# Bio-Linux : Past and Future

# Packaging the Galaxy Server

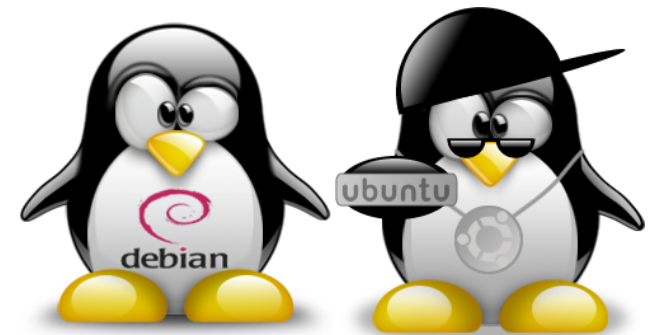Tim Booth *<tbooth@ceh.ac.uk>*
H. Soon Gweon *<soonio@gmail.com>*

**Centre for Ecology & Hydrology**
NATURAL ENVIRONMENT RESEARCH COUNCIL

http://nebc.nerc.ac.uk/tools/bio-linux

# A Bioinformatics Workstation

"one-stop-shop computational solution
-- from desktop to the cloud --
for biologists and bioinformaticians alike"

- Free, open source operating system

- Based on Linux (Ubuntu LTS) (64-bit)

- Free and Open Source

- 8 years + 6 major releases

- 143 packages, >500 bioinformatics tools

Centre for
Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

http://nebc.nerc.ac.uk/tools/bio-linux

Bio-Linux

# Community Outreach

- Free Download, USB Sticks

- Mailing lists (Announce, Developers, Users)

- Helpdesk for User Support: helpdesk@nebc.ac.uk

- Teaching: "Introduction to Bio-Linux"

- Website – e.g. Code Catalogue
  http://nebc.nerc.ac.uk/tools/code-corner

Centre for
Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

http://nebc.nerc.ac.uk/tools/bio-linux

Bio-Linux

# Usage Figures

- >2500 <u>registered</u> ISO downloads (so far)

- 150 active people on discussion list (so far)

- **>5000** distinct IP hits on package repository (last month)
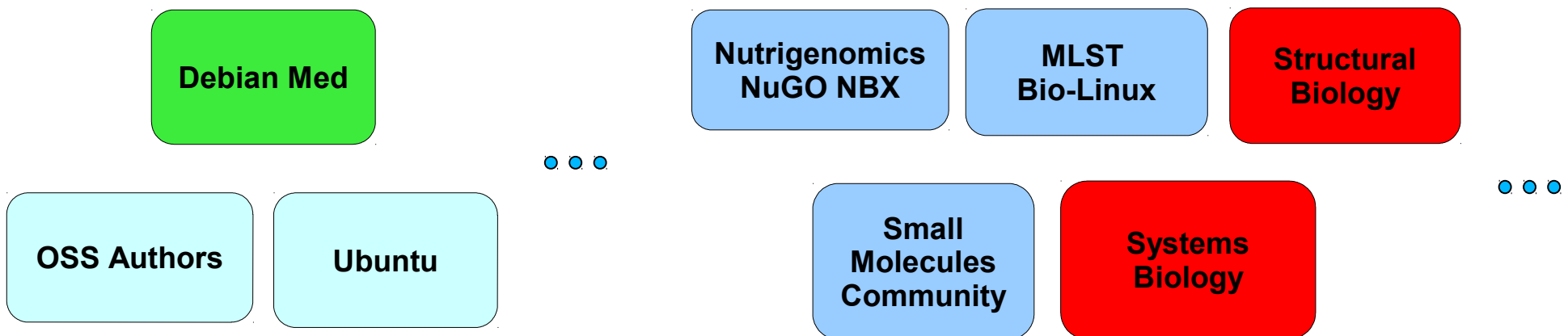
- 1200 page views (last month)

# FUTURE of Bio-Linux

Centre for
Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

http://nebc.nerc.ac.uk/tools/bio-linux

# From "a Bioinformatics Workstation" to "a Computational Biology Workstation"

## Compile software bundles from **Expert Groups**

Debian Med

OSS Authors

Ubuntu

Nutrigenomics NuGO NBX

MLST Bio-Linux

Structural Biology

Small Molecules Community

Systems Biology

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NEBC

http://nebc.nerc.ac.uk/tools/bio-linux

Bio-Linux

# Community Project: CloudBioLinux

Harvard University, School of Public Health

J. Craig Venter Institute, Washington

Galaxy CloudMan, Emory University & Croatia

UK National Grid Service

- Grid computing support ✔
- Amazon EC2 ✔
- Academic clouds ⬅

**http://cloudbiolinux.org**

CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NEBC

Bio-Linux

# New methods for coping with "Big Data"

- **Use Case: InterPro** (a tool set as well as a major reference database)
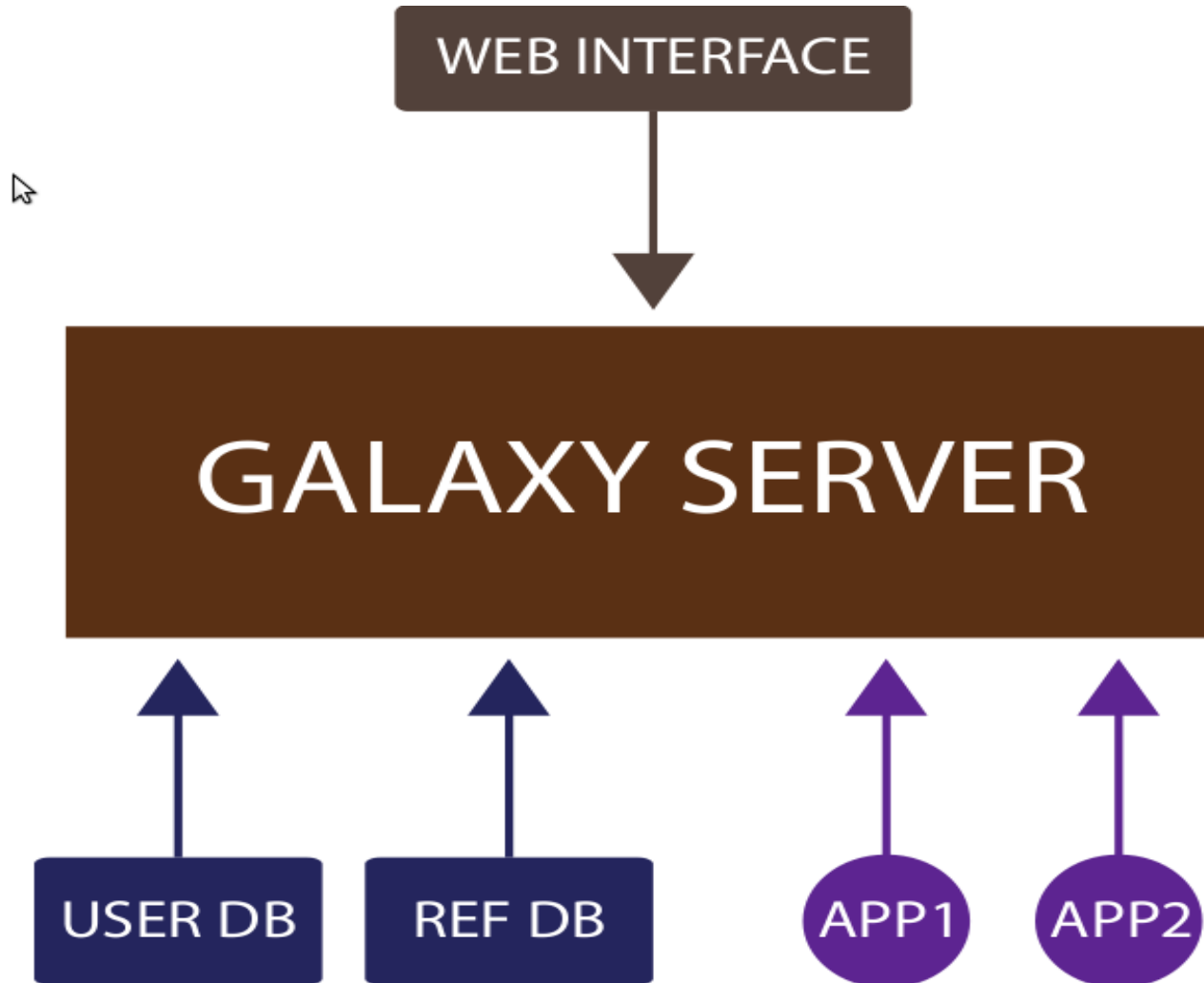
  Project Partner, EBI

- '**getData' system for packaging reference with associated tools**

  Project Partner, Steffen Moeller (Lubeck Uni) & Dr. Charles Plessy (RIKEN, Japan)

- **How to secure data** in a local cloud context

  Industrial Collaborator, Recombinant Data

# Basic Galaxy Architecture

| | Dependency (Version) | Section | Tools | BL Status | Comments | Essential | Esser |
|---|---|---|---|---|---|---|---|
| 2 | add_scores () | Evolution | Add scores | New package | Tiny C app - https://bitbucket.org/natefoo/add_scores | | |
| 3 | beam2 | Human Genome Variation | ? | New package | | | |
| 4 | Blast+ (NCBI BLAST+) | NGS: Mapping | Blast+ | Done | Packaged recently to Deb standard, but needs more testing. | 1 | |
| 5 | bowtie (0.12.7) | NGS: Mapping | Map with Bowtie for Illumina, SOLiD | Needs work | Easy backport?? | 1 | |
| 6 | BWA (0.5.6) | NGS: Mapping | Map with BWA for Illumina, SOLiD | New package | | 1 | |
| 7 | ClustalW (2.1) | Multiple Alignments | Align with Clustal | Done | | 1 | |
| 8 | cufflinks (1.0.1) | NGS: RNA Analysis | Cufflinks, Cuffcompare, Cuffdiff | Needs work | Easy backport | 1 | |
| 9 | cElementTree (Python) () | NGS: Mapping | Parse megablast xml output | New package | | 1 | |
| 10 | EMBOSS () | EMBOSS | ALL | Needs work | At least needs testing, maybe reconfiguratioin | 1 | |
| 11 | Gnuplot (Python) () | Graph/Display Data | Bar chart | Done | | | |
| 12 | HYPHY () | Evolution | ALL | New package | | 1 | |
| 13 | LAJ | ? | ? | New package | | 1 | |
| 14 | R package 'leaps' () | Multiple Regression | Perform Best-subsets Regression | New package | | | |
| 15 | Lastz (1.01.88) | NGS: Mapping | Lastz, Lastz paired reads | New package | | 1 | |
| 16 | liftOver (UCSC) () | Lift-Over | Convert genome coordinates | New package | | 1 | |
| 17 | lps_tool | ? | ? | New package | | | |
| 18 | macs () | MACS | NGS: Peak Calling | New package | | 1 | |
| 19 | megablast (2.2.22) () | NGS: Mapping | Megablast | Needs work | | 1 | |
| 20 | Memesuite | Motif tools | ? | New package | | | |
| 21 | NumPy () | Various | Various | Done | | | |
| 22 | Pass2 | Human Genome Variation | ? | New package | | | |
| 23 | Perl () | Regional Variation, Text Manuipulation | Various | Done | | | |
| 24 | PerM (0.2.6) | NGS: Mapping | Map with PerM | New package | | 1 | |
| 25 | ps2pdf (Ghostscript) () | Evolution: HyPhy, Metagenomic analyses | Neighbor Joining Tree, Draw Phylogeny | Done | | | |
| 26 | PyTables () | Get Genomic Scores | Compute phastOdds score | Done | | | |
| 27 | R package 'prettyR' () | Join, Subtract and Group | Group | New package | | | |
| 28 | R package 'leaps' () | Multiple Regression | Perform Best-subsets Regression | Needs work | | | |
| 29 | R package 'tcltk' () | Statistics | Compute q-values | New package | | | |
| 30 | R package 'qvalue' () | Statistics | Compute q-values | Done | | | |
| 31 | R package 'Rwave' () | Discrete Wavelet Analysis | all | New package | | | |
| 32 | R package 'wavethresh' () | Discrete Wavelet Analysis | all | New package | | | |
| 33 | R package 'waveslim' () | Discrete Wavelet Analysis | all | New package | | | |
| 34 | RPy () | Graph/Display Data | Histogram | Done | | | |
| 35 | samtools (0.1.12a) | NGS: SAM Tools | SAM-to-BAM, BAM-to-SAM, merge BAM, generate pileup | Needs work | Needs backport | 1 | |
| 36 | sift | Human Genome Variation | ? | New package | | | |
| 37 | sputnik () | Regional Variation | Extract Orthologous Microsatellites | New package | | 1 | |
| 38 | taxBuilder () | Metagenomic analyses | Fetch taxonomic representation | New package | | 1 | |
| 39 | taxonomy2tree () | Metagenomic analyses | Summarize taxonomy, draw phylogent | New package | | | |
| 40 | tophat (1.2.0) | NGS: RNA Analysis | Tophat | Needs work | Needs backport | 1 | |
| 41 | tree2PS-fast () | Metagenomic analyses | Draw phylogeny | New package | | 1 | |
| 42 | twoBitToFa (UCSC) () | Evolution | Add scores | New package | Casey advises packaging the whole tree for UCSC | 1 | |
| 43 | velvet () | NGS: Assembly | velvetg, veleveth | Needs work | Easy backport | 1 | |
| 44 | weblogo3 | Sequence Logo generator | Weblogo3 | New package | | | |
| 45 | mothur | Metagenomic analyses | | | | | |
| 46 | | | | | NGS and evolution packages | 22 | |
| 47 | | | | | # hours at 2 hrs/package | 44 | |

Tabs: Bio-Linux packages | Milestone | **Galaxy deps orig** ▾ | Galaxy deps progress | Casey's prioritize list/notes | Related Links    Dependency (Version)

http://nebc.nerc.ac.uk/tools/bio-linux

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Dependency (Version)** | **Section** | **Tools** | **BL Status** | **Comments** | **Essential** Es |
| 2 | add_scores () | Evolution | Add scores | Done 23/12 | Tiny C app - https://bitbucket.org/natefoo/add_scores | |
| 3 | beam2 | Human Genome Variation | ? | Done 6/1/2012 | | |
| 4 | Blast+ (NCBI BLAST+) | NGS: Mapping | Blast+ | Done | Packaged recently to Deb standard, but needs more testing. | 1 |
| 5 | bowtie (0.12.7) | NGS: Mapping | Map with Bowtie for Illumina, SOLiD | Done 21/12 | In PPA - needs test | 1 |
| 6 | BWA (0.5.6) | NGS: Mapping | Map with BWA for Illumina, SOLiD | Done 5/1/2012 | | 1 |
| 7 | ClustalW (2.1) | Multiple Alignments | Align with Clustal | Done | native | 1 |
| 8 | cufflinks (1.0.1) | NGS: RNA Analysis | Cufflinks, Cuffcompare, Cuffdiff | Done 22/12 | Version 1.2.1 packaged - is this OK? Problem found in testing - needs rebuild. | 1 |
| 9 | cElementTree (Python) () | NGS: Mapping | Parse megablast xml output | Done | Standard component of python2.6 | 1 |
| 10 | EMBOSS () | EMBOSS | ALL | Needs work | At least needs testing, maybe reconfiguratioin | |
| 11 | Gnuplot (Python) () | Graph/Display Data | Bar chart | Done | | |
| 12 | HYPHY () | Evolution | ALL | Done 21/12 | In PPA - needs test | 1 |
| 13 | LAJ | ? | ? | New package | Source not available, requested by email | 1 |
| 14 | R package 'leaps' () | Multiple Regression | Perform Best-subsets Regression | New package | | |
| 15 | Lastz (1.01.88) | NGS: Mapping | Lastz, Lastz paired reads | Done 20/12 | In PPA - needs test | 1 |
| 16 | liftOver (UCSC) () | Lift-Over | Convert genome coordinates | New package | Not free software, cannot upload to PPA | 1 |
| 17 | lps_tool | ? | ? | New package | | |
| 18 | macs () | MACS | NGS: Peak Calling | Done 5/1/2012 | | 1 |
| 19 | megablast (2.2.22) () | NGS: Mapping | Megablast | Needs work | | 1 |
| 20 | Memesuite | Motif tools | ? | New package | | |
| 21 | NumPy () | Various | Various | Done | | |
| 22 | Pass2 | Human Genome Variation | ? | Done 6/1/2012 | | |
| 23 | Perl () | Regional Variation, Text Manuipulation | Various | Done | | |
| 24 | PerM (0.2.6) | NGS: Mapping | Map with PerM | Done 5/1/2012 | | 1 |
| 25 | ps2pdf (Ghostscript) () | Evolution: HyPhy, Metagenomic analyses | Neighbor Joining Tree, Draw Phylogeny | Done | | 1 |
| 26 | PyTables () | Get Genomic Scores | Compute phastOdds score | Done | | |
| 27 | R package 'prettyR' () | Join, Subtract and Group | Group | New package | | |
| 28 | R package 'leaps' () | Multiple Regression | Perform Best-subsets Regression | Needs work | | |
| 29 | R package 'tcltk' () | Statistics | Compute q-values | New package | | |
| 30 | R package 'qvalue' () | Statistics | Compute q-values | Done | | |
| 31 | R package 'Rwave' () | Discrete Wavelet Analysis | all | New package | | |
| 32 | R package 'wavethresh' () | Discrete Wavelet Analysis | all | New package | | |
| 33 | R package 'waveslim' () | Discrete Wavelet Analysis | all | New package | | |
| 34 | RPy () | Graph/Display Data | Histogram | Done | Sensitive to both R and Python version - fixed in PPA | |
| 35 | samtools (0.1.12a) | NGS: SAM Tools | SAM-to-BAM, BAM-to-SAM, merge BAM, generate pileup | Done 20/12 | Packaged 0.1.18 - needs test | 1 |
| 36 | sift | Human Genome Variation | ? | New package | | |
| 37 | sputnik () | Regional Variation | Extract Orthologous Microsatellites | Done 5/1/2012 | | |
| 38 | taxBuilder () | Metagenomic analyses | Fetch taxonomic representation | Done on 20/12 | See tree2PS-fast | 1 |
| 39 | taxonomy2tree () | Metagenomic analyses | Summarize taxonomy, draw phylogent | Done on 20/12 | See tree2PS-fast | 1 |
| 40 | tophat (1.2.0) | NGS: RNA Analysis | Tophat | Needs work | Need to align Tim's build with DM Git version | 1 |
| 41 | tree2PS-fast () | Metagenomic analyses | Draw phylogeny | Done 23/12 | Tim working on this - https://bitbucket.org/natefoo/taxonomy/overview | 1 |
| 42 | twoBitToFa (UCSC) () | Evolution | Add scores | New package | Casey advises packaging the whole tree for UCSC - but Not free software, cannot upload to PPA. Will have to bundle non-free binaries. | |
| 43 | velvet () | NGS: Assembly | velvetg, veleveth | Done on 20/12 | 1.1.05 in PPA - needs test | |
| 44 | weblogo3 | Sequence Logo generator | Weblogo3 | Done 6/1/2012 | | |
| 45 | mothur | Metagenomic analyses | | Done 7/12 | | |
| 46 | segan | | library needed by tophat | Needs work | Current version insufficient to build tophat - need 1.3 - Tim working on it - see timsppa | |

Bio-Linux packages | Milestone | Galaxy deps orig | **Galaxy deps progress** ▼ | Casey's prioritize list/notes | Related Links

Dependency (Version)

http://nebc.nerc.ac.uk/tools/bio-linux