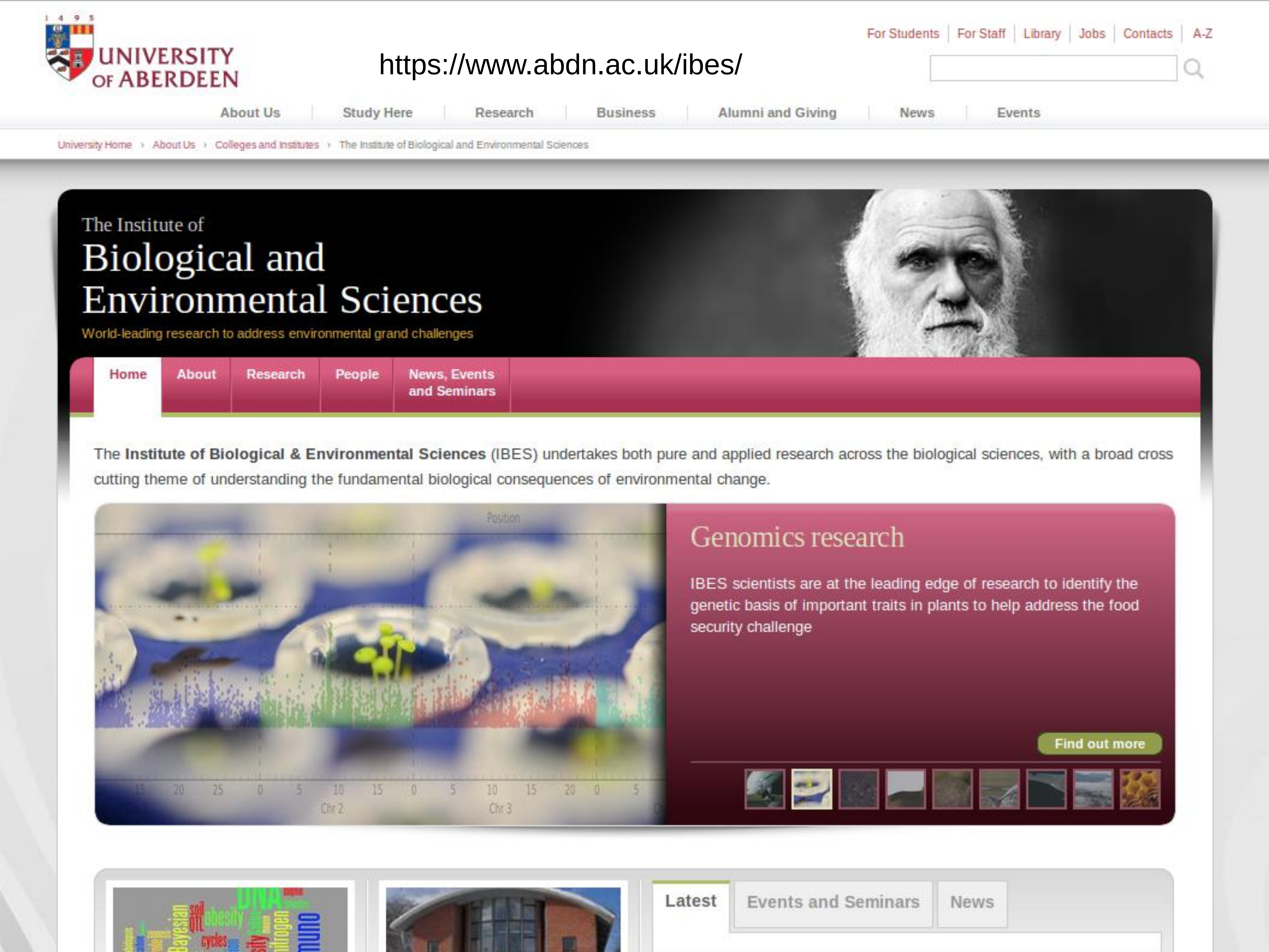
University of Aberdeen Institute of Biological and Environmental Sciences
and
Minke Informatics Limited

Basel Life Science Week

Next Generation Sequencing:
Clinical and research applications

Thu 24 Sep 2015

https://www.abdn.ac.uk/ibes/



1 4 9 5
UNIVERSITY OF ABERDEEN
For Students | For Staff | Library | Jobs | Contacts | A-Z
About Us
Study Here
Research
Business
Alumni and Giving
News
Events
University Home › About Us › Colleges and Institutes › The Institute of Biological and Environmental Sciences

The Institute of
Biological and
Environmental Sciences

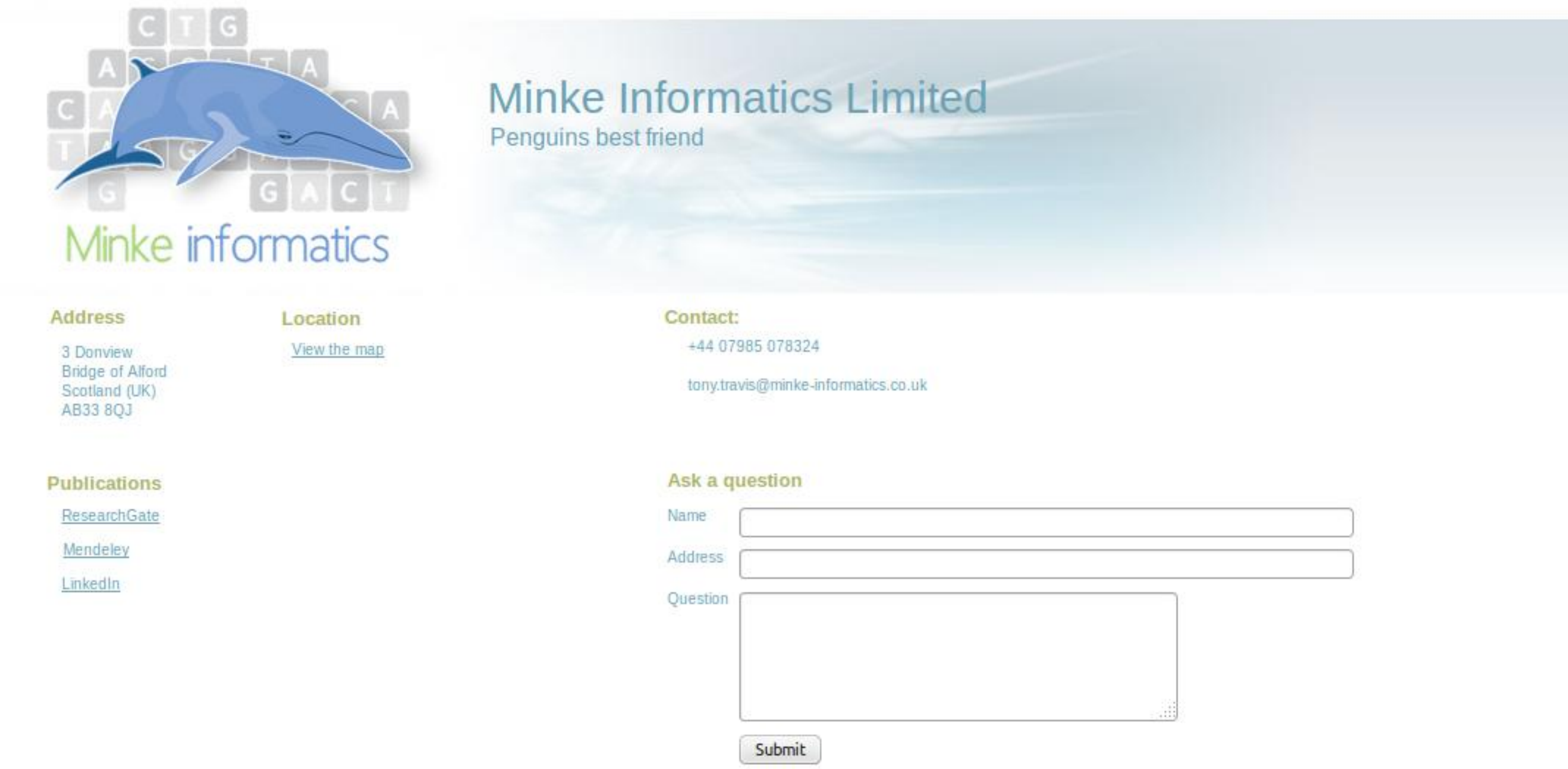
World-leading research to address environmental grand challenges

Home
About
Research
People
News, Events and Seminars

The **Institute of Biological & Environmental Sciences** (IBES) undertakes both pure and applied research across the biological sciences, with a broad cross cutting theme of understanding the fundamental biological consequences of environmental change.

Genomics research

IBES scientists are at the leading edge of research to identify the genetic basis of important traits in plants to help address the food security challenge



Minke Informatics Limited

Penguins best friend

Address

3 Donview
Bridge of Alford
Scotland (UK)
AB33 8QJ

Location

View the map

Contact:

+44 07985 078324

tony.travis@minke-informatics.co.uk

Publications

ResearchGate

Mendeley

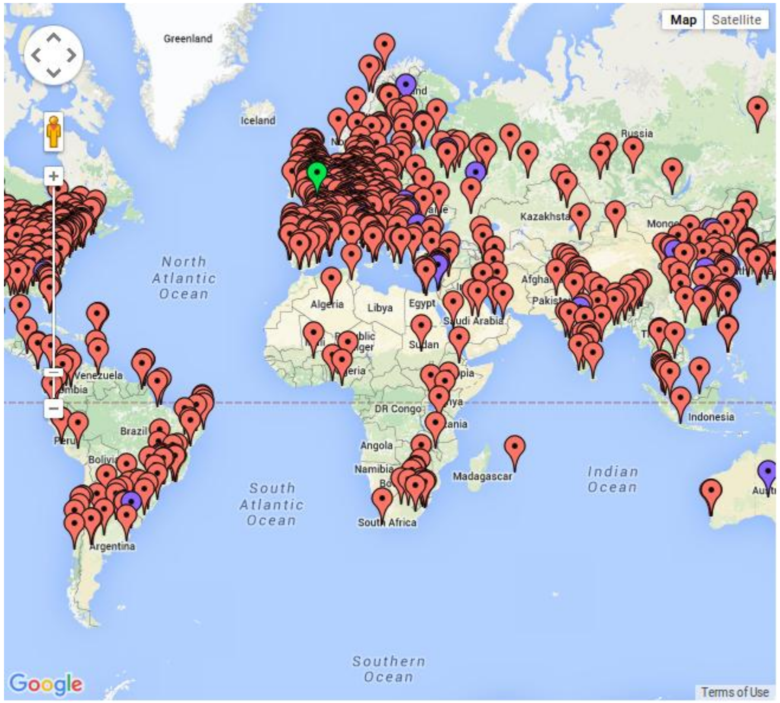
LinkedIn

Ask a question

http://minke-informatics.co.uk/

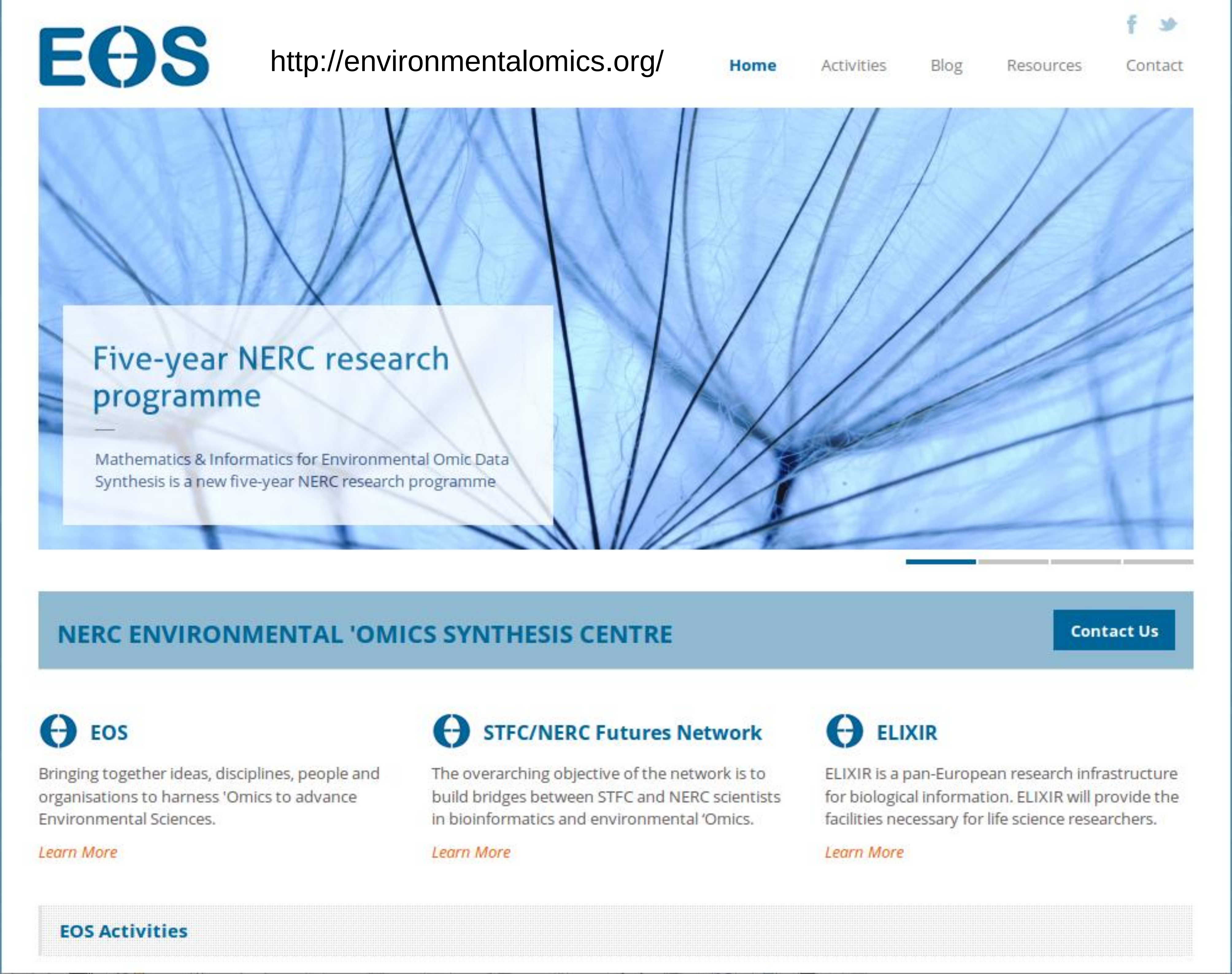
NERC EOS Bio-Linux

- NERC
  - Natural Environment Research Council (UK)
- EOS
  - NERC Environmental 'Omics Synthesis Centre
- NEBC
  - NERC Envirnomental Bioinformatics Centre



http://environmentalomics.org/bio-linux/

http://environmentalomics.org/



Five-year NERC research programme

Mathematics & Informatics for Environmental Omic Data Synthesis is a new five-year NERC research programme

NERC ENVIRONMENTAL 'OMICS SYNTHESIS CENTRE

EOS

Bringing together ideas, disciplines, people and organisations to harness 'Omics to advance Environmental Sciences.

STFC/NERC Futures Network

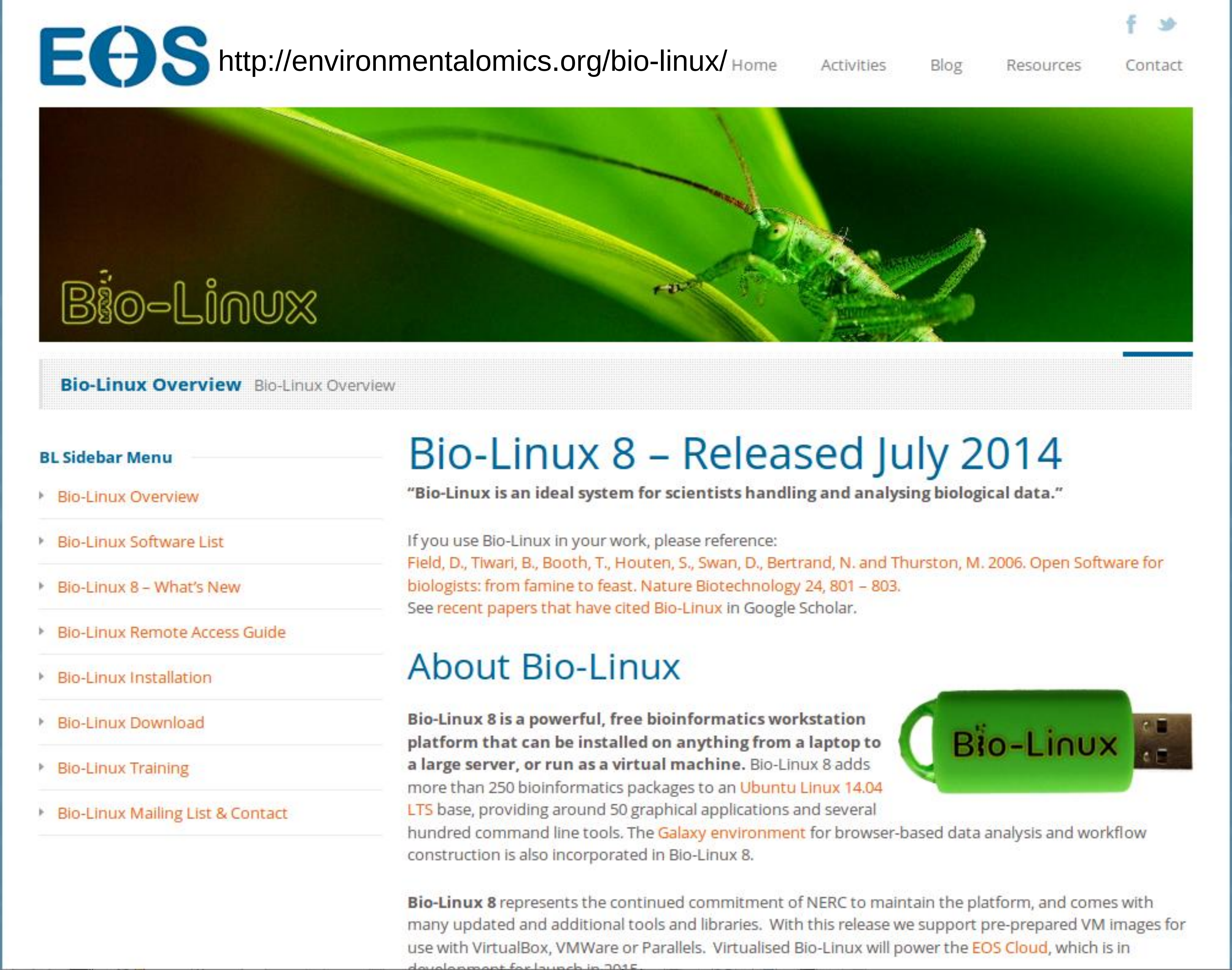
The overarching objective of the network is to build bridges between STFC and NERC scientists in bioinformatics and environmental 'Omics.

ELIXIR

ELIXIR is a pan-European research infrastructure for biological information. ELIXIR will provide the facilities necessary for life science researchers.

EOS Activities

http://environmentalomics.org/bio-linux/



Bio-Linux Overview Bio-Linux Overview

BL Sidebar Menu

- Bio-Linux Overview
- Bio-Linux Software List
- Bio-Linux 8 – What's New
- Bio-Linux Remote Access Guide
- Bio-Linux Installation
- Bio-Linux Download
- Bio-Linux Training
- Bio-Linux Mailing List & Contact

Bio-Linux 8 – Released July 2014

"Bio-Linux is an ideal system for scientists handling and analysing biological data."

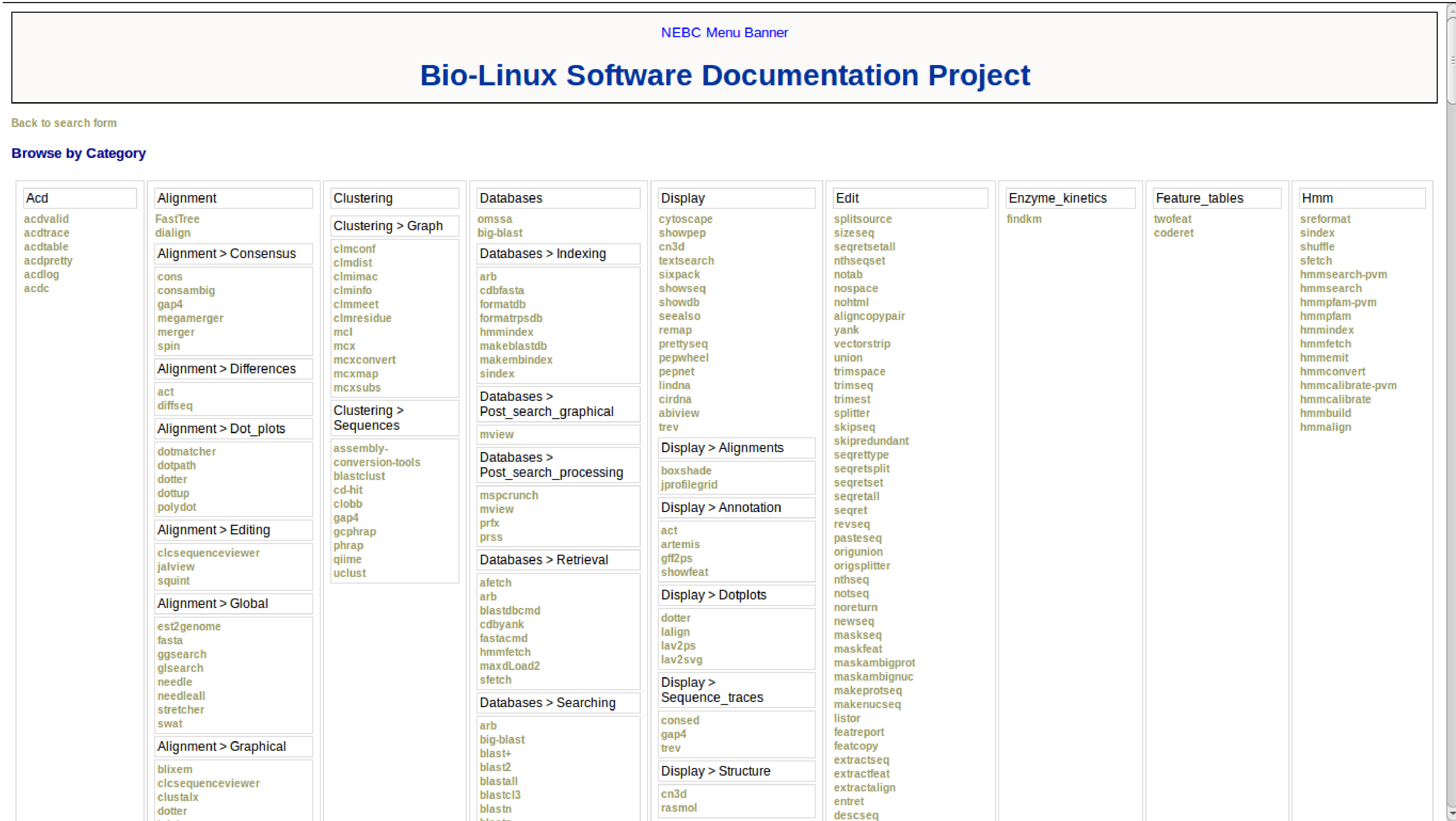
If you use Bio-Linux in your work, please reference:
Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N. and Thurston, M. 2006. Open Software for biologists: from famine to feast. Nature Biotechnology 24, 801 – 803.
See recent papers that have cited Bio-Linux in Google Scholar.

About Bio-Linux

Bio-Linux 8 is a powerful, free bioinformatics workstation platform that can be installed on anything from a laptop to a large server, or run as a virtual machine. Bio-Linux 8 adds more than 250 bioinformatics packages to an Ubuntu Linux 14.04 LTS base, providing around 50 graphical applications and several hundred command line tools. The Galaxy environment for browser-based data analysis and workflow construction is also incorporated in Bio-Linux 8.

Bio-Linux 8 represents the continued commitment of NERC to maintain the platform, and comes with many updated and additional tools and libraries. With this release we support pre-prepared VM images for use with VirtualBox, VMWare or Parallels. Virtualised Bio-Linux will power the EOS Cloud, which is in development for launch in 2015.

Bio-Linux contains over 250 software packages



NEBC Menu Banner

Bio-Linux Software Documentation Project

Back to search form

Browse by Category

Acd

acdvalid
acdtrace
acdtable
acdpretty
acdlog
acdc

Alignment

FastTree
dialign

Alignment > Consensus

cons
consambig
gap4
megamerger
merger
spin

Alignment > Differences

act
diffseq

Alignment > Dot_plots

dotmatcher
dotpath
dotter
dottup
polydot

Alignment > Editing

clcsequenceviewer
jalview
squint

Alignment > Global

est2genome
fasta
ggsearch
glsearch
needle
needleall
stretcher
swat

Alignment > Graphical

blixem
clcsequenceviewer
clustalx
dotter

Clustering

Clustering > Graph

clmconf
clmdist
clmimac
clminfo
clmmeet
clmresidue
mcl
mcx
mcxconvert
mcxmap
mcxsubs

Clustering > Sequences

assembly-conversion-tools
blastclust
cd-hit
clobb
gap4
gcphrap
phrap
qiime
uclust

Databases

omssa
big-blast

Databases > Indexing

arb
cdbfasta
formatdb
formatrpsdb
hmmindex
makeblastdb
makembindex
sindex

Databases > Post_search_graphical

mview

Databases > Post_search_processing

mspcrunch
mview
prfx
prss

Databases > Retrieval

afetch
arb
blastdbcmd
cdbyank
fastacmd
hmmfetch
maxdLoad2
sfetch

Databases > Searching

arb
big-blast
blast+
blast2
blastall
blastcl3
blastn

Display

cytoscape
showpep
cn3d
textsearch
sixpack
showseq
showdb
seealso
remap
prettyseq
pepwheel
pepnet
lindna
cirdna
abiview
trev

Display > Alignments

boxshade
jprofilegrid

Display > Annotation

act
artemis
gff2ps
showfeat

Display > Dotplots

dotter
lalign
lav2ps
lav2svg

Display > Sequence_traces

consed
gap4
trev

Display > Structure

cn3d
rasmol

Edit

splitsource
sizeseq
seqretsetall
nthseqset
notab
nospace
nohtml
aligncopypair
yank
vectorstrip
union
trimspace
trimseq
trimest
splitter
skipseq
skipredundant
seqrettype
seqretsplit
seqretset
seqretall
seqret
revseq
pasteseq
origunion
origsplitter
nthseq
notseq
noreturn
newseq
maskseq
maskfeat
maskambigprot
maskambignuc
makeprotseq
makenucseq
listor
featreport
featcopy
extractseq
extractfeat
extractalign
entret
descseq

Enzyme_kinetics

findkm

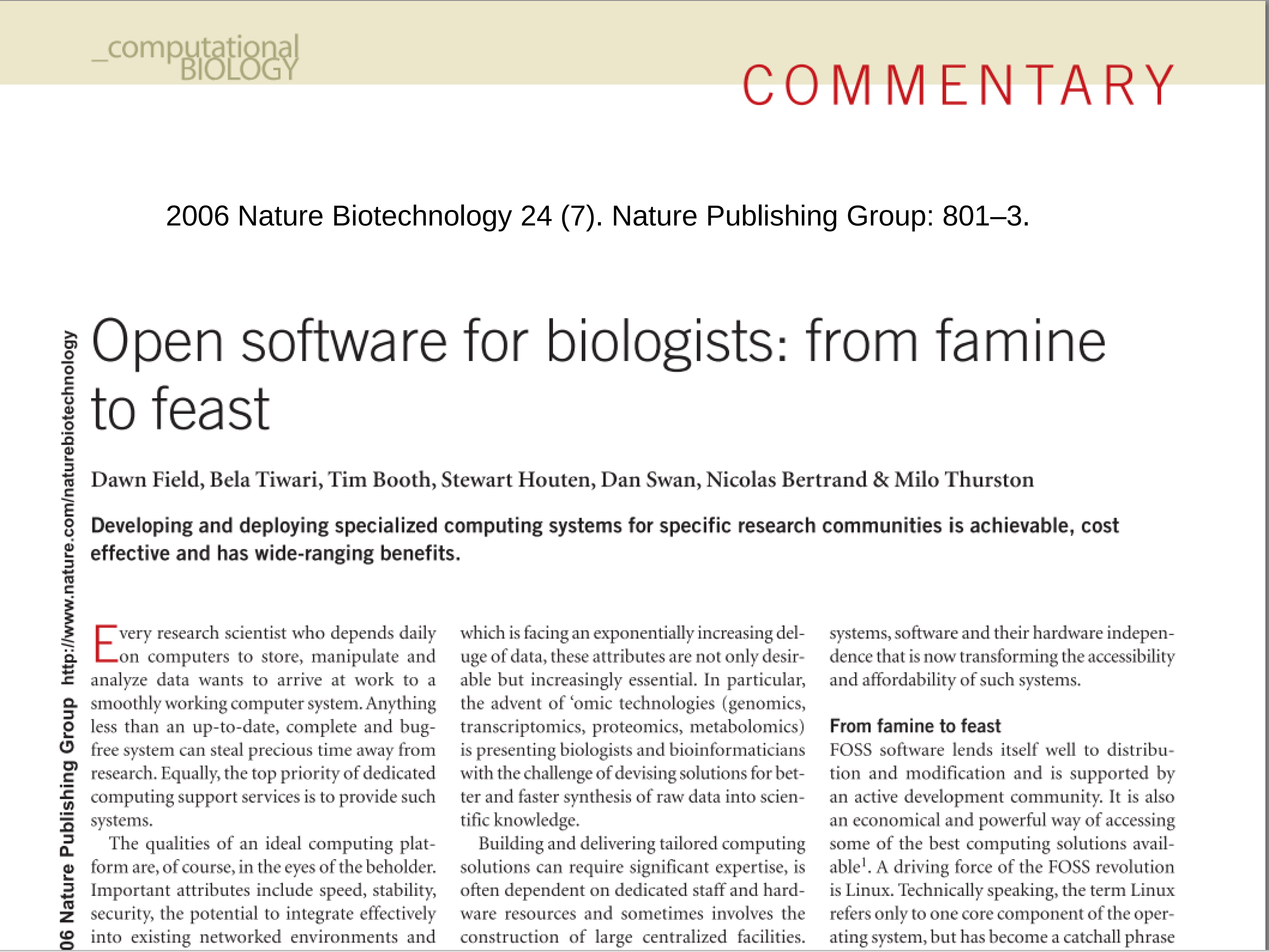
Feature_tables

twofeat
coderet

Hmm

sreformat
sindex
shuffle
sfetch
hmmsearch-pvm
hmmsearch
hmmpfam-pvm
hmmpfam
hmmindex
hmmfetch
hmmemit
hmmconvert
hmmcalibrate-pvm
hmmcalibrate
hmmbuild
hmmalign

2006 Nature Biotechnology 24 (7). Nature Publishing Group: 801–3.



Open software for biologists: from famine to feast

Dawn Field, Bela Tiwari, Tim Booth, Stewart Houten, Dan Swan, Nicolas Bertrand & Milo Thurston

Developing and deploying specialized computing systems for specific research communities is achievable, cost effective and has wide-ranging benefits.

Every research scientist who depends daily on computers to store, manipulate and analyze data wants to arrive at work to a smoothly working computer system. Anything less than an up-to-date, complete and bug-free system can steal precious time away from research. Equally, the top priority of dedicated computing support services is to provide such systems.

The qualities of an ideal computing platform are, of course, in the eyes of the beholder. Important attributes include speed, stability, security, the potential to integrate effectively into existing networked environments and which is facing an exponentially increasing deluge of data, these attributes are not only desirable but increasingly essential. In particular, the advent of 'omic technologies (genomics, transcriptomics, proteomics, metabolomics) is presenting biologists and bioinformaticians with the challenge of devising solutions for better and faster synthesis of raw data into scientific knowledge.

Building and delivering tailored computing solutions can require significant expertise, is often dependent on dedicated staff and hardware resources and sometimes involves the construction of large centralized facilities. systems, software and their hardware independence that is now transforming the accessibility and affordability of such systems.

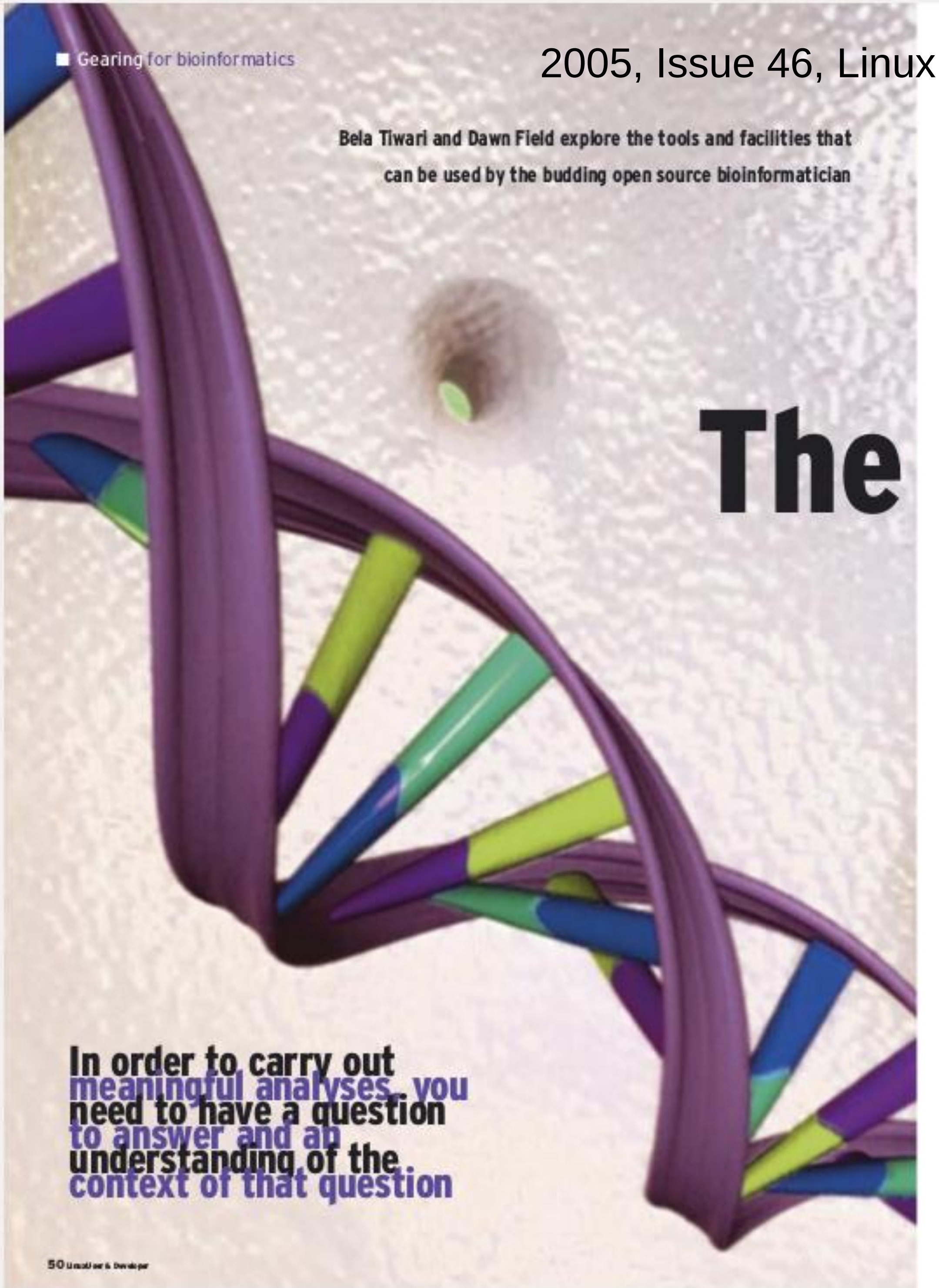
From famine to feast

FOSS software lends itself well to distribution and modification and is supported by an active development community. It is also an economical and powerful way of accessing some of the best computing solutions available[1]. A driving force of the FOSS revolution is Linux. Technically speaking, the term Linux refers only to one core component of the operating system, but has become a catchall phrase

Origins of Bio-Linux

- NERC requirements
  - Software platform to support the diverse bioinformatics used in research they fund
  - Cost-effective alternative to proprietary bioinformatics software
  - Used by biologists to analyse their *own* data
- Freely available bioinformatics software
  - Packaged, tested and documented

2005, Issue 46, Linux User and Developer



Bela Tiwari and Dawn Field explore the tools and facilities that can be used by the budding open source bioinformatician

In order to carry out meaningful analyses, you need to have a question to answer and an understanding of the context of that question.

The bioinformatics playground

‘Bioinformatics’ is a buzz word that is becoming increasingly audible in the Linux world. Fast, economical, flexible, and extensible computing power is making Linux increasingly attractive to scientists in many areas of research, including biology. More generally, the open source movement has greatly benefited biological research; the most publicised project being the publicly funded effort to sequence and make freely available the human genome. Less well publicised is the huge amount of biological data that can be freely accessed. The combination of data availability and free software is revolutionising this field.

The ability to redistribute Linux, the existence of online documentation, active user and developer communities, and the fact that much bioinformatics software is developed for Linux/Unix systems, has opened the way for individual users without access to large centralised resources to be able to install and run bioinformatics software to analyse data, and to start developing for the wider community.

Here we outline projects that can help to significantly ease the experience of trying out, using, and providing computing platforms appropriate for bioinformatics analyses.

KNOWING WHEN

Turning data into knowledge is a complex task that demands data manipulation, comparison, statistical analysis, visualisation, as well as data storage and dissemination. Usually, the weight of many lines of evidence must be combined to answer a scientific question, and the interpretation of the output of many different software tools plays a key role in discerning and assembling data from which biological knowledge is born.

Finding and installing common tools for bioinformatics on your own machine, especially for those new to Linux, can be a daunting task. Projects with enough funding are able to hire dedicated system administrators to provide sustainable bioinformatics computing systems, but many of us are not that lucky and have to go it alone.

To add to the challenge, much bioinformatics software is written by academics, and while there are some very good, well tested packages out there, there are also many that were intended to answer a particular question, on a particular machine, for a particular group. Such packages were often not built with portability, future use or further development in mind. Knowing when to persevere or give up with a piece of software is all part of the key skills of a bioinformatician or bioinformatics systems provider. Even very experienced system administrators can sometimes find installing and integrating bioinformatics software and databases frustrating and tedious.

Many developers have faced these challenges already and taking advantage of the resources some of them have made freely available can greatly reduce the overheads involved in establishing a new system for bioinformatics. Some of these resources are described in this article including CD and DVD-based Live Linux distributions customized for bioinformatics analyses, full distributions that can be installed from iso images or installed over the network, and also specialised package repositories. Each of these solutions has its particular attractions for users with different requirements.

PICKING YOUR SOLUTION

Whether you plan to use a system yourself or provide it for others, give thought to your long-term requirements. Questions you might be asking yourself include how much computing power you are likely to need, whether you require a cluster-based solution, how many databases need to be stored locally, how many users will depend on the system, how they will access it, etc. Live CD or DVD distributions may be good for an individual and for demonstration purposes, but they are probably not the right choice for the provision of tools to a whole department.

LIVE DISTRIBUTIONS

Live Linux distributions are a relatively new phenomenon and offer some big advantages. You don't have to install anything to run them. Just slot the CD or DVD into the drive and boot your machine. Et voila! If the developers have done their jobs correctly, the software should be configured to run properly without any further configuration. Live distributions may appeal to people who want to try a system out, those who want to demonstrate software to others, or those who want a portable Linux system for their own purposes. It is unlikely, however, that a live distribution will suffice as your primary bioinformatics system if you want to undertake serious bioinformatics work.

FULL SYSTEMS

Full systems customised for bioinformatics work are offered freely by a number of groups. Installed systems are very flexible. Unlike a Live distribution, you can always add extra software and customise to your hearts content. The distributions reviewed here are available either by downloadable iso files (BioBrew and BioLand) or by network installation (Bio-Linux). Currently, BioBrew is the only distribution of the three reviewed that can also be purchased on DVD.

By nature a certain degree of knowledge is required for maintaining a machine running Linux, with the level required varying between the systems reviewed here. For example, if you are a biologist with little computing or systems knowledge, but you require access to a high

Bio-Linux is FLOSS

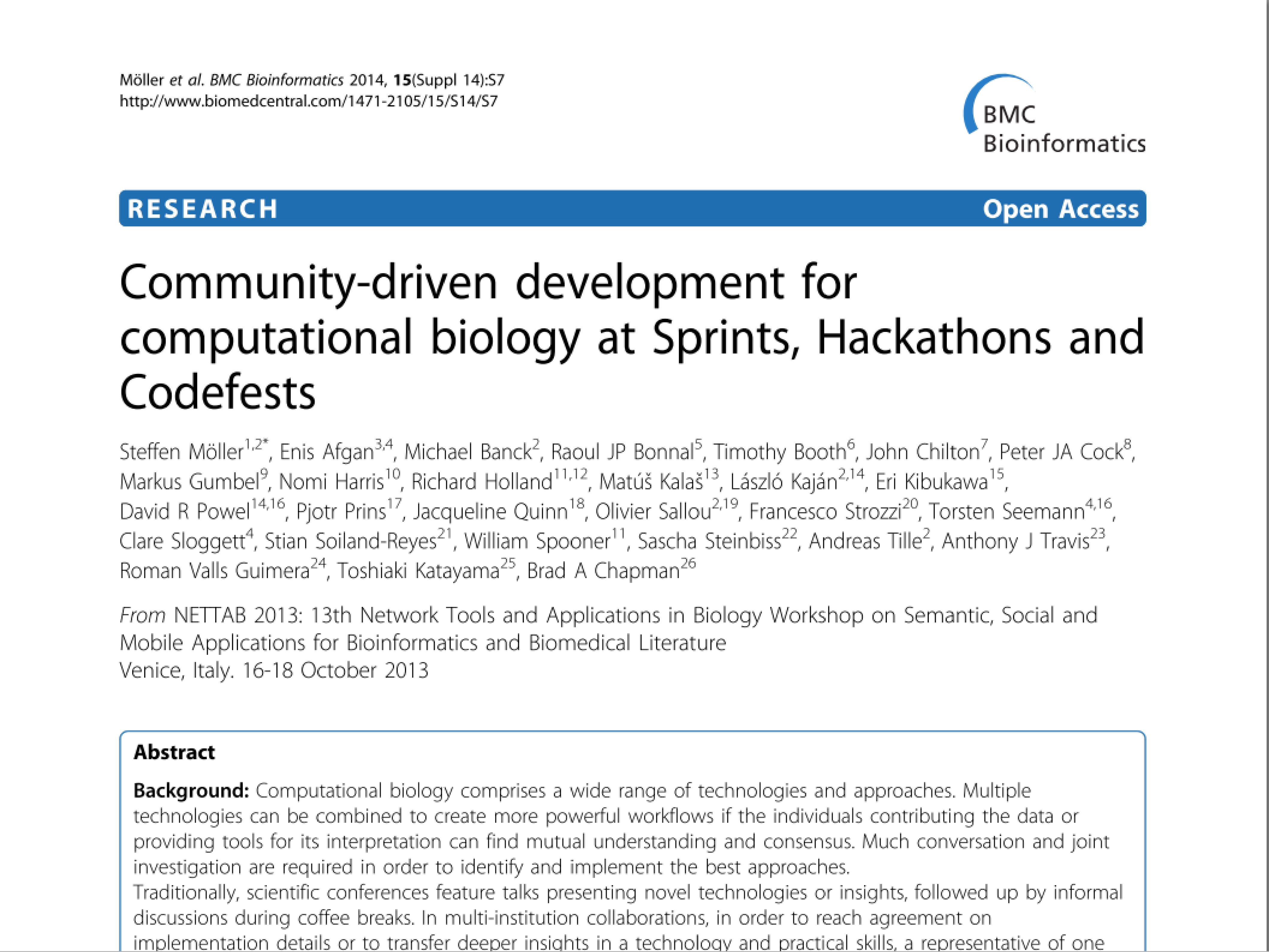
- Free - as in beer
- *Libre* - as in speech
- Open
- Source
- Software

Does FLOSS matter?

- Yes!
- Free software is software that gives you the user the freedom to share, study and modify it. We call this free software because the user is free

Why does this matter for bioinformatics?

- Intellectual freedom is important in biology
- Share software with other people legally
- Develop new versions of old software legally



Möller *et al. BMC Bioinformatics* 2014, **15**(Suppl 14):S7
http://www.biomedcentral.com/1471-2105/15/S14/S7

BMC Bioinformatics

RESEARCH Open Access

Community-driven development for computational biology at Sprints, Hackathons and Codefests

Steffen Möller[1,2*], Enis Afgan[3,4], Michael Banck[2], Raoul JP Bonnal[5], Timothy Booth[6], John Chilton[7], Peter JA Cock[8], Markus Gumbel[9], Nomi Harris[10], Richard Holland[11,12], Matúš Kalaš[13], László Kaján[2,14], Eri Kibukawa[15], David R Powel[14,16], Pjotr Prins[17], Jacqueline Quinn[18], Olivier Sallou[2,19], Francesco Strozzi[20], Torsten Seemann[4,16], Clare Sloggett[4], Stian Soiland-Reyes[21], William Spooner[11], Sascha Steinbiss[22], Andreas Tille[2], Anthony J Travis[23], Roman Valls Guimera[24], Toshiaki Katayama[25], Brad A Chapman[26]

From NETTAB 2013: 13th Network Tools and Applications in Biology Workshop on Semantic, Social and Mobile Applications for Bioinformatics and Biomedical Literature
Venice, Italy. 16-18 October 2013

Abstract

Background: Computational biology comprises a wide range of technologies and approaches. Multiple technologies can be combined to create more powerful workflows if the individuals contributing the data or providing tools for its interpretation can find mutual understanding and consensus. Much conversation and joint investigation are required in order to identify and implement the best approaches.
Traditionally, scientific conferences feature talks presenting novel technologies or insights, followed up by informal discussions during coffee breaks. In multi-institution collaborations, in order to reach agreement on implementation details or to transfer deeper insights in a technology and practical skills, a representative of one

Bio-Linux training and support

- Bioinformatics 'core' services
  - Typically overstretched and under-resourced
  - Better to teach biologists about bioinformatics
    - Biologists are advised how to analyse their data
    - Biologists better understand their own analysis
- Training environment
  - Based on guided self-study
  - Workshop or training course

2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), 578–82. IEEE.

Bio-Linux as a Tool for Bioinformatics Training

Timothy Booth, Mesude Bicak*, Hyun Soon Gweon,
Dawn Field
Molecular Evolution and Bioinformatics Group
NERC Centre for Ecology and Hydrology
Wallingford, United Kingdom
tbooth@ceh.ac.uk, mbicak@ceh.ac.uk, hyugwe@ceh.ac.uk,
dfield@ceh.ac.uk

Enis Afgan
Center for Informatics and Computing
Ruđer Bošković Institute
Zagreb, Croatia
enis.afgan@irb.hr

***Abstract*—Because of the ever-increasing application of next-generation sequencing (NGS) in research, and the expectation of faster experiment turn-around, it is becoming unfeasible and unscalable for analysis to be done exclusively by existing trained bioinformaticians. Instead, researchers and bench biologists are performing at least parts of most analyses. In order for this to be realized, two conditions must be satisfied: (1) well designed and accessible tools need to be made available, and (2) researchers and biologists need to be trained to use such tools in order to confidently handle high volumes of NGS data. Bio-Linux helps on both counts by offering well over one hundred bioinformatics tools packaged into a single distribution, easily accessible and readily usable. Bio-Linux is also accessible in the form of virtual images or on the cloud, thus providing researchers with immediate access to scalable compute infrastructure required to run the analysis. Furthermore this paper discusses how bioinformatics training on Bio-Linux is helping to bridge the data production and analysis gap.**

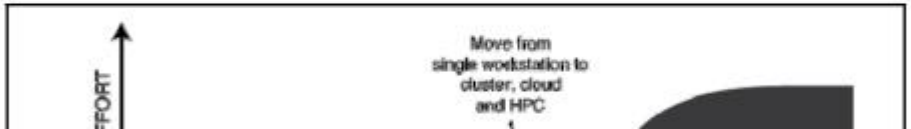
Keywords—bioinformatics; next-generation sequencing; training; cloud computing.

I. Introduction

deal with errors and subtleties in data and understand the tools, as well as their strengths/weaknesses for a given problem. But with the system set-up taken care of, the researcher is free to focus on these problems.

B. *Learning hurdles in bioinformatics*

Anyone wishing to develop bioinformatics skills and to analyse NGS data effectively faces many learning challenges. Here, we identify two particular hurdles - steep learning curves that a user must overcome to progress. As illustrated in Fig. 1, these are: 1) the move from manual, interactive data manipulation to programmatic manipulation, e.g. from manually editing a batch of files to writing a simple shell loop to make the edits, and 2) the move from working on a single system to working on a Grid or Cluster system, e.g. from running a big set of BLAST searches in series to splitting the query and submitting it to a compute cluster.



Bio-Linux 'live' DVD

- SystemImager
  - Used by NEBC for Bio-Linux network installation
    - Difficult to use on slow/unreliable (2GB download)
- BioKnoppix
  - Developed from Knoppix Debian 'live' DVD
    - Difficult to customise and impossible to upgrade
- Zen Linux
  - Knoppix derivative Debian 'live' DVD
    - Easy to customise and upgrade

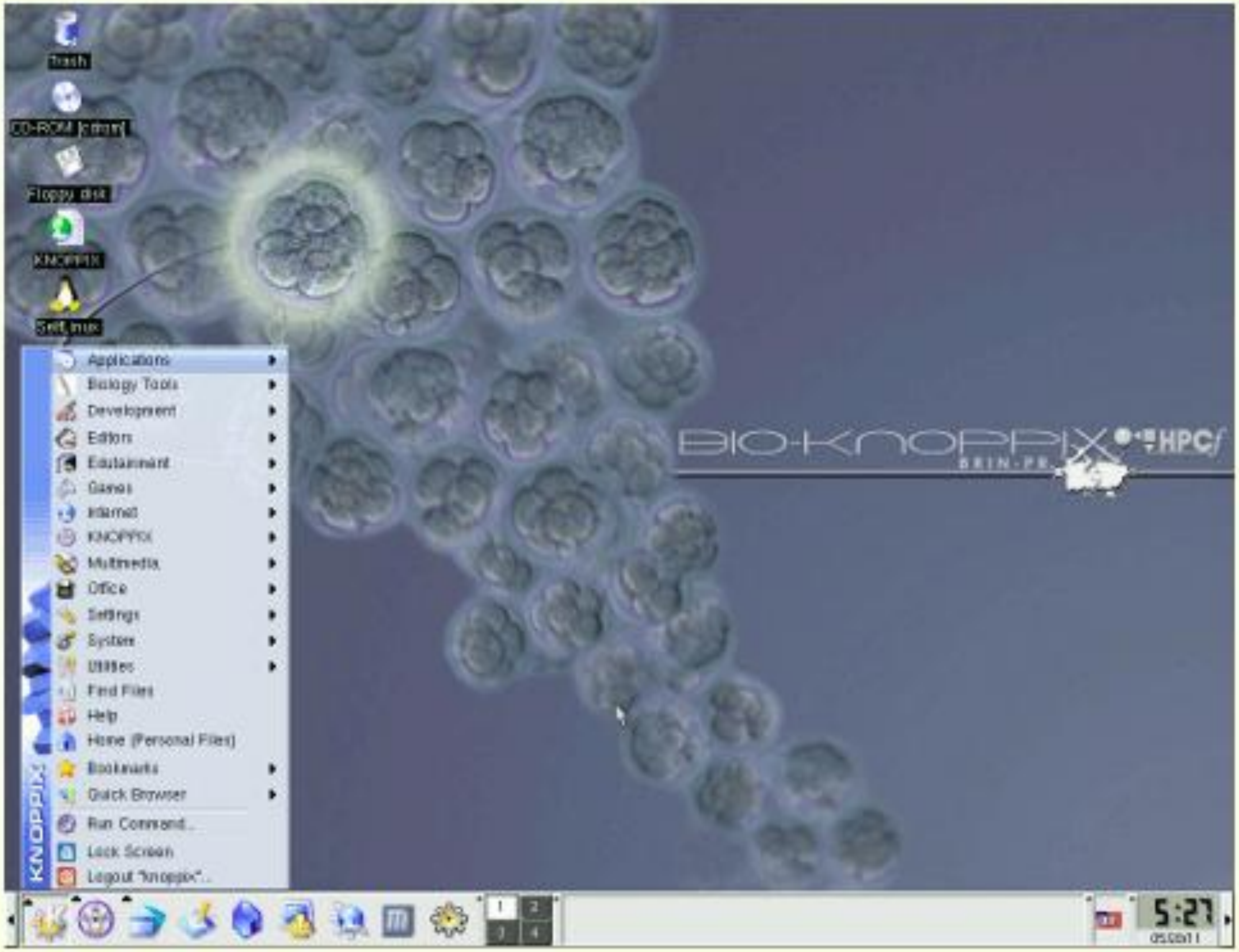
DistroWatch.com: Bioknoppix

Bioknoppix

Last Update: 2015-06-08 01:36 UTC

BIOKNOPPIX

- **OS Type:** Linux
- **Based on:** Debian, KNOPPIX
- **Origin:** Puerto Rico
- **Architecture:** i486
- **Desktop:** Fluxbox, Fluxbox, IceWM, IceWM, KDE, WMaker, Xfce
- **Category:** Live Medium
- **Status:** Discontinued
- **Popularity:** Not ranked



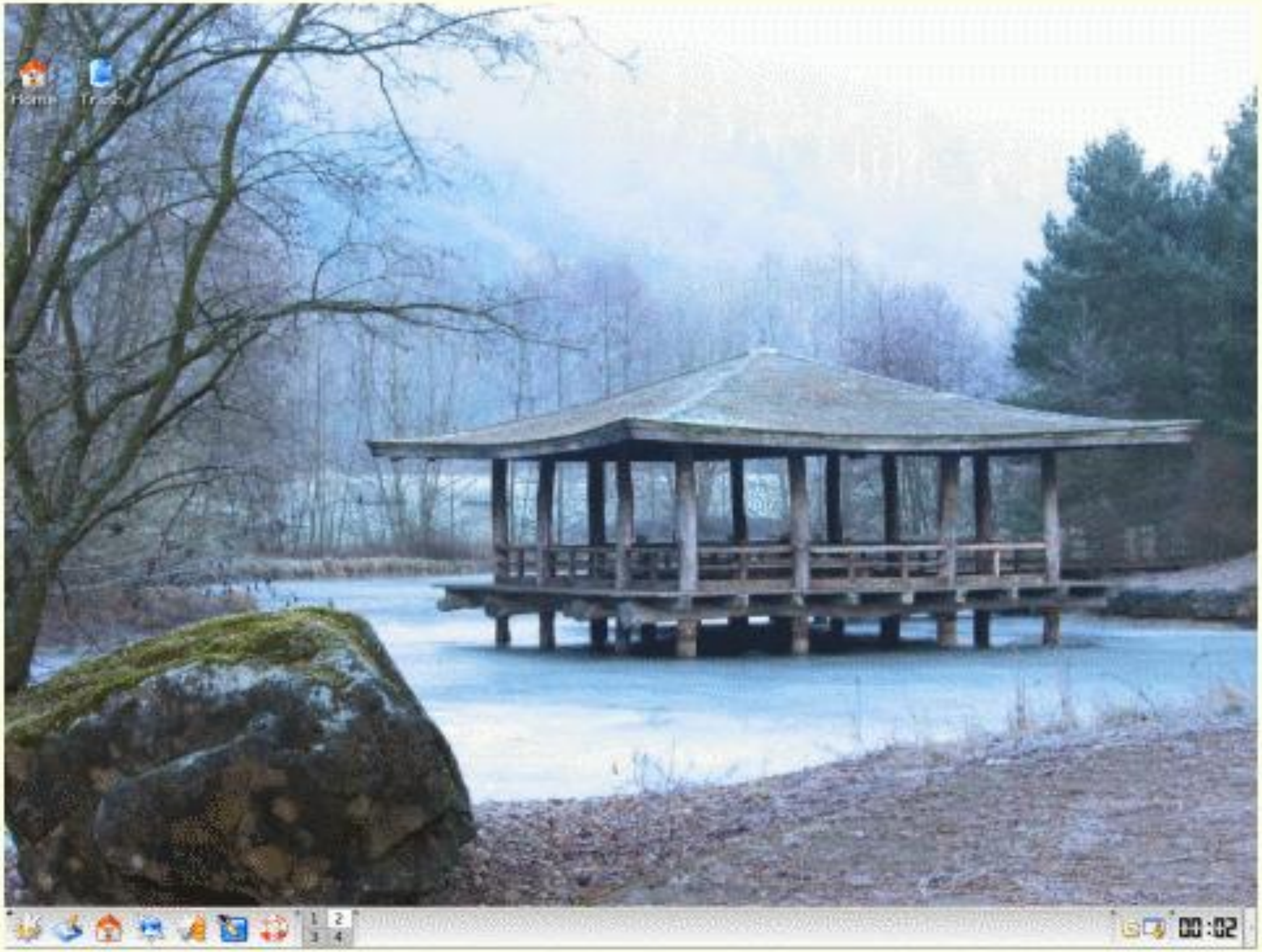
Bioknoppix is a customised distribution of the KNOPPIX live CD. With this distribution you just boot from the CD and you have a fully functional Linux OS with open source applications targeted at the molecular biologist. Besides using some RAM, Bioknoppix doesn't touch the host computer, being ideal for demonstrations, molecular biology students, workshops, etc.

DistroWatch.com: Zen Linux

Zen Linux

Last Update: 2015-06-08 01:36 UTC

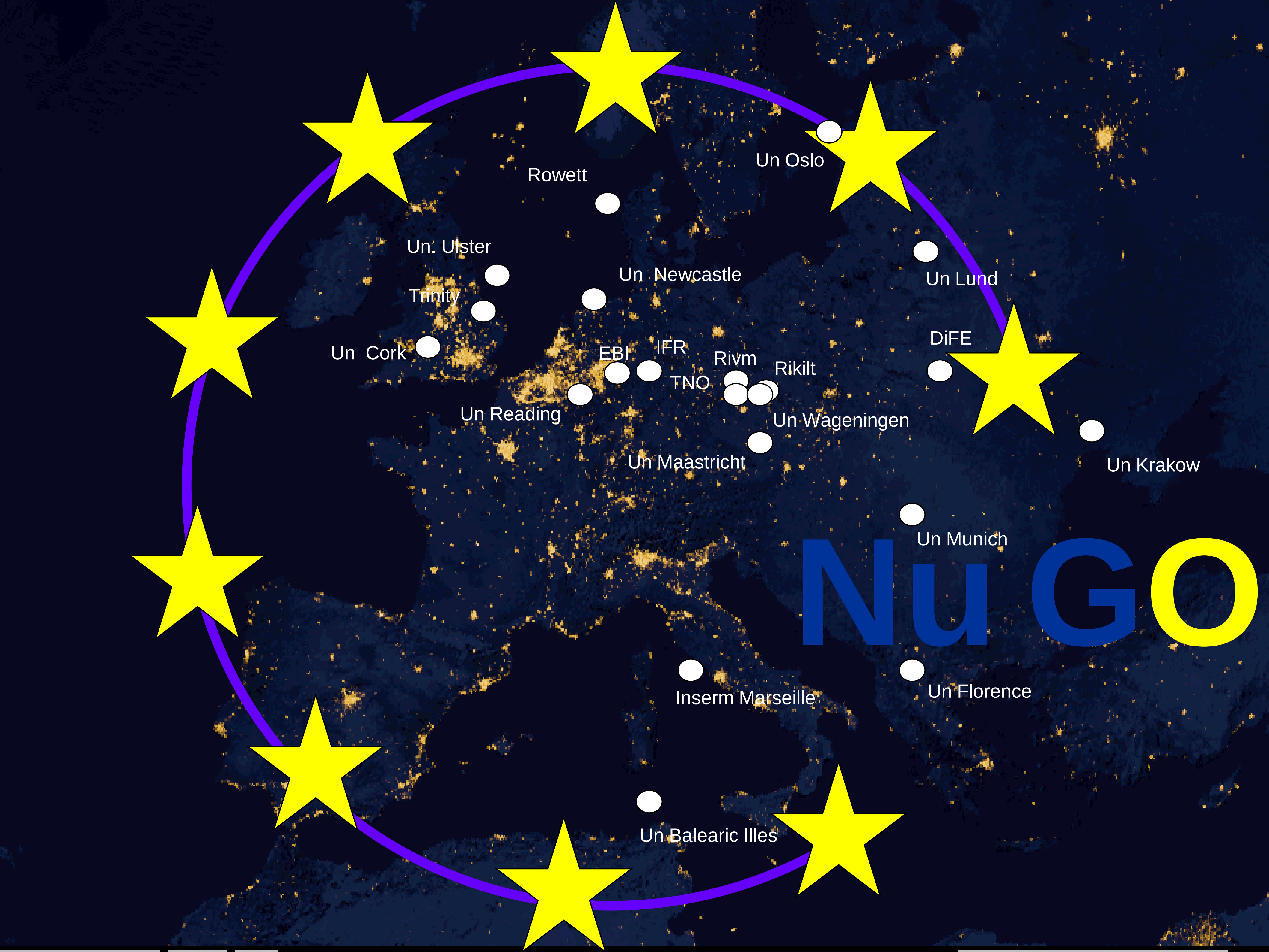
- **OS Type:** Linux
- **Based on:** Debian, KNOPPIX
- **Origin:** USA
- **Architecture:** i486
- **Desktop:** Fluxbox, KDE
- **Category:** Live Medium
- **Status:** Discontinued
- **Popularity:** Not ranked



Zen Linux is a bootable live CD distribution. Most configuration is done automatically upon boot and requires no user interaction. It includes the ability to to create remastered, personalised editions of the product.

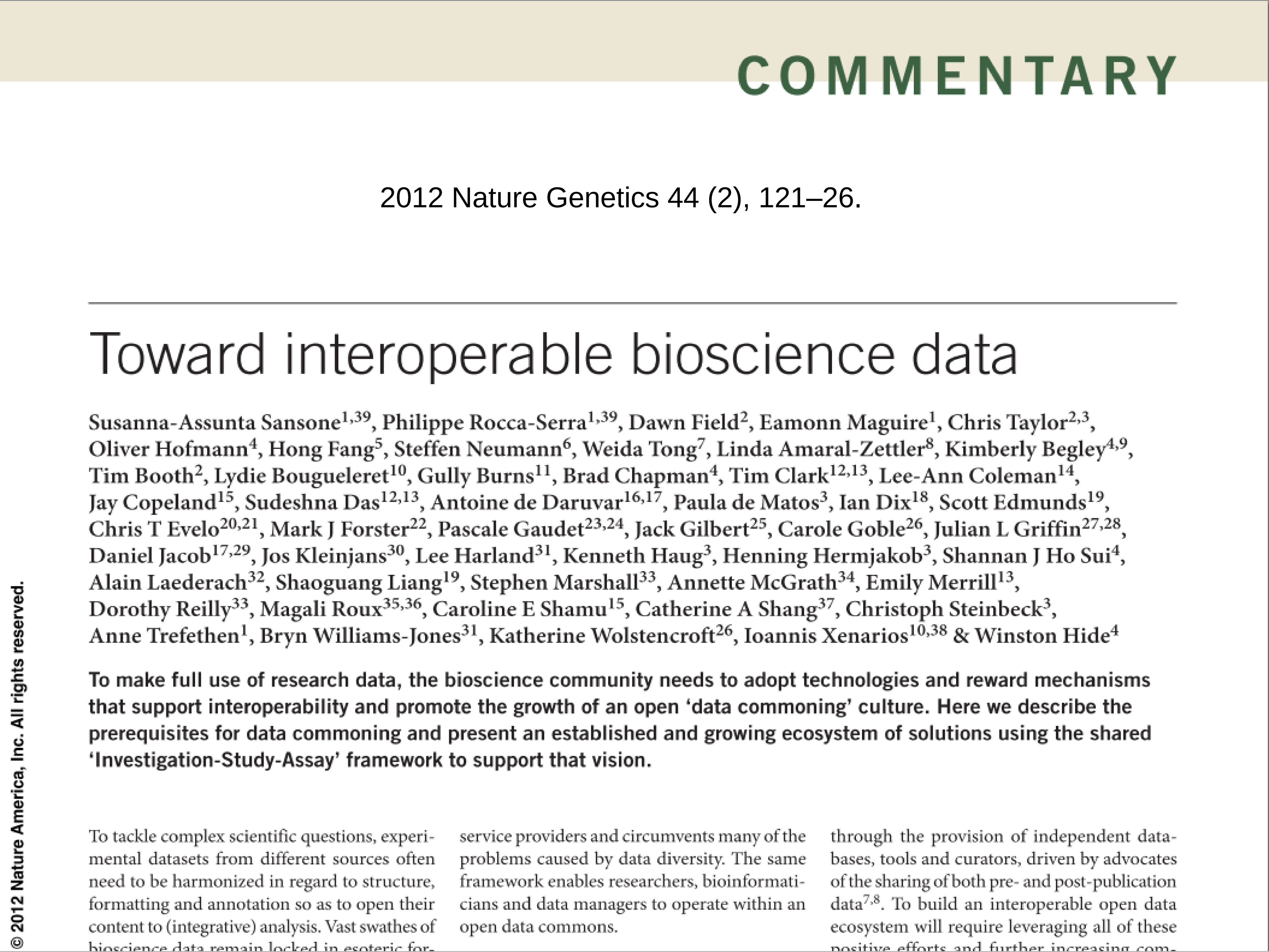
Zen Summary

Distribution	Zen Linux
Home Page	http://www.zenlinux.org/


Rowett
Un Oslo
Un. Ulster
Un Newcastle
Un Lund
Trinity
Un Cork
EBI
IFR
Rivm
Rikilt
DiFE
TNO
Un Reading
Un Wageningen
Un Maastricht
Un Krakow
Un Munich
NuGO
Inserm Marseille
Un Florence
Un Balearic Illes

Data sharing

- Data sharing
  - What is possible?
  - What needs to be improved?
- Manage expectations
  - Network latency
  - Storage capacity



COMMENTARY

2012 Nature Genetics 44 (2), 121–26.

Toward interoperable bioscience data

Susanna-Assunta Sansone[1,39], Philippe Rocca-Serra[1,39], Dawn Field[2], Eamonn Maguire[1], Chris Taylor[2,3], Oliver Hofmann[4], Hong Fang[5], Steffen Neumann[6], Weida Tong[7], Linda Amaral-Zettler[8], Kimberly Begley[4,9], Tim Booth[2], Lydie Bougueleret[10], Gully Burns[11], Brad Chapman[4], Tim Clark[12,13], Lee-Ann Coleman[14], Jay Copeland[15], Sudeshna Das[12,13], Antoine de Daruvar[16,17], Paula de Matos[3], Ian Dix[18], Scott Edmunds[19], Chris T Evelo[20,21], Mark J Forster[22], Pascale Gaudet[23,24], Jack Gilbert[25], Carole Goble[26], Julian L Griffin[27,28], Daniel Jacob[17,29], Jos Kleinjans[30], Lee Harland[31], Kenneth Haug[3], Henning Hermjakob[3], Shannan J Ho Sui[4], Alain Laederach[32], Shaoguang Liang[19], Stephen Marshall[33], Annette McGrath[34], Emily Merrill[13], Dorothy Reilly[33], Magali Roux[35,36], Caroline E Shamu[15], Catherine A Shang[37], Christoph Steinbeck[3], Anne Trefethen[1], Bryn Williams-Jones[31], Katherine Wolstencroft[26], Ioannis Xenarios[10,38] & Winston Hide[4]

To make full use of research data, the bioscience community needs to adopt technologies and reward mechanisms that support interoperability and promote the growth of an open 'data commoning' culture. Here we describe the prerequisites for data commoning and present an established and growing ecosystem of solutions using the shared 'Investigation-Study-Assay' framework to support that vision.

To tackle complex scientific questions, experimental datasets from different sources often need to be harmonized in regard to structure, formatting and annotation so as to open their content to (integrative) analysis. Vast swathes of bioscience data remain locked in esoteric for-

service providers and circumvents many of the problems caused by data diversity. The same framework enables researchers, bioinformaticians and data managers to operate within an open data commons.

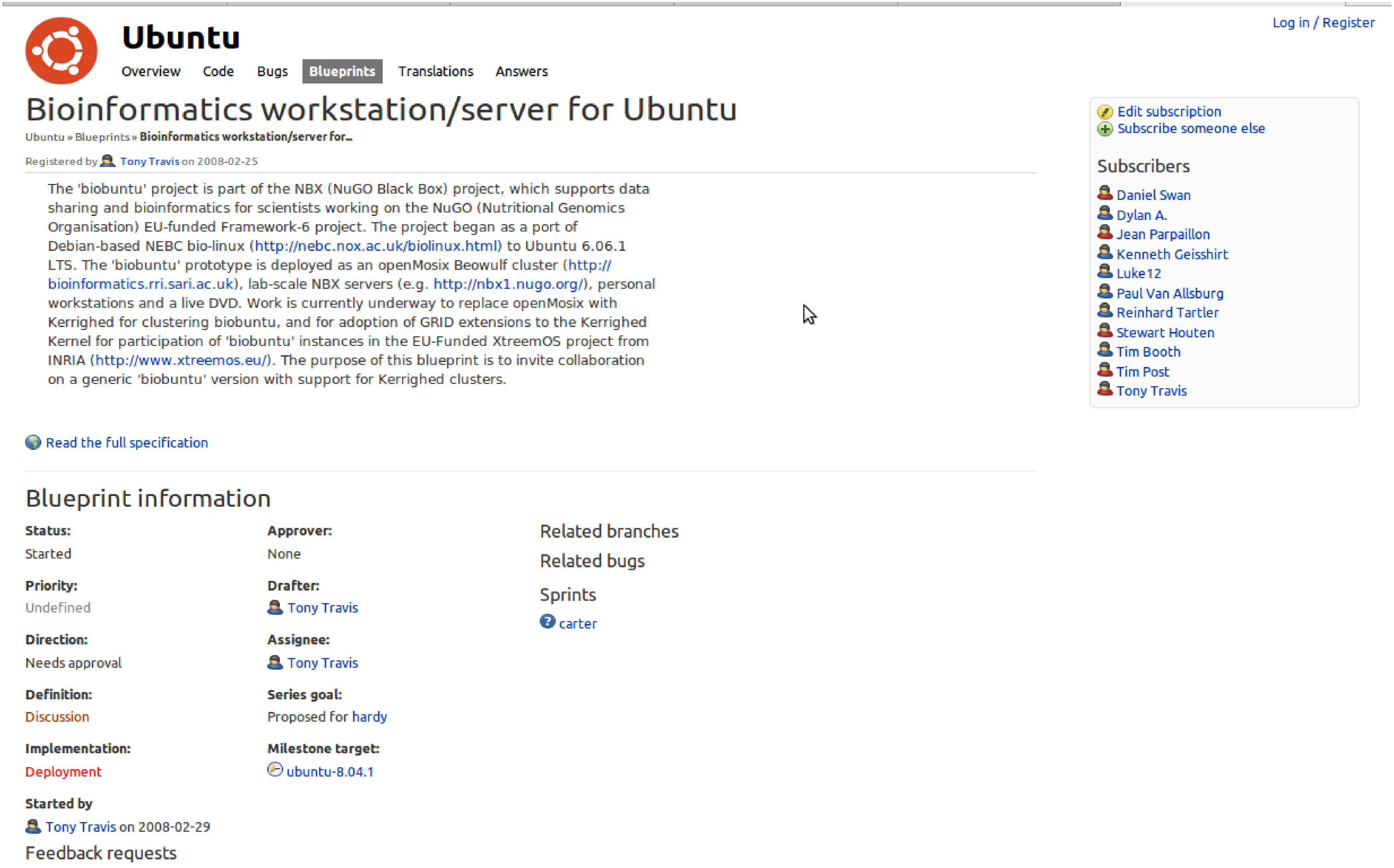
through the provision of independent databases, tools and curators, driven by advocates of the sharing of both pre- and post-publication data[7,8]. To build an interoperable open data ecosystem will require leveraging all of these positive efforts and further increasing com-

Bio-Linux + Ubuntu 'biobuntu'

- NEBC (UK) Bio-Linux4 (Debian “Sarge”)
  - NERC Environmental Bioinformatics Centre
- Ubuntu 6.06.1 LTS operating system
  - Open source software based on Debian Linux

+

=



Bioinformatics workstation/server for Ubuntu

Ubuntu » Blueprints » **Bioinformatics workstation/server for...**

Registered by Tony Travis on 2008-02-25

The 'biobuntu' project is part of the NBX (NuGO Black Box) project, which supports data sharing and bioinformatics for scientists working on the NuGO (Nutritional Genomics Organisation) EU-funded Framework-6 project. The project began as a port of Debian-based NEBC bio-linux (http://nebc.nox.ac.uk/biolinux.html) to Ubuntu 6.06.1 LTS. The 'biobuntu' prototype is deployed as an openMosix Beowulf cluster (http://bioinformatics.rri.sari.ac.uk), lab-scale NBX servers (e.g. http://nbx1.nugo.org/), personal workstations and a live DVD. Work is currently underway to replace openMosix with Kerrighed for clustering biobuntu, and for adoption of GRID extensions to the Kerrighed Kernel for participation of 'biobuntu' instances in the EU-Funded XtreemOS project from INRIA (http://www.xtreemos.eu/). The purpose of this blueprint is to invite collaboration on a generic 'biobuntu' version with support for Kerrighed clusters.

Read the full specification

Blueprint information

Status:
Started

Priority:
Undefined

Direction:
Needs approval

Definition:
Discussion

Implementation:
Deployment

Started by
Tony Travis on 2008-02-29

Approver:
None

Drafter:
Tony Travis

Assignee:
Tony Travis

Series goal:
Proposed for hardy

Milestone target:
ubuntu-8.04.1

Related branches

Related bugs

Sprints

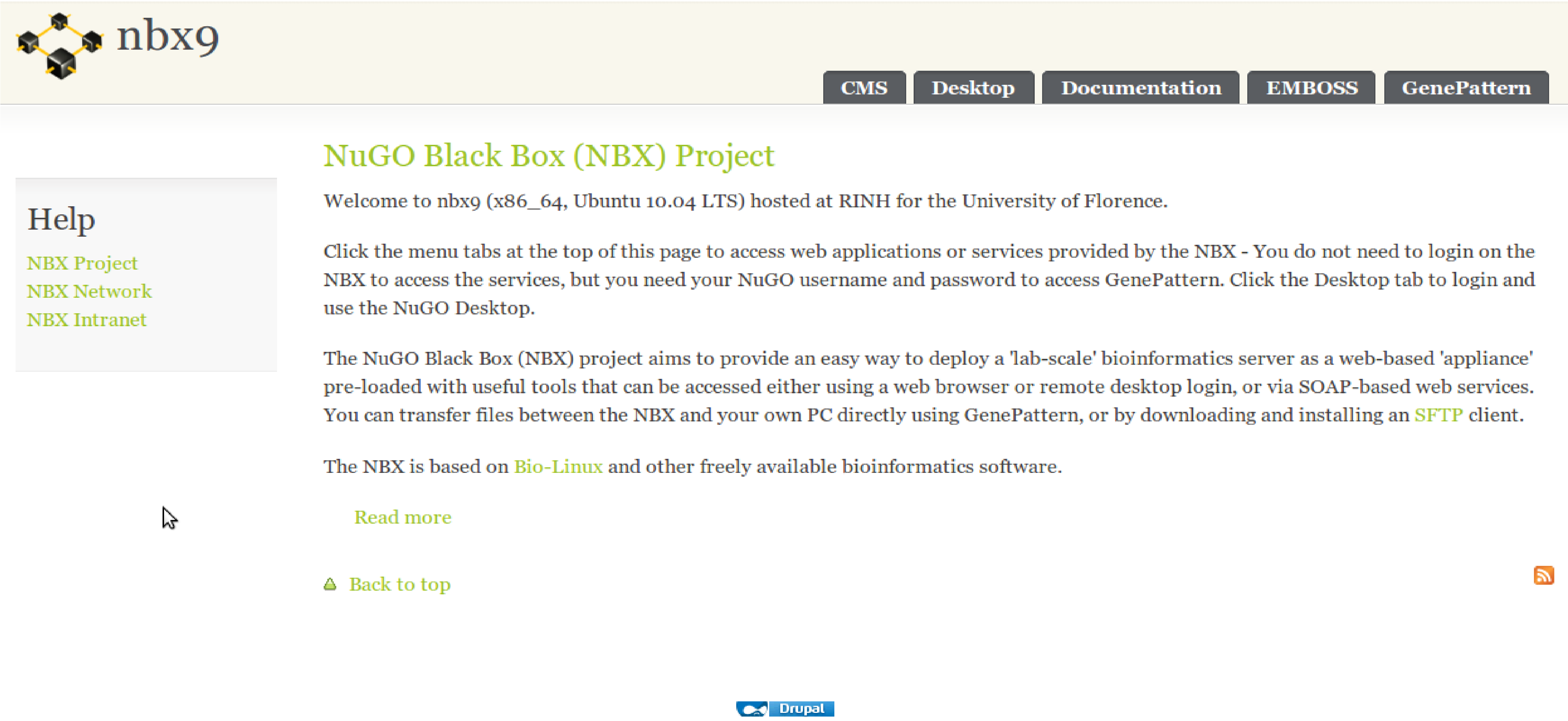
carter

Feedback requests

Edit subscription
Subscribe someone else

Subscribers

Daniel Swan
Dylan A.
Jean Parpaillon
Kenneth Geisshirt
Luke12
Paul Van Allsburg
Reinhard Tartler
Stewart Houten
Tim Booth
Tim Post
Tony Travis



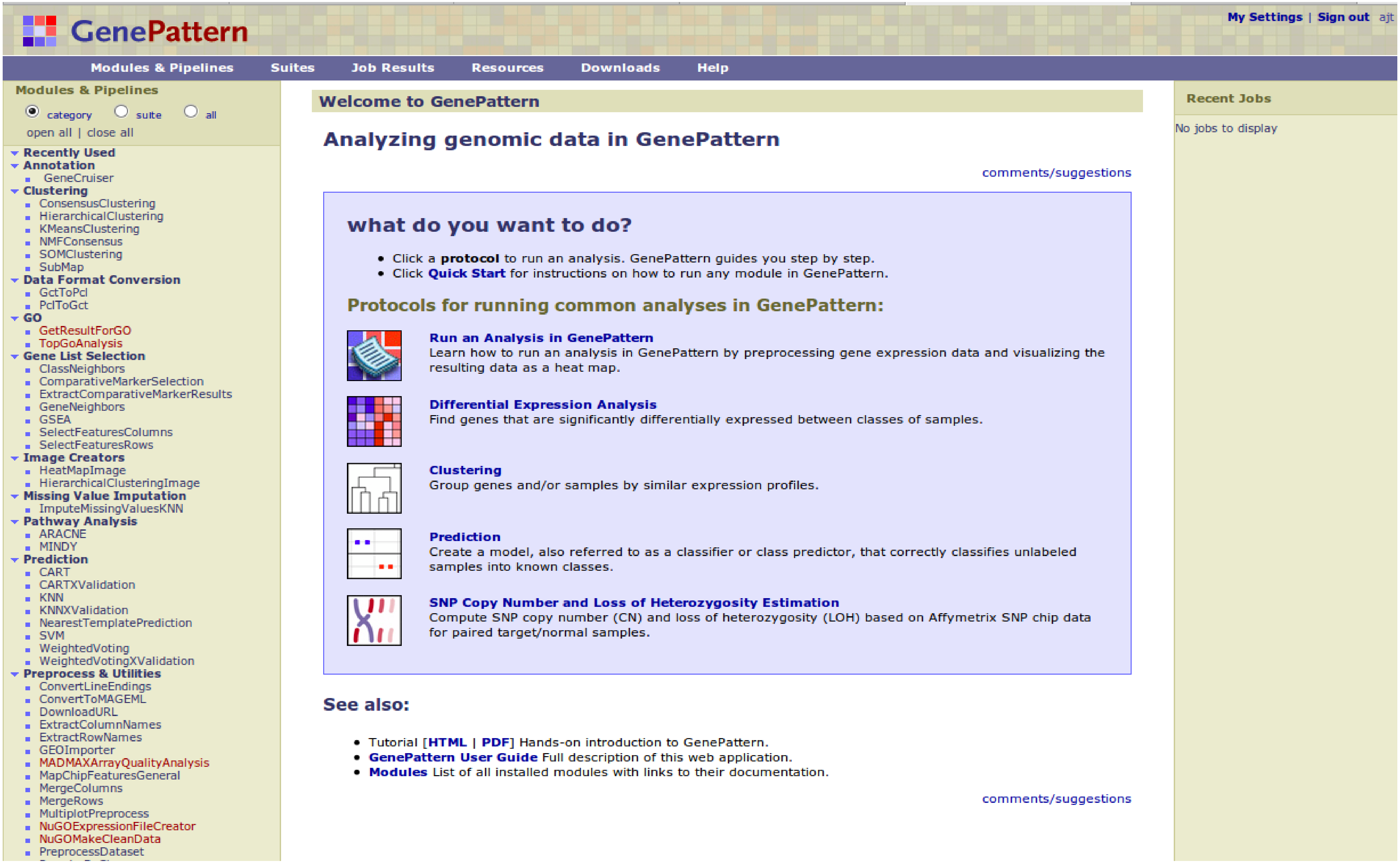
NuGO Black Box (NBX) Project

Welcome to nbx9 (x86_64, Ubuntu 10.04 LTS) hosted at RINH for the University of Florence.

Click the menu tabs at the top of this page to access web applications or services provided by the NBX - You do not need to login on the NBX to access the services, but you need your NuGO username and password to access GenePattern. Click the Desktop tab to login and use the NuGO Desktop.

The NuGO Black Box (NBX) project aims to provide an easy way to deploy a 'lab-scale' bioinformatics server as a web-based 'appliance' pre-loaded with useful tools that can be accessed either using a web browser or remote desktop login, or via SOAP-based web services. You can transfer files between the NBX and your own PC directly using GenePattern, or by downloading and installing an SFTP client.

The NBX is based on Bio-Linux and other freely available bioinformatics software.



Modules & Pipelines

● category ○ suite ○ all

open all | close all

- **Recently Used**
- **Annotation**
  - GeneCruiser
- **Clustering**
  - ConsensusClustering
  - HierarchicalClustering
  - KMeansClustering
  - NMFConsensus
  - SOMClustering
  - SubMap
- **Data Format Conversion**
  - GctToPcl
  - PclToGct
- **GO**
  - GetResultForGO
  - TopGoAnalysis
- **Gene List Selection**
  - ClassNeighbors
  - ComparativeMarkerSelection
  - ExtractComparativeMarkerResults
  - GeneNeighbors
  - GSEA
  - SelectFeaturesColumns
  - SelectFeaturesRows
- **Image Creators**
  - HeatMapImage
  - HierarchicalClusteringImage
- **Missing Value Imputation**
  - ImputeMissingValuesKNN
- **Pathway Analysis**
  - ARACNE
  - MINDY
- **Prediction**
  - CART
  - CARTXValidation
  - KNN
  - KNNXValidation
  - NearestTemplatePrediction
  - SVM
  - WeightedVoting
  - WeightedVotingXValidation
- **Preprocess & Utilities**
  - ConvertLineEndings
  - ConvertToMAGEML
  - DownloadURL
  - ExtractColumnNames
  - ExtractRowNames
  - GEOImporter
  - MADMAXArrayQualityAnalysis
  - MapChipFeaturesGeneral
  - MergeColumns
  - MergeRows
  - MultiplotPreprocess
  - NuGOExpressionFileCreator
  - NuGOMakeCleanData
  - PreprocessDataset

Welcome to GenePattern

Analyzing genomic data in GenePattern

comments/suggestions

what do you want to do?

- Click a **protocol** to run an analysis. GenePattern guides you step by step.
- Click **Quick Start** for instructions on how to run any module in GenePattern.

Protocols for running common analyses in GenePattern:

Run an Analysis in GenePattern
Learn how to run an analysis in GenePattern by preprocessing gene expression data and visualizing the resulting data as a heat map.

Differential Expression Analysis
Find genes that are significantly differentially expressed between classes of samples.

Clustering
Group genes and/or samples by similar expression profiles.

Prediction
Create a model, also referred to as a classifier or class predictor, that correctly classifies unlabeled samples into known classes.

SNP Copy Number and Loss of Heterozygosity Estimation
Compute SNP copy number (CN) and loss of heterozygosity (LOH) based on Affymetrix SNP chip data for paired target/normal samples.

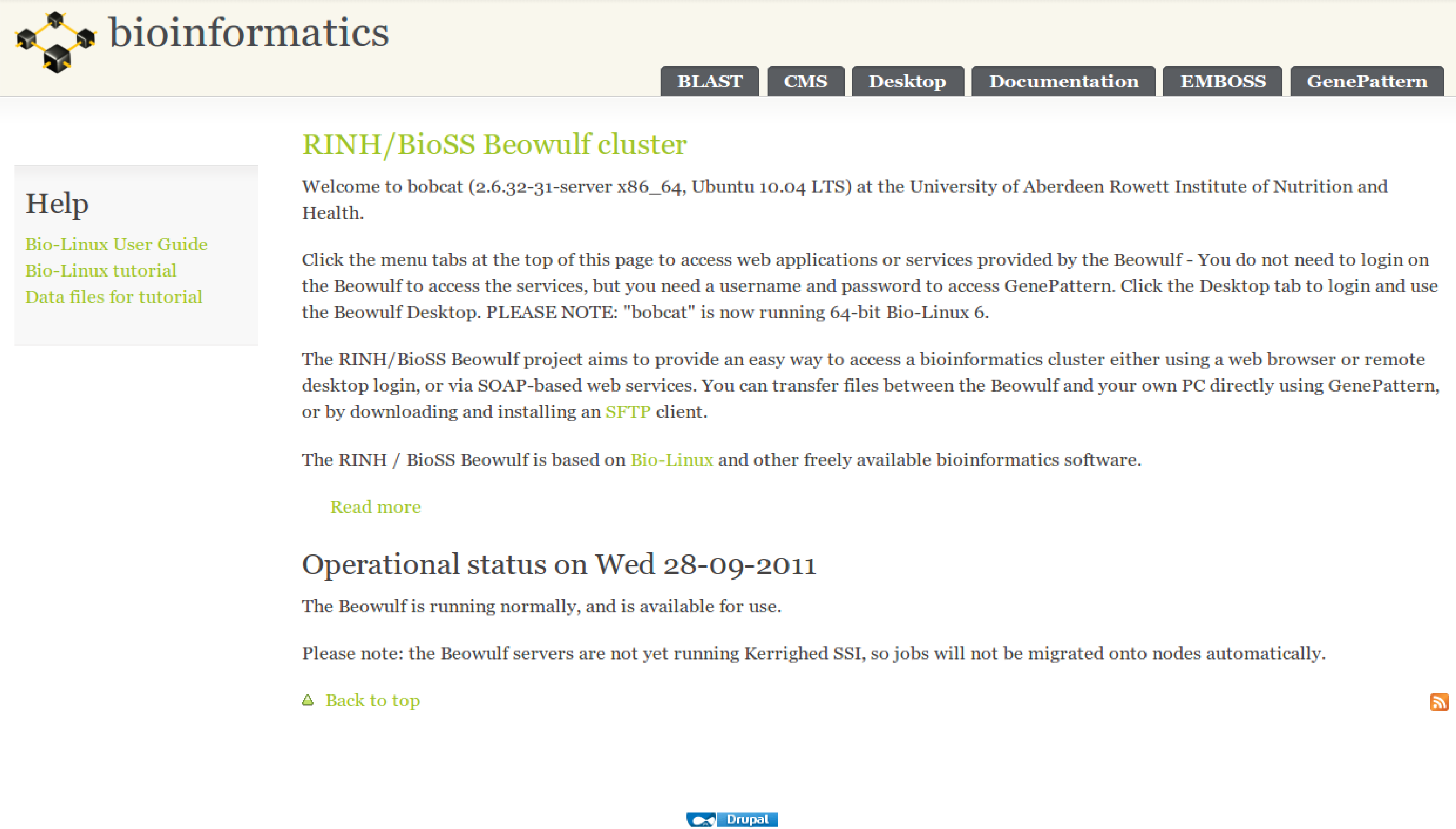
See also:

- Tutorial [**HTML** | **PDF**] Hands-on introduction to GenePattern.
- **GenePattern User Guide** Full description of this web application.
- **Modules** List of all installed modules with links to their documentation.

comments/suggestions

Recent Jobs

No jobs to display



RINH/BioSS Beowulf cluster

Welcome to bobcat (2.6.32-31-server x86_64, Ubuntu 10.04 LTS) at the University of Aberdeen Rowett Institute of Nutrition and Health.

Click the menu tabs at the top of this page to access web applications or services provided by the Beowulf - You do not need to login on the Beowulf to access the services, but you need a username and password to access GenePattern. Click the Desktop tab to login and use the Beowulf Desktop. PLEASE NOTE: "bobcat" is now running 64-bit Bio-Linux 6.

The RINH/BioSS Beowulf project aims to provide an easy way to access a bioinformatics cluster either using a web browser or remote desktop login, or via SOAP-based web services. You can transfer files between the Beowulf and your own PC directly using GenePattern, or by downloading and installing an SFTP client.

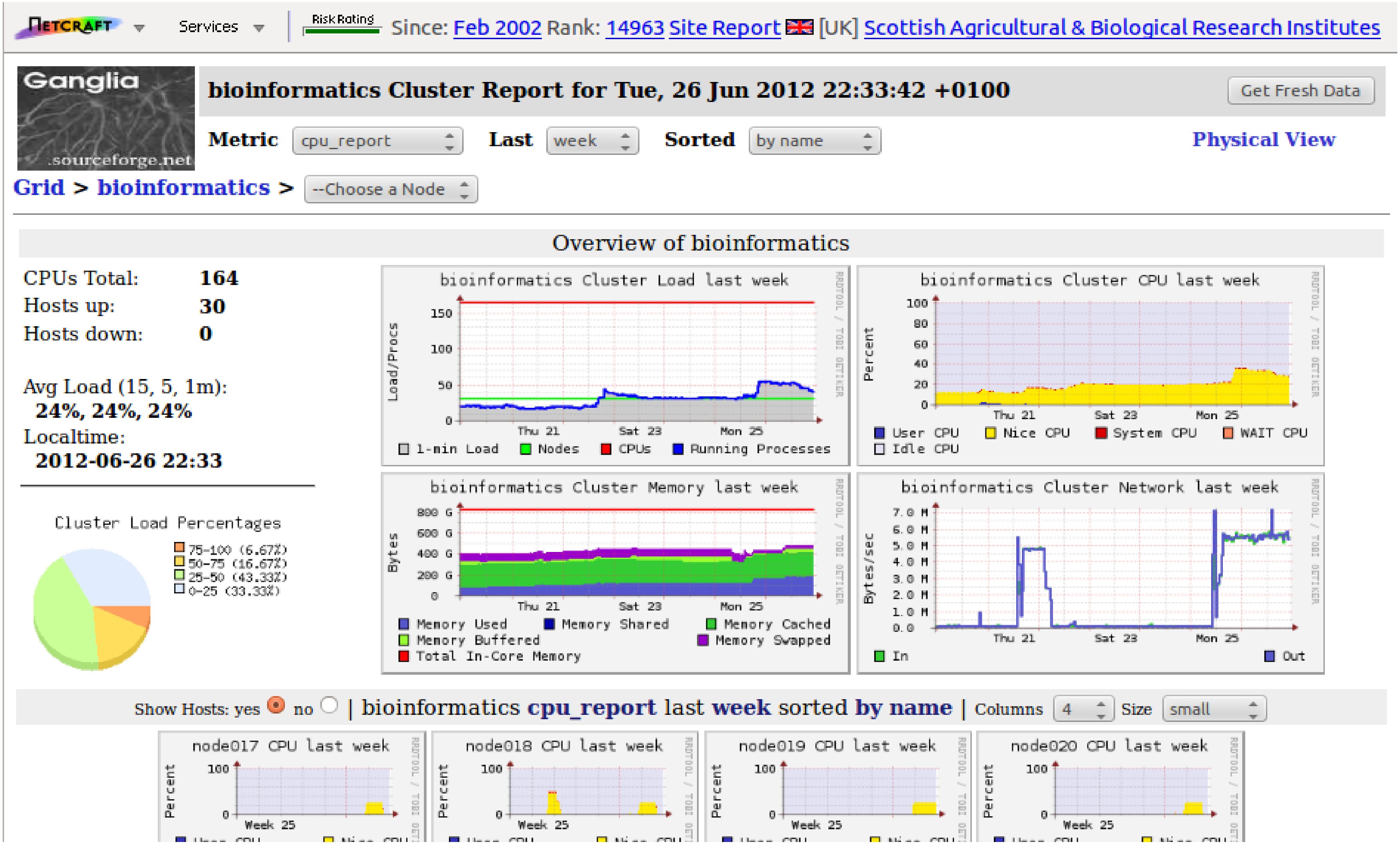
The RINH / BioSS Beowulf is based on Bio-Linux and other freely available bioinformatics software.

Read more

Operational status on Wed 28-09-2011

The Beowulf is running normally, and is available for use.

Please note: the Beowulf servers are not yet running Kerrighed SSI, so jobs will not be migrated onto nodes automatically.



Since: Feb 2002 Rank: 14963 Site Report [UK] Scottish Agricultural & Biological Research Institutes

bioinformatics Cluster Report for Tue, 26 Jun 2012 22:33:42 +0100

Metric cpu_report **Last** week **Sorted** by name **Physical View**

Grid > bioinformatics > --Choose a Node

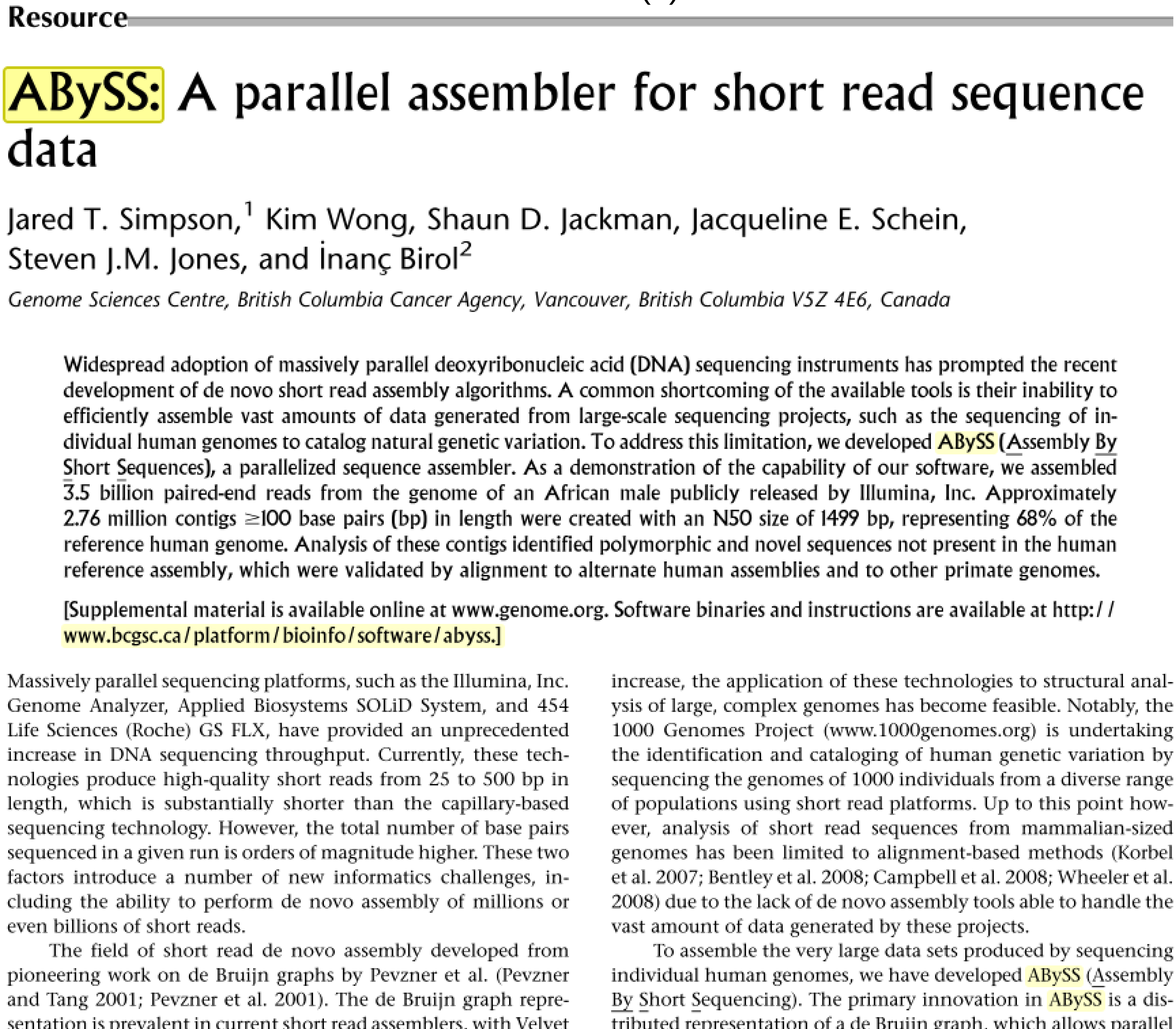
Overview of bioinformatics

CPUs Total: **164**
Hosts up: **30**
Hosts down: **0**

Avg Load (15, 5, 1m):
24%, 24%, 24%
Localtime:
2012-06-26 22:33

Show Hosts: yes ● no ○ | bioinformatics **cpu_report** last **week** sorted **by name** | Columns 4 Size small

Genome Research 19 (6): 1117–23.



Resource

ABySS: A parallel assembler for short read sequence data

Jared T. Simpson,[1] Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and İnanç Birol[2]

Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada

Widespread adoption of massively parallel deoxyribonucleic acid (DNA) sequencing instruments has prompted the recent development of de novo short read assembly algorithms. A common shortcoming of the available tools is their inability to efficiently assemble vast amounts of data generated from large-scale sequencing projects, such as the sequencing of individual human genomes to catalog natural genetic variation. To address this limitation, we developed ABySS (Assembly By Short Sequences), a parallelized sequence assembler. As a demonstration of the capability of our software, we assembled 3.5 billion paired-end reads from the genome of an African male publicly released by Illumina, Inc. Approximately 2.76 million contigs ≥100 base pairs (bp) in length were created with an N50 size of 1499 bp, representing 68% of the reference human genome. Analysis of these contigs identified polymorphic and novel sequences not present in the human reference assembly, which were validated by alignment to alternate human assemblies and to other primate genomes.

[Supplemental material is available online at www.genome.org. Software binaries and instructions are available at http://www.bcgsc.ca/platform/bioinfo/software/abyss.]

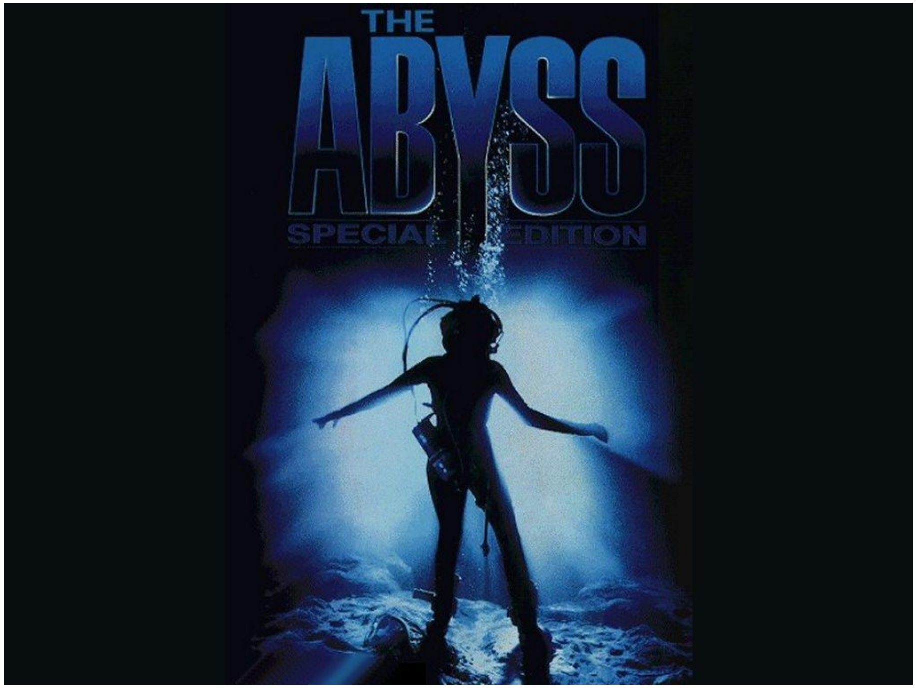
Massively parallel sequencing platforms, such as the Illumina, Inc. Genome Analyzer, Applied Biosystems SOLiD System, and 454 Life Sciences (Roche) GS FLX, have provided an unprecedented increase in DNA sequencing throughput. Currently, these technologies produce high-quality short reads from 25 to 500 bp in length, which is substantially shorter than the capillary-based sequencing technology. However, the total number of base pairs sequenced in a given run is orders of magnitude higher. These two factors introduce a number of new informatics challenges, including the ability to perform de novo assembly of millions or even billions of short reads.

The field of short read de novo assembly developed from pioneering work on de Bruijn graphs by Pevzner et al. (Pevzner and Tang 2001; Pevzner et al. 2001). The de Bruijn graph representation is prevalent in current short read assemblers, with Velvet increase, the application of these technologies to structural analysis of large, complex genomes has become feasible. Notably, the 1000 Genomes Project (www.1000genomes.org) is undertaking the identification and cataloging of human genetic variation by sequencing the genomes of 1000 individuals from a diverse range of populations using short read platforms. Up to this point however, analysis of short read sequences from mammalian-sized genomes has been limited to alignment-based methods (Korbel et al. 2007; Bentley et al. 2008; Campbell et al. 2008; Wheeler et al. 2008) due to the lack of de novo assembly tools able to handle the vast amount of data generated by these projects.

To assemble the very large data sets produced by sequencing individual human genomes, we have developed ABySS (Assembly By Short Sequencing). The primary innovation in ABySS is a distributed representation of a de Bruijn graph, which allows parallel

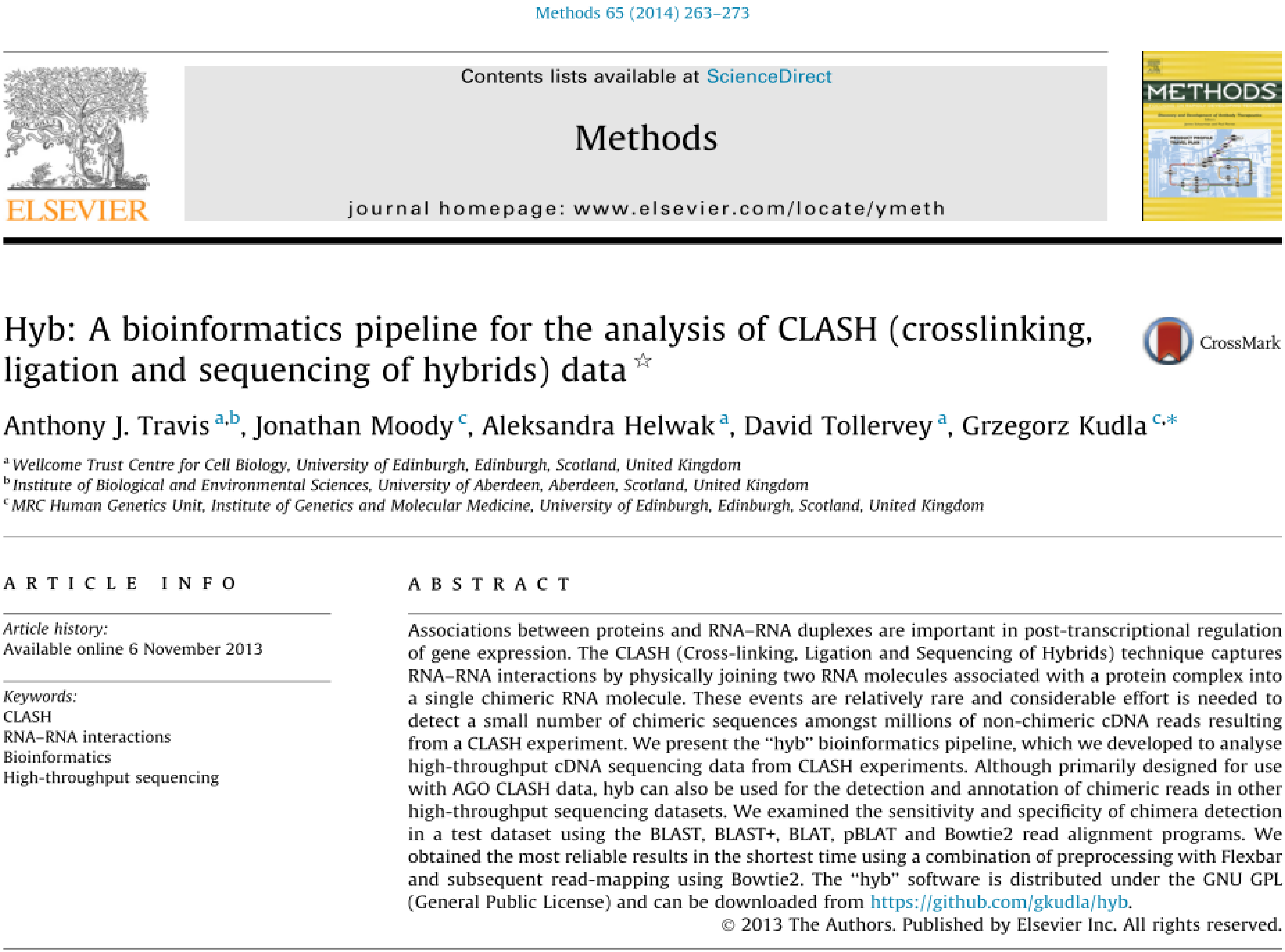
ABySS assember based on *make*

- Assembly By Short Sequences
- *de novo*, parallel, paired-end sequence assembler
- AbySS *pipeline* is implemented as an executable 'Makefile'



make bioinformatics easier!

- *GNU 'make'*
  - Unix + GNU/Linux
  - Software utility
- Builds projects
  - According to *rules*
  - File *dependencies*
  - Concurrent *workflow*



Methods 65 (2014) 263–273

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Hyb: A bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data ☆

Anthony J. Travis [a,b], Jonathan Moody [c], Aleksandra Helwak [a], David Tollervey [a], Grzegorz Kudla [c,*]

[a] *Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh, Scotland, United Kingdom*
[b] *Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, Scotland, United Kingdom*
[c] *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Scotland, United Kingdom*

ARTICLE INFO

Article history:
Available online 6 November 2013

Keywords:
CLASH
RNA–RNA interactions
Bioinformatics
High-throughput sequencing

ABSTRACT

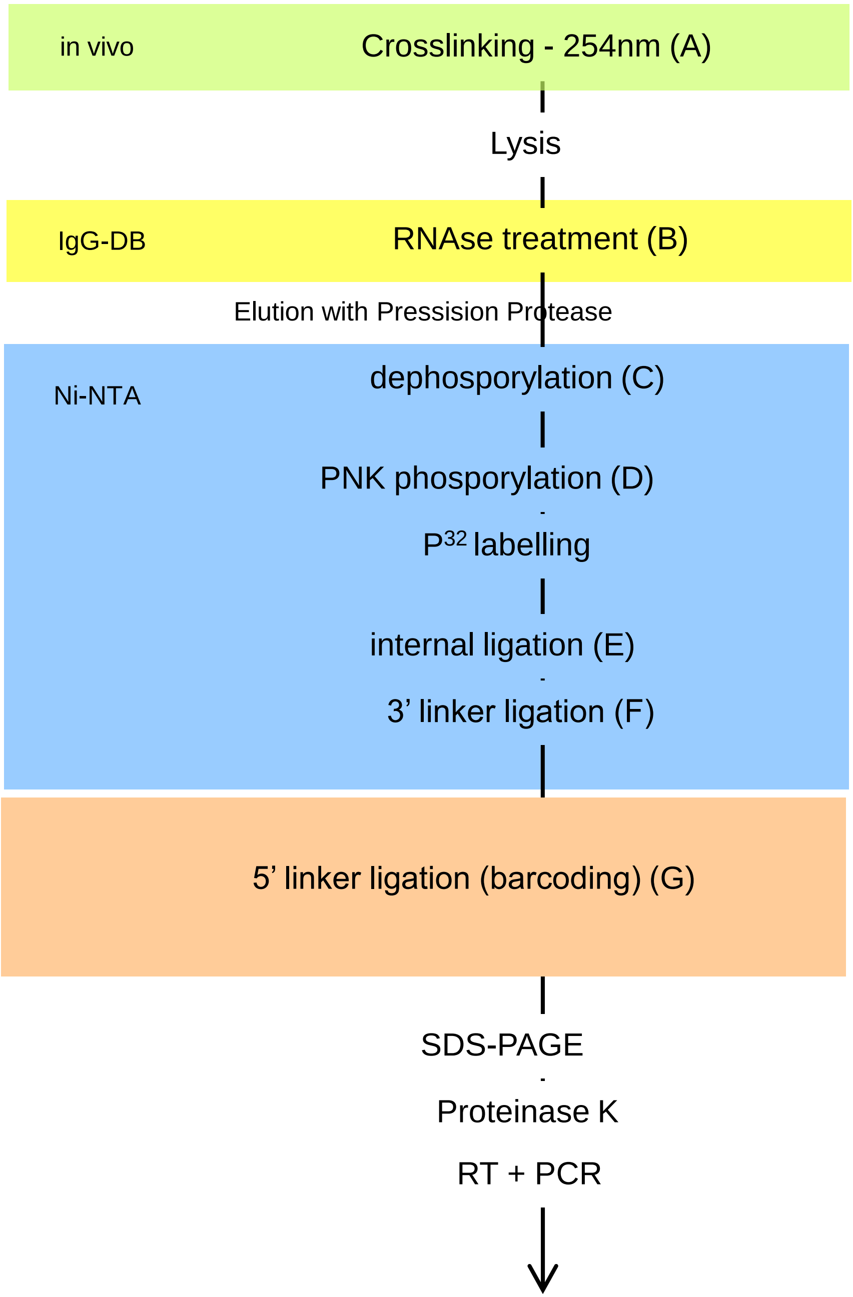
Associations between proteins and RNA–RNA duplexes are important in post-transcriptional regulation of gene expression. The CLASH (Cross-linking, Ligation and Sequencing of Hybrids) technique captures RNA–RNA interactions by physically joining two RNA molecules associated with a protein complex into a single chimeric RNA molecule. These events are relatively rare and considerable effort is needed to detect a small number of chimeric sequences amongst millions of non-chimeric cDNA reads resulting from a CLASH experiment. We present the "hyb" bioinformatics pipeline, which we developed to analyse high-throughput cDNA sequencing data from CLASH experiments. Although primarily designed for use with AGO CLASH data, hyb can also be used for the detection and annotation of chimeric reads in other high-throughput sequencing datasets. We examined the sensitivity and specificity of chimera detection in a test dataset using the BLAST, BLAST+, BLAT, pBLAT and Bowtie2 read alignment programs. We obtained the most reliable results in the shortest time using a combination of preprocessing with Flexbar and subsequent read-mapping using Bowtie2. The "hyb" software is distributed under the GNU GPL (General Public License) and can be downloaded from https://github.com/gkudla/hyb.

“hyb” - an executable Makefile

- Named “hyb” to avoid confusion with wet-lab CLASH
- GNU “make”
  - Machine *reasoning*
    - Makefile contains rules

```
target: dependencies
        actions
```

- *Orchestration*
  - Target *independence* allows *concurrency*
  - Avoids unnecessary re-analysis of results



in vivo

Crosslinking - 254nm (A)

Lysis

IgG-DB

RNAse treatment (B)

Elution with Pressision Protease

Ni-NTA

dephosporylation (C)

PNK phosporylation (D)

P^{32} labelling

internal ligation (E)

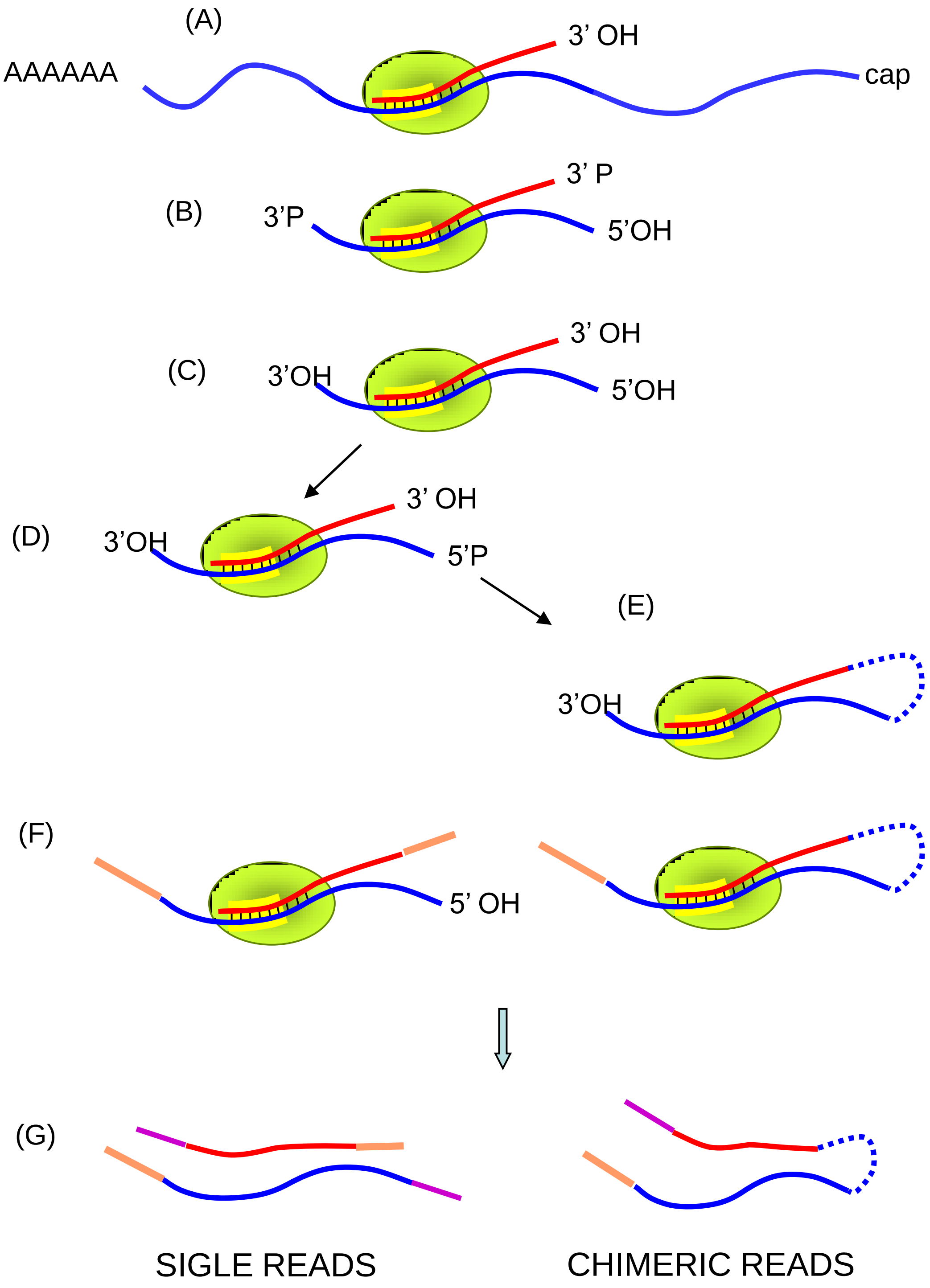
3’ linker ligation (F)

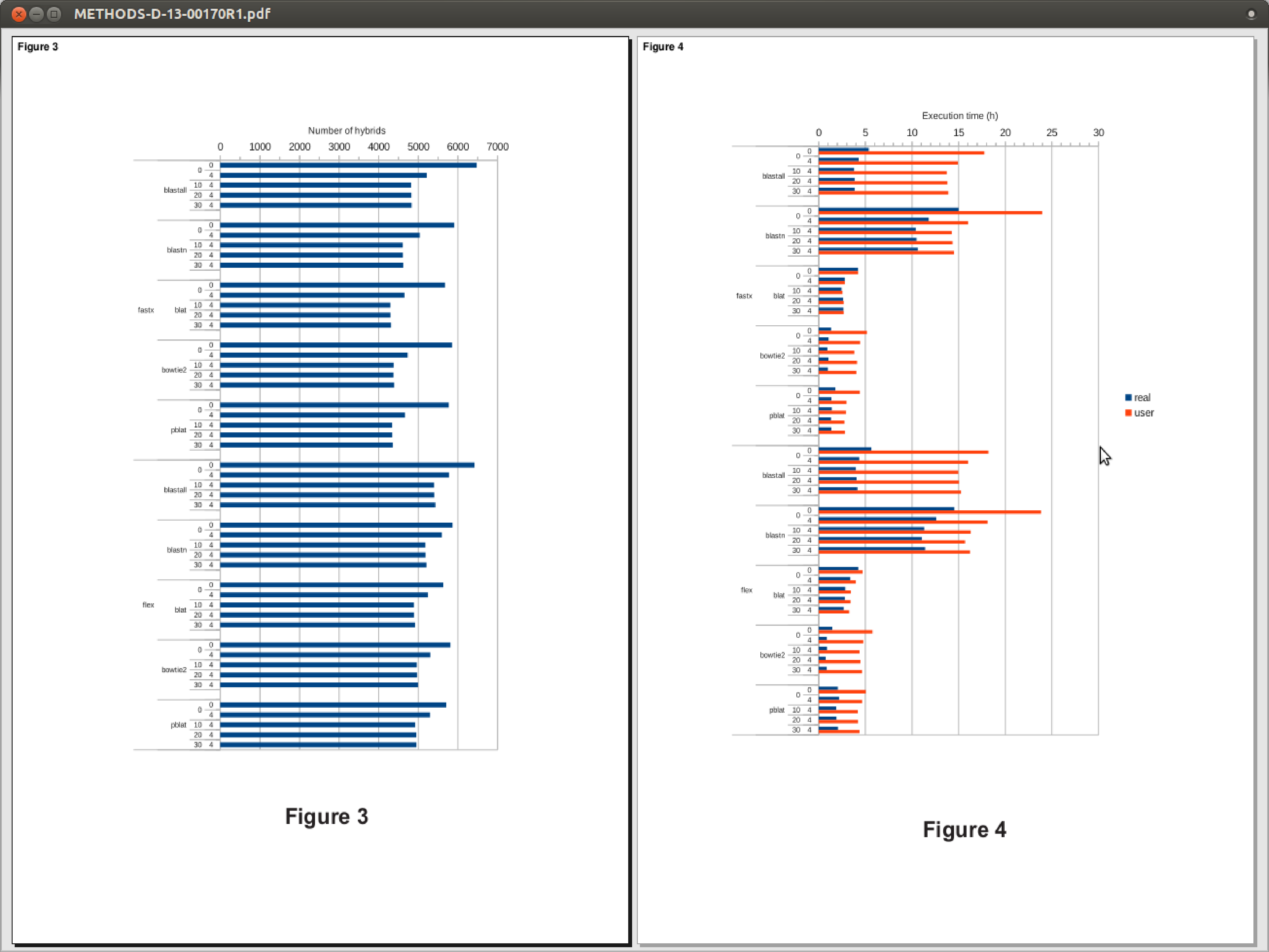
5’ linker ligation (barcoding) (G)

SDS-PAGE

Proteinase K

RT + PCR





METHODS-D-13-00170R1.pdf

Figure 3

Figure 3

Figure 4

Figure 4

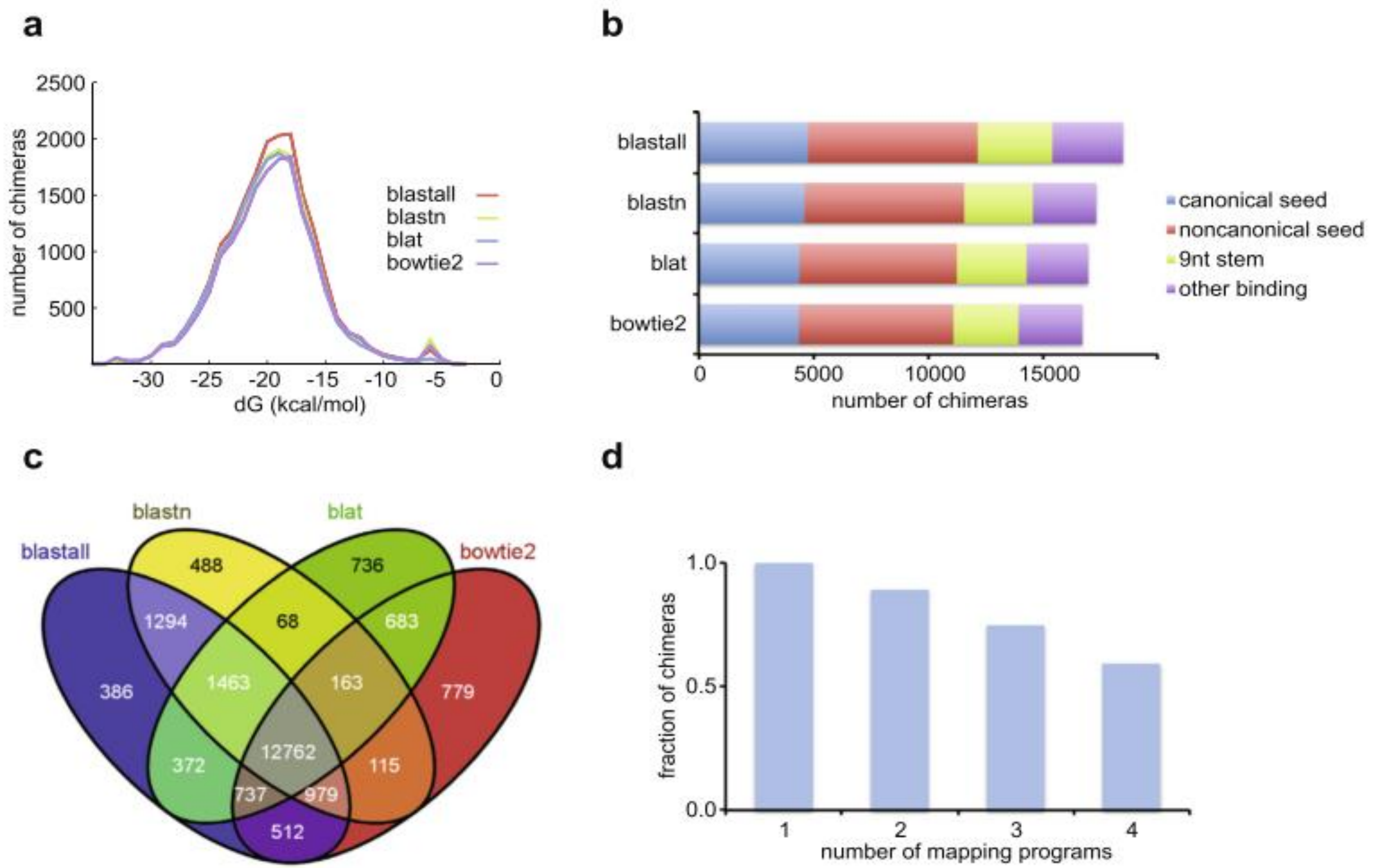


Fig. 5. Characteristics of chimeras recovered as a function of the mapping program used. (a) Distribution of folding energies of miRNA–mRNA chimeras identified with blastall, blastn, blat, and bowtie2. (b) Types of RNA–RNA interactions recovered with each mapping program. (c) Numbers of chimeras recovered with different combinations of mapping programs, analysed with VENNY [29]. A total of 12762 interactions are found with all four mapping programs, whereas 21537 interactions are found with at least one of the programs. (d) Fractions of chimeras recovered with one or more, two or more, three or more, and four mapping programs, respectively. Analyses were performed on dataset E4 (Ref. [11]), with the following parameters: trim = 0 filt = 0 min = 4 len = 17.

Make has been used in other bioinformatics pipelines. For example, the "PredictProtein" server [26] invokes Make programmatically by a Perl driver script to process jobs submitted via a

miRNA–mRNA interactions can be distinguished from false positives by the following characteristics:

(1) Average predicted folding energy of chimeras (stronger

So, how do we run Bio-Linux?

Boot the Bio-Linux USB-stick

- Try Bio-Linux out on your *own* laptop
- FLOSS platform for bioinformatics work
- Contact Tony Travis to obtain a Bio-Linux USB-stick at BLSW

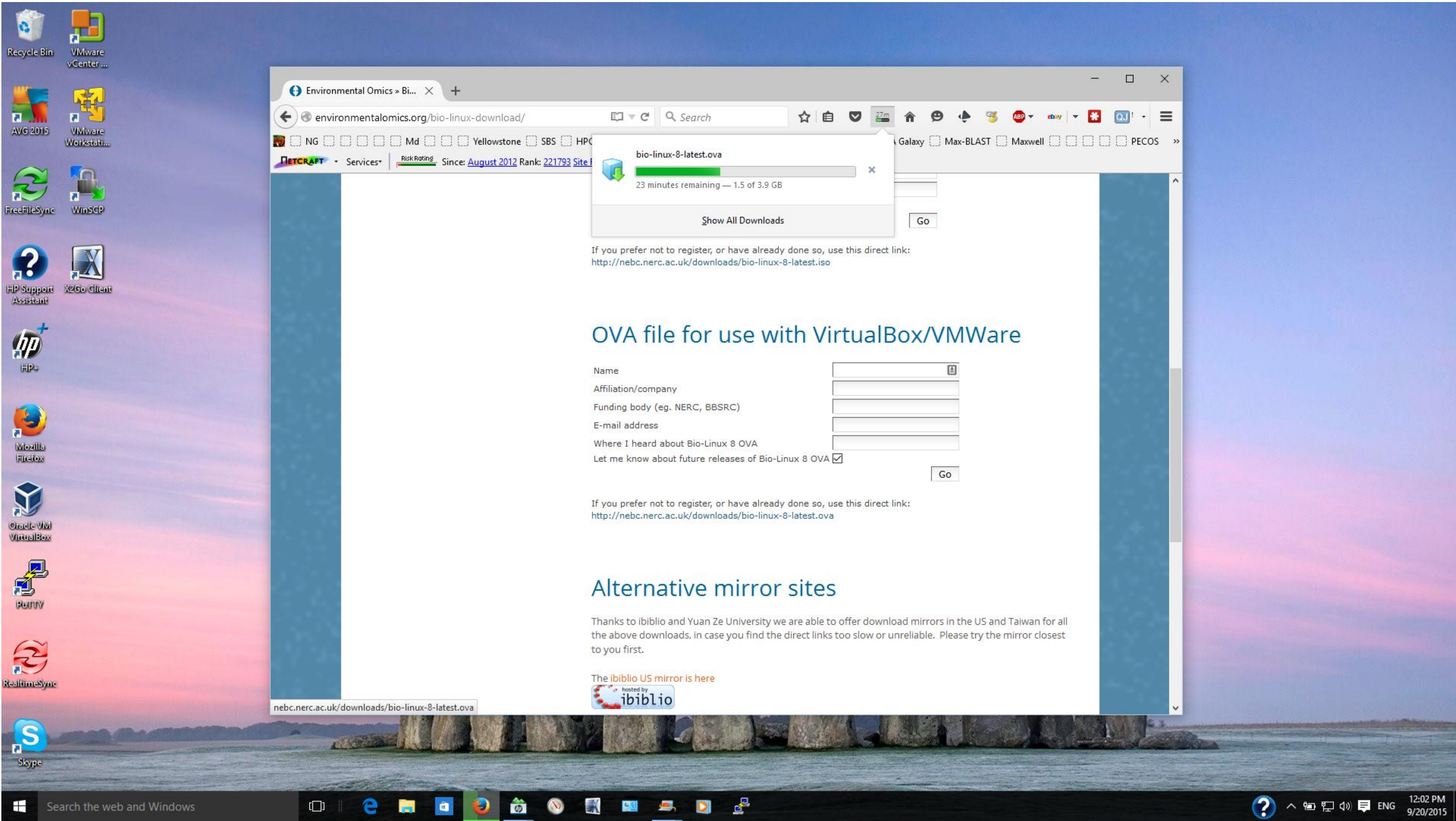
tony.travis@minke-informatics.co.uk

Run Bio-Linux under Windows

- Virtual Machine
  - Windows host
  - Bio-Linux guest
- Bio-Linux OVA
  - Hypervisor neutral
  - Vmware
  - VirtualBox

Download the Bio-Linux OVA file

http://environmentalomics.org/bio-linux-download/



bio-linux-8-latest.ova

23 minutes remaining — 1.5 of 3.9 GB

If you prefer not to register, or have already done so, use this direct link:
http://nebc.nerc.ac.uk/downloads/bio-linux-8-latest.iso

OVA file for use with VirtualBox/VMWare

Name

Affiliation/company

Funding body (eg. NERC, BBSRC)

E-mail address

Where I heard about Bio-Linux 8 OVA

Let me know about future releases of Bio-Linux 8 OVA ☑

Go

If you prefer not to register, or have already done so, use this direct link:
http://nebc.nerc.ac.uk/downloads/bio-linux-8-latest.ova

Alternative mirror sites

Thanks to ibiblio and Yuan Ze University we are able to offer download mirrors in the US and Taiwan for all the above downloads, in case you find the direct links too slow or unreliable. Please try the mirror closest to you first.

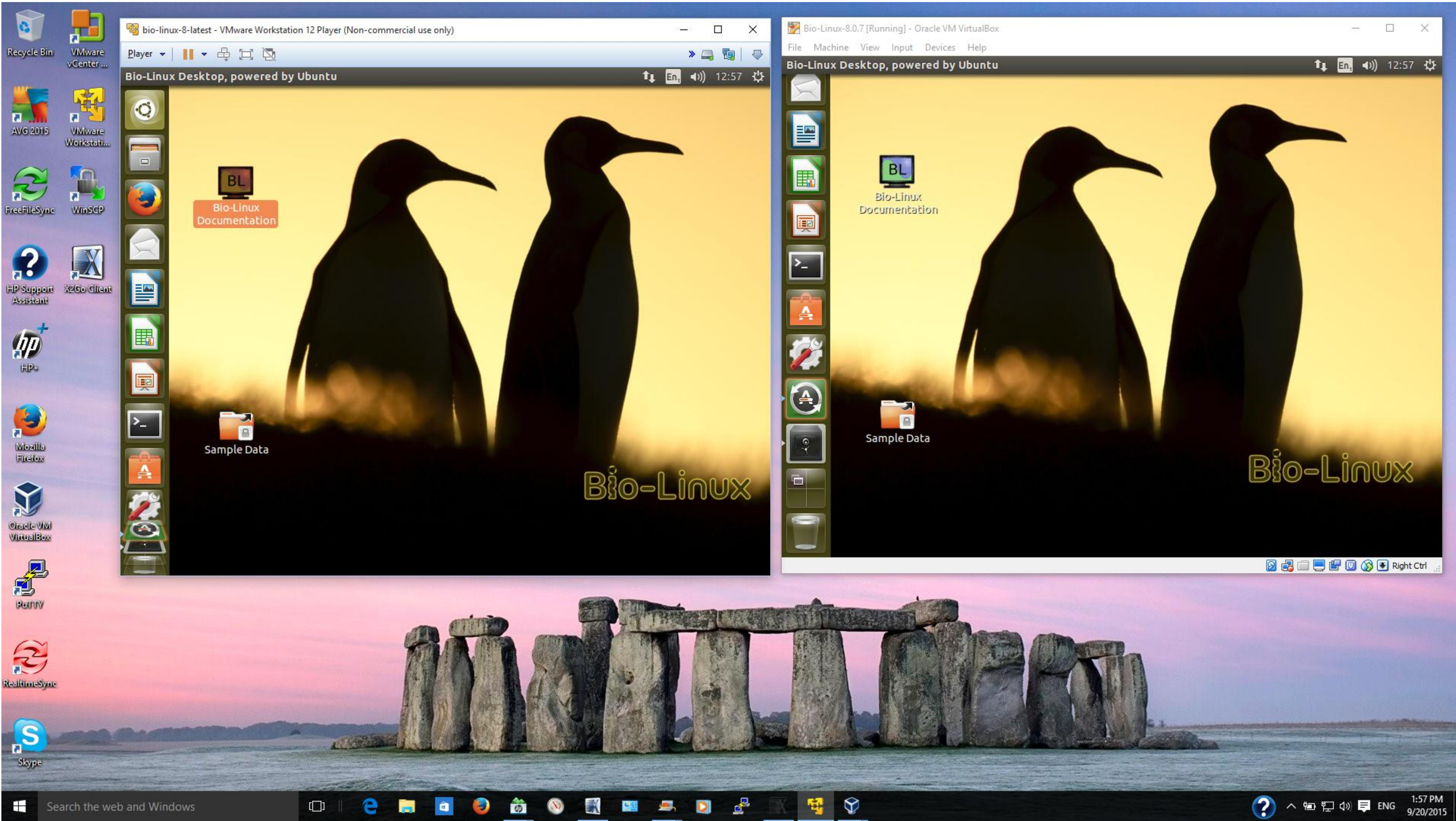
The ibiblio US mirror is here

hosted by ibiblio

nebc.nerc.ac.uk/downloads/bio-linux-8-latest.ova

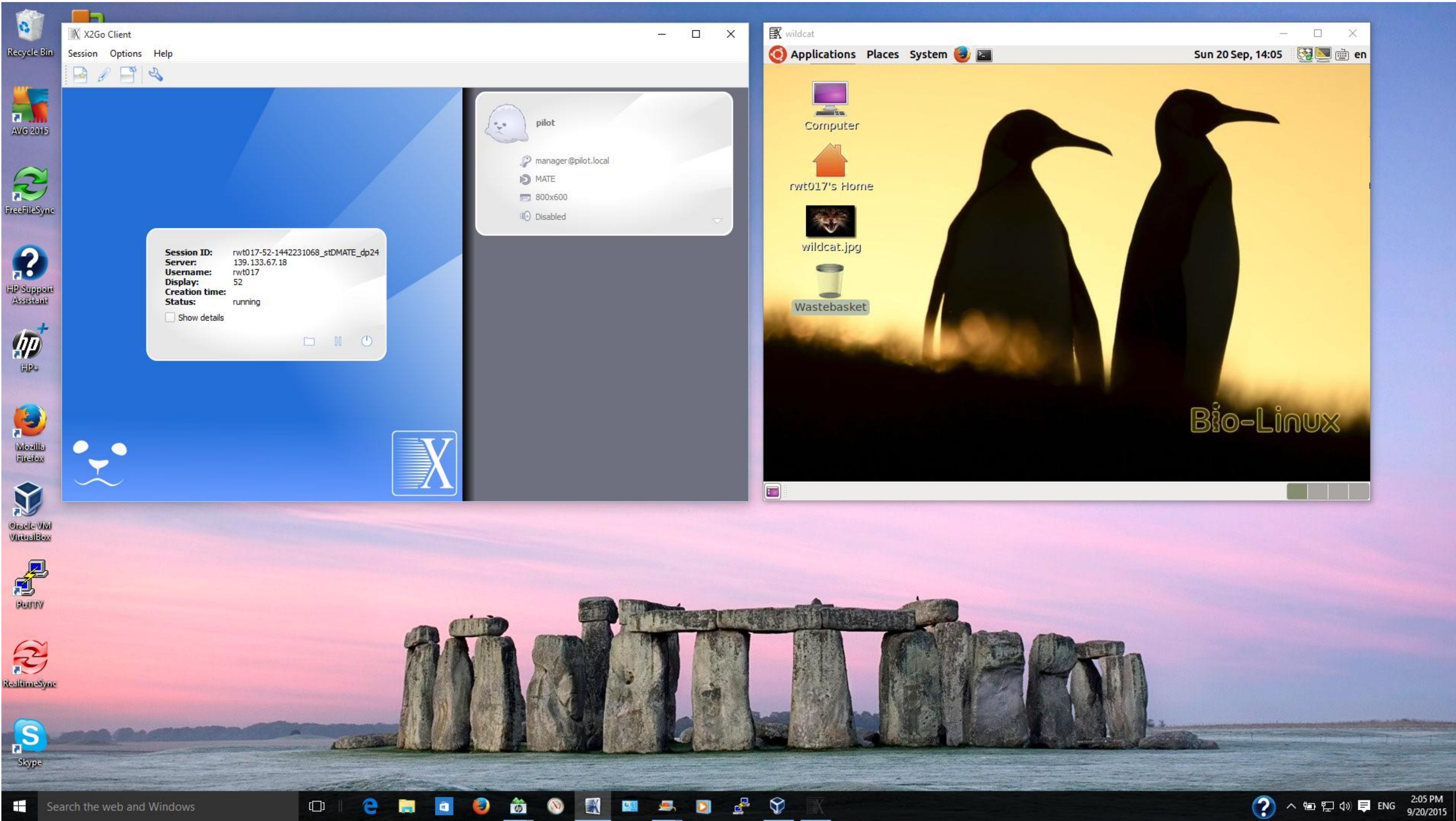
Import into Vmware or VirtualBox

Run Bio-Linux under Microsoft Windows



Use a Bio-Linux terminal server

Connect to a remote MATE desktop using “x2go”



Session ID: rwt017-52-1442231068_stDMATE_dp24
Server: 139.133.67.18
Username: rwt017
Display: 52
Creation time:
Status: running

□ Show details

pilot

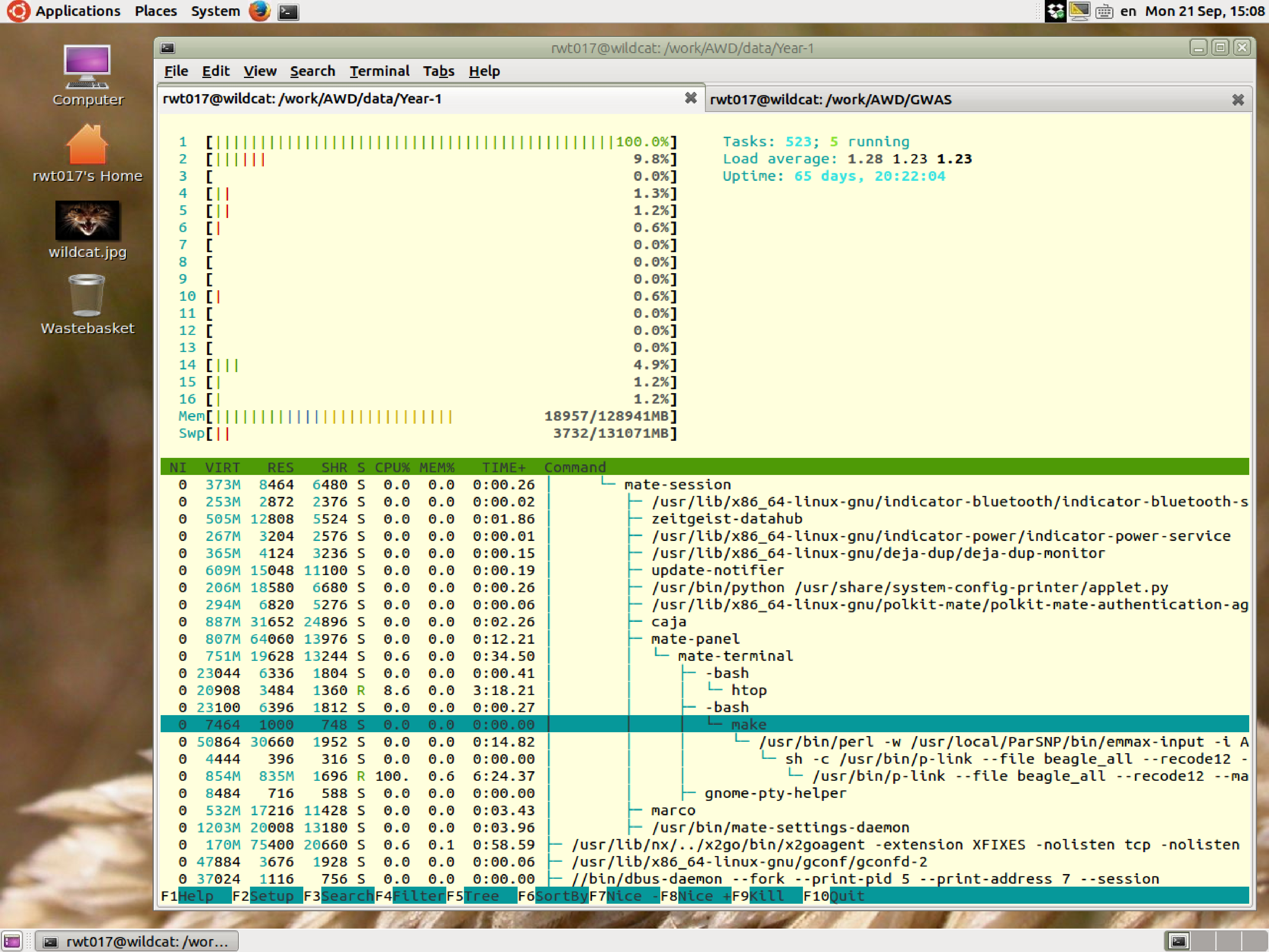
manager@pilot.local
MATE
800x600
Disabled

Bio-Linux terminal server

- Ubuntu 14.04 LTS
  - 2@Opteron 6128
    - 16 cores
  - 128GiB RAM
  - 18TB disk space
    - 2TB system
    - 8TB user
    - 8TB backup
- Bio-Linux 8.0.7

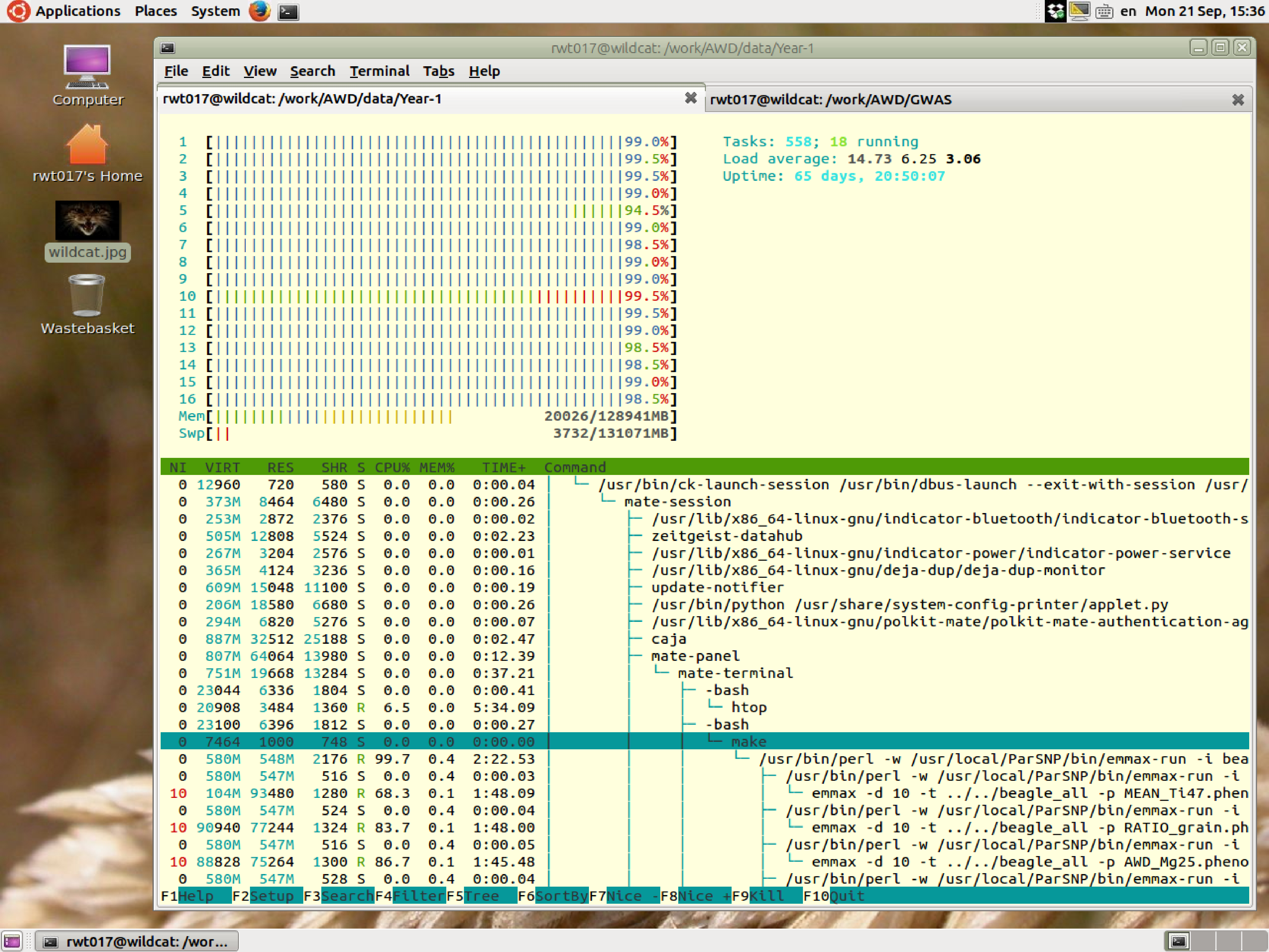
Beware!

wildcat



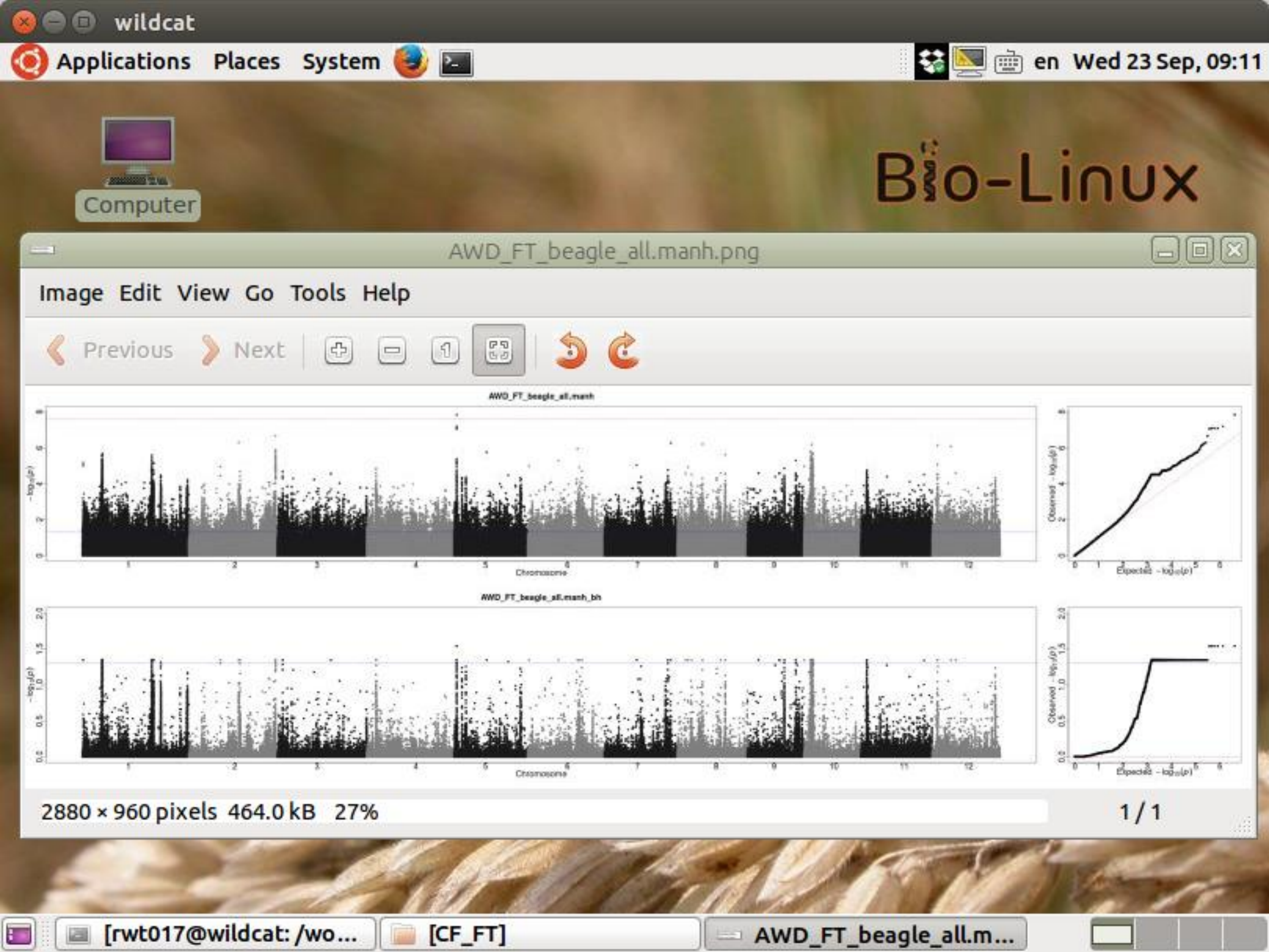
```
  1  [||||||||||||||||||||||||||||||||||||||||||||||100.0%]     Tasks: 523; 5 running
  2  [||||||                                        9.8%]     Load average: 1.28 1.23 1.23
  3  [                                              0.0%]     Uptime: 65 days, 20:22:04
  4  [||                                            1.3%]
  5  [||                                            1.2%]
  6  [|                                             0.6%]
  7  [                                              0.0%]
  8  [                                              0.0%]
  9  [                                              0.0%]
  10 [|                                             0.6%]
  11 [                                              0.0%]
  12 [                                              0.0%]
  13 [                                              0.0%]
  14 [|||                                           4.9%]
  15 [|                                             1.2%]
  16 [|                                             1.2%]
  Mem[||||||||||||||||||||||||||||          18957/128941MB]
  Swp[||                                     3732/131071MB]

 NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
  0  373M  8464  6480 S  0.0  0.0  0:00.26 │      └─ mate-session
  0  253M  2872  2376 S  0.0  0.0  0:00.02 │         ├─ /usr/lib/x86_64-linux-gnu/indicator-bluetooth/indicator-bluetooth-s
  0  505M 12808  5524 S  0.0  0.0  0:01.86 │         ├─ zeitgeist-datahub
  0  267M  3204  2576 S  0.0  0.0  0:00.01 │         ├─ /usr/lib/x86_64-linux-gnu/indicator-power/indicator-power-service
  0  365M  4124  3236 S  0.0  0.0  0:00.15 │         ├─ /usr/lib/x86_64-linux-gnu/deja-dup/deja-dup-monitor
  0  609M 15048 11100 S  0.0  0.0  0:00.19 │         ├─ update-notifier
  0  206M 18580  6680 S  0.0  0.0  0:00.26 │         ├─ /usr/bin/python /usr/share/system-config-printer/applet.py
  0  294M  6820  5276 S  0.0  0.0  0:00.06 │         ├─ /usr/lib/x86_64-linux-gnu/polkit-mate/polkit-mate-authentication-ag
  0  887M 31652 24896 S  0.0  0.0  0:02.26 │         ├─ caja
  0  807M 64060 13976 S  0.0  0.0  0:12.21 │         ├─ mate-panel
  0  751M 19628 13244 S  0.6  0.0  0:34.50 │         │  └─ mate-terminal
  0 23044  6336  1804 S  0.0  0.0  0:00.41 │         │     ├─ -bash
  0 20908  3484  1360 R  8.6  0.0  3:18.21 │         │     │  └─ htop
  0 23100  6396  1812 S  0.0  0.0  0:00.27 │         │     ├─ -bash
  0  7464  1000   748 S  0.0  0.0  0:00.00 │         │     │  └─ make
  0 50864 30660  1952 S  0.0  0.0  0:14.82 │         │     │     └─ /usr/bin/perl -w /usr/local/ParSNP/bin/emmax-input -i A
  0  4444   396   316 S  0.0  0.0  0:00.00 │         │     │        └─ sh -c /usr/bin/p-link --file beagle_all --recode12 -
  0  854M  835M  1696 R 100.  0.6  6:24.37 │         │     │           └─ /usr/bin/p-link --file beagle_all --recode12 --ma
  0  8484   716   588 S  0.0  0.0  0:00.00 │         │     └─ gnome-pty-helper
  0  532M 17216 11428 S  0.0  0.0  0:03.43 │         ├─ marco
  0 1203M 20008 13180 S  0.0  0.0  0:03.96 │         ├─ /usr/bin/mate-settings-daemon
  0  170M 75400 20660 S  0.6  0.1  0:58.59 ├─ /usr/lib/nx/../x2go/bin/x2goagent -extension XFIXES -nolisten tcp -nolisten
  0 47884  3676  1928 S  0.0  0.0  0:00.06 ├─ /usr/lib/x86_64-linux-gnu/gconf/gconfd-2
  0 37024  1116   756 S  0.0  0.0  0:00.00 ├─ //bin/dbus-daemon --fork --print-pid 5 --print-address 7 --session
F1Help  F2Setup  F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```



```
  1  [||||||||||||||||||||||||||||||||||||||||||||||99.0%]     Tasks: 558; 18 running
  2  [||||||||||||||||||||||||||||||||||||||||||||||99.5%]     Load average: 14.73 6.25 3.06
  3  [||||||||||||||||||||||||||||||||||||||||||||||99.5%]     Uptime: 65 days, 20:50:07
  4  [||||||||||||||||||||||||||||||||||||||||||||||99.0%]
  5  [||||||||||||||||||||||||||||||||||||||||||||||94.5%]
  6  [||||||||||||||||||||||||||||||||||||||||||||||99.0%]
  7  [||||||||||||||||||||||||||||||||||||||||||||||98.5%]
  8  [||||||||||||||||||||||||||||||||||||||||||||||99.0%]
  9  [||||||||||||||||||||||||||||||||||||||||||||||99.0%]
  10 [||||||||||||||||||||||||||||||||||||||||||||||99.5%]
  11 [||||||||||||||||||||||||||||||||||||||||||||||99.5%]
  12 [||||||||||||||||||||||||||||||||||||||||||||||99.0%]
  13 [||||||||||||||||||||||||||||||||||||||||||||||98.5%]
  14 [||||||||||||||||||||||||||||||||||||||||||||||98.5%]
  15 [||||||||||||||||||||||||||||||||||||||||||||||99.0%]
  16 [||||||||||||||||||||||||||||||||||||||||||||||98.5%]
  Mem[|||||||||||||||||||||||||||         20026/128941MB]
  Swp[||                                   3732/131071MB]

 NI  VIRT   RES   SHR S  CPU% MEM%   TIME+  Command
  0 12960   720   580 S  0.0  0.0  0:00.04  └─ /usr/bin/ck-launch-session /usr/bin/dbus-launch --exit-with-session /usr/
  0  373M  8464  6480 S  0.0  0.0  0:00.26     └─ mate-session
  0  253M  2872  2376 S  0.0  0.0  0:00.02        ├─ /usr/lib/x86_64-linux-gnu/indicator-bluetooth/indicator-bluetooth-s
  0  505M 12808  5524 S  0.0  0.0  0:02.23        ├─ zeitgeist-datahub
  0  267M  3204  2576 S  0.0  0.0  0:00.01        ├─ /usr/lib/x86_64-linux-gnu/indicator-power/indicator-power-service
  0  365M  4124  3236 S  0.0  0.0  0:00.16        ├─ /usr/lib/x86_64-linux-gnu/deja-dup/deja-dup-monitor
  0  609M 15048 11100 S  0.0  0.0  0:00.19        ├─ update-notifier
  0  206M 18580  6680 S  0.0  0.0  0:00.26        ├─ /usr/bin/python /usr/share/system-config-printer/applet.py
  0  294M  6820  5276 S  0.0  0.0  0:00.07        ├─ /usr/lib/x86_64-linux-gnu/polkit-mate/polkit-mate-authentication-ag
  0  887M 32512 25188 S  0.0  0.0  0:02.47        ├─ caja
  0  807M 64064 13980 S  0.0  0.0  0:12.39        ├─ mate-panel
  0  751M 19668 13284 S  0.0  0.0  0:37.21        │  └─ mate-terminal
  0 23044  6336  1804 S  0.0  0.0  0:00.41        │     ├─ -bash
  0 20908  3484  1360 R  6.5  0.0  5:34.09        │     │  └─ htop
  0 23100  6396  1812 S  0.0  0.0  0:00.27        │     ├─ -bash
  0  7464  1000   748 S  0.0  0.0  0:00.00        │     │  └─ make
  0  580M  548M  2176 R 99.7  0.4  2:22.53        │     │     └─ /usr/bin/perl -w /usr/local/ParSNP/bin/emmax-run -i bea
  0  580M  547M   516 S  0.0  0.4  0:00.03        │     │        ├─ /usr/bin/perl -w /usr/local/ParSNP/bin/emmax-run -i
 10  104M 93480  1280 R 68.3  0.1  1:48.09        │     │        │  └─ emmax -d 10 -t ../../beagle_all -p MEAN_Ti47.phen
  0  580M  547M   524 S  0.0  0.4  0:00.04        │     │        ├─ /usr/bin/perl -w /usr/local/ParSNP/bin/emmax-run -i
 10 90940 77244  1324 R 83.7  0.1  1:48.00        │     │        │  └─ emmax -d 10 -t ../../beagle_all -p RATIO_grain.ph
  0  580M  547M   516 S  0.0  0.4  0:00.05        │     │        ├─ /usr/bin/perl -w /usr/local/ParSNP/bin/emmax-run -i
 10 88828 75264  1300 R 86.7  0.1  1:45.48        │     │        │  └─ emmax -d 10 -t ../../beagle_all -p AWD_Mg25.pheno
  0  580M  547M   528 S  0.0  0.4  0:00.04        │     │        ├─ /usr/bin/perl -w /usr/local/ParSNP/bin/emmax-run -i
F1Help  F2Setup  F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```



2880 × 960 pixels 464.0 kB 27%

1 / 1

Conclusions

- Bio-Linux is useful for reproducible research
  - Common platform with well-defined environment
- Biologists get better insight into their data by doing their *own* bioinformatics
- Bioinformaticians can be more effective by training and supporting biologists
  - Peer role for more advanced research projects
- Intellectual freedom really does matter
  - How you use your computer is part of that

Acknowledgements

- Bio-Linux
  - Tim Booth (NEBC)
  - Bela Tiwari (NEBC/CLCbio)
- NuGO
  - Ben van Ommen (TNO)
  - Chris Evelo (BigCat, Maastricht University, NL)
  - Philip de Groot (Wageningen University, NL)
- Molecular genetics of drought tolerance in rice
  - Adam Price (University of Aberdeen, UK)