

Deploying bioinformatics tools with CloudBioLinux

Brad Chapman
Bioinformatics Core,
Harvard School of Public Health
<https://github.com/chapmanb>

30 June 2014

Outline

- Overview of CloudBioLinux
- Create custom set of tools
- Install from custom flavor
- Add a new tool

What is CloudBioLinux?

Infrastructure for installing biological software

- deb/rpm packages
- Bio-Linux
- Linuxbrew with homebrew-science
- Python, Ruby, R package management
- Conda + Binstar <https://conda.binstar.org/>
- Custom installation scripts

Key Features

- Community
- Curation


Community

PUBLIC  chapmanb / cloudbiolinux


 Unwatch 22  Unstar 123  Fork 79

CloudBioLinux: configure virtual (or real) machines with tools for biological analyses <http://cloudbiolinux.org> —
Edit

1,987 commits 2 branches 0 releases 26 contributors

 branch: master cloudbiolinux / +

Use GitHub mirror of RNA-SeQC jar to avoid Broad download issues with...

 chapmanb authored 6 days ago latest commit a4bb6f3390

cloudbio	Use GitHub mirror of RNA-SeQC jar to avoid Broad download issues with...	6 days ago
config	updated yamli; sacCer2 => sacCer3	15 days ago
contrib	Update to latest version of snpEff, GEMINI and VEP. Prefer brew insta...	23 days ago
deploy	Deployer: Bug fixes and clean up of deploy_bourne.sh.	10 months ago
doc	Initial work on new CloudBioLinux as a framework documentation write...	11 months ago
installed_files	For the remote desktop, switch to xfce4 (vs. jwm) window manager beca...	2 months ago
manifest	updated yamli; sacCer2 => sacCer3	15 days ago
test	Simplify customization: allow flavors to specify a directory containi...	2 years ago

<> Code

Issues 1

Pull Requests 1


Pulse

Graphs

Network

Settings

HTTPS clone URL



You can clone with HTTPS, SSH,
or Subversion.

 Download ZIP

<https://github.com/chapmanb/cloudbiolinux>

History

Integration of multiple efforts

- JCVI Cloud Bio-Linux
- Bioperl Max
- Infochimps machetEC2
- Bio-Linux
- DebianMed

Original goal

Overcome bare-metal problem with AWS images

- Ubuntu
- Single AMI with biological tools
- Automated build infrastructure
- Bring in developer community
- Ready to use for researchers

Biological data

- Genomes, organized and indexed
- Associated data files: dbSNP, reference transcripts
- S3 bucket
- Tools with organized data
- GEMINI: <https://github.com/arq5x/gemini>

Local installation

- Multiple platforms: Ubuntu, RedHat/CentOS, Debian, ScientificLinux
- Isolated installations: no sudo, non-VM environments
- Rapid turnaround for fixes

Flavors: customized installations

- Target specific use case
- Sub-collection of packages from full distribution
- Example:
`cloudbiolinux/contrib/flavor/biopython`

Pjotr Prins

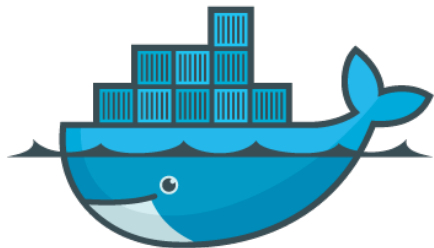
Hidden infrastructure

bcbio-nextgen

- CloudBioLinux drives fully automated installation
- Reproducible build scripts for docker migration

<https://github.com/chapmanb/bcbio-nextgen>

Containers



docker

<http://docker.io/>

<https://github.com/chapmanb/bcbio-nextgen-vm>

Galaxy toolshed integration vision

- CloudBioLinux flavor to install tools
- Install in isolated Docker container
- Galaxy support for Docker

[https://bitbucket.org/galaxy/galaxy-central/
pull-request/401/
allow-tools-and-deployers-to-specify](https://bitbucket.org/galaxy/galaxy-central/pull-request/401/allow-tools-and-deployers-to-specify)

Manifest

- Full manifest of installed software
- Prioritize biological software
- YAML format for parsing and downstream queries

Outline

- Overview of CloudBioLinux
- Create custom set of tools
- Install from custom flavor
- Add a new tool












CloudBioLinux architecture

- YAML configuration
- Flavors
- Fabric scripts
- Documentation

What is a flavor?

- Subset of full CloudBioLinux packages
- Defined set of packages for a task

YAML configuration: directory


 packages-debian.yaml	Remap debian packages not available from latest ubuntu list. Thanks t...
 packages-homebrew.yaml	Update to latest version of snpEff, GEMINI and VEP. Prefer brew insta...
 packages-nix.yaml	Disable default use of Nix Packages
 packages-scientificlinux.yaml	Fill in some missing gaps between Ubuntu and CentOS/SL (base R). Add ...
 packages-yum.yaml	Ensure ruby installed for bootstrapping homebrew on bare machines dur...
 packages.yaml	Avoid estscan install which is problematic on Ubuntu 14.04. Fixes #163
 perl-libraries.yaml	Mega patch of fixes to get installation working cleanly on localhost ...
 puppet_classes.yaml	Initial work LWR integration.
 python-libraries.yaml	fixed issues with libraries missing from pypi by adding --allow-unver...
 r-libraries.yaml	updated yaml; sacCer2 => sacCer3
 ruby-libraries.yaml	Mega patch of fixes to get installation working cleanly on localhost ...

YAML configuration: example

```
1  # Packages available in the Homebrew and Linuxbrew package manager
2  ---
3  bio_nextgen:
4    alignment:
5      - bwa
6      - bowtie2
7      - novoalign
8      - rna-star
9    utilities:
10     - bamtools
11     - bedtools
12     - cramtools
13     - libmaus
14     - biobambam
15     - fastqc
16     - fastx_toolkit
17     - qualimap
18     - sambamba
19     - staden_io_lib
20   analysis:
21     #- cufflinks
22     - samtools
23     - htlib
24     - bcftools
25     #- tophat
26   variant:
```

Example flavor





 branch: **master** ▾ [cloudbiolinux](#) / [contrib](#) / [flavor](#) / [ngs_pipeline_minimal](#) / +

Update to latest version of snpEff, GEMINI and VEP. Prefer brew insta... 



chapmanb authored 22 days ago

..

 custom.yaml	Remove pbzip which fails to compile on Mac and is not currently used...
 main.yaml	Add bioconductor libraries to bcbio-nextgen installation flavor to st...
 packages-homebrew.yaml	Update to latest version of snpEff, GEMINI and VEP. Prefer brew insta...
 r-libs.yaml	Add additional R libraries for cn.mops: snow and rtracklayer. Update ...

Edit main.yaml

```
1  ---
2  # Flavor containing with minimal instructions to install tools for
3  # running next-generation sequencing pipelines.
4  packages:
5    - minimal
6    - libraries
7    - python
8    - java
9    - r
10   - ruby
11   - bio_nextgen
12  libraries:
13    - r-libs
```

Edit set of brew installed packages

```
1  # Packages available in the Homebrew and Linuxbrew package mar
2  ---
3  bio_nextgen:
4      alignment:
5          - bwa
6      #      - bowtie2          (2.2.0 doesn't work with Tophat, so use 1
7          - novoalign
8          - rna-star
9      utilities:
10         - bamtools
11         - bedtools
12         - cramtools
13         - libmaus
14         - biobambam
15         - fastqc
16         - qualimap
17         - sambamba
18         - samblaster
19         - seqtk==HEAD
20         - speedseq
21         - staden_io_lib
```

Edit fabricrc.txt

```
35 # Global installation directory for packages and standard programs
36 system_install = /usr/local
37
38 # Local install directory for versioned software that will not
39 # be included in the path by default
40 local_install = /usr/local/share
41
42 # Shell to be used by CBL scripts during runtime
43 shell_config = ~/.bashrc
44 shell = /bin/bash -i -c
45
46 # Global setting for using sudo; allows installation of custom packages
47 # by non-privileged users.
48 # *Note*: ``system_install`` needs to point to a user-writeable directory if
49 # ``use_sudo`` is set to ``False``
50 use_sudo = True
51
52 # -- Details about reference data installation
53
54 # Path where biological reference data files should be retrieved to
55 data_files = /mnt/biodata
56
```

Outline

- Overview of CloudBioLinux
- Create custom set of tools
- Install from custom flavor
- Add a new tool

Setup: get CloudBioLinux and Fabric

Retrieve source and fabric for execution


```
$ git clone https://github.com/chapmanb/cloudbiolinux.git  
$ pip install fabric
```

<https://github.com/fabric/fabric>






Short demonstration flavor

branch: master **cloudbiolinux** / contrib / flavor / **demo** / +

Add demo flavor for showing ability to install tools locally. Support... ...

 **chapmanb** authored 8 minutes ago

..

 README.md	Add demo flavor for showing ability to install tools locally. Support...
 custom.yaml	Add demo flavor for showing ability to install tools locally. Support...
 fabricrc.txt	Add demo flavor for showing ability to install tools locally. Support...
 main.yaml	Add demo flavor for showing ability to install tools locally. Support...
 packages-homebrew.yaml	Add demo flavor for showing ability to install tools locally. Support...

contrib/flavor/demo

Install

Single command

```
$ cd cloudbiolinux
```

```
$ fab -H localhost install_biolinux:flavor=demo
```

Isolated install directory

```
$ tree -d -L 2 ~/tmp/cbl_demo/
/home/chapmanb/tmp/cbl_demo
|-- bin
|-- Cellar
|   |-- bedtools
|   |-- bwa
|   |-- gatk-framework
|   '-- samtools
|-- include
|   '-- bam -> ../Cellar/samtools/0.1.19/include/bam
|-- lib
|   '-- pkgconfig
|-- Library
|   |-- Aliases
|   |-- Contributions
|   |-- ENV
|   |-- Formula
|   |-- Homebrew
|   |-- LinkedKegs
|   '-- Taps
|-- opt
|   |-- bedtools -> ../Cellar/bedtools/2.19.1
|   |-- bwa -> ../Cellar/bwa/0.7.9a
|   |-- gatk-framework -> ../Cellar/gatk-framework/3.1-1
|   '-- samtools -> ../Cellar/samtools/0.1.19
'-- share
    |-- doc
    |-- java -> ../Cellar/gatk-framework/3.1-1/share/java
    |-- man
    '-- samtools -> ../Cellar/samtools/0.1.19/share/samtools
```

Update paths to include automatically

```
export PATH=~ /tmp/cbl_demo/bin:$PATH
export LD_LIBRARY_PATH=~ /tmp/cbl_demo/lib:$LD_LIBRARY_PATH
export PERL5LIB=~ /tmp/cbl_demo/lib/perl5:
                ~/tmp/cbl_demo/lib/perl5/site_perl:${PERL5LIB}
```

Do some science

```
$ bedtools
```

```
$ samtools
```

```
$ gatk-framework
```

Outline

- Overview of CloudBioLinux
- Create custom set of tools
- Install from custom flavor
- Add a new tool

Tool add options

- Add to any existing packaging community
 - DebianMed
 - Bio-Linux
 - Homebrew
- Custom python code
- This example: Homebrew recipe

Homebrew/Linuxbrew

Linuxbrew

A fork of Homebrew for Linux

Install Linuxbrew (tl;dr)

Paste at a Terminal prompt:

```
ruby -e "$(wget -O- https://raw.githubusercontent.com/Homebrew/linuxbrew/go/install)"
```

See [Dependencies](#) and [Installation](#) below for more details.


Features

- Can install software to a home directory and so does not require sudo
- Install software not packaged by the native distribution
- Install up-to-date versions of software when the native distribution is old
- Use the same package manager to manage both your Mac and Linux machines

<https://github.com/Homebrew/homebrew>

<https://github.com/Homebrew/linuxbrew>

homebrew-science

PUBLIC  Homebrew / **homebrew-science**


Watch 30 Star 324 Fork 336









Scientific formulae for the Homebrew package manager <http://brew.sh>

1,690 commits 17 branches 0 releases 250 contributors

branch: master **homebrew-science** / +

Maxima 5.33.0

 rinx authored 10 hours ago latest commit e7a4573e32
* dpo committed 4 hours ago

 CONTRIBUTING.md	CONTRIBUTING: add Table of Contents, adjust heading style	2 months ago
 README.md	README: update links	23 days ago
 abyss-explorer.rb	ABYSS-Explorer 1.3.4: New formula	8 months ago
 abyss.rb	abyss 1.5.1	2 months ago
 adol-c.rb	Adol-C: upgrade to 2.5.0.	7 days ago
 alembic.rb	alembic: use hg release tag	a month ago
 allpaths-lg.rb	allpaths-lg: fix url	3 months ago
 amos.rb	Batch convert http download urls from SourceForge to https	4 months ago

Code

Issues 30

Pull Requests 13

Wiki

Pulse

Graphs

Network

HTTPS clone URL


<https://github.com>

You can clone with **HTTPS**, **SSH**, or **Subversion**.

Download ZIP

<https://github.com/Homebrew/homebrew-science>

homebrew-cbl

PUBLIC  chapmanb / homebrew-cbl

Unwatch 2 Star 3 Fork 3











Homebrew repository for CloudBioLinux: incubator for formulas to end up in homebrew-science — Edit

67 commits 1 branch 0 releases 3 contributors

branch: master homebrew-cbl / +

Versioned build target for samtools-0.2.0-rc9

chapmanb authored 3 days ago latest commit c3c6399b8b

 README.md	Add builds for freebayes associated tools from GitHub revisions: vcfl...	7 months ago
 biobambam.rb	VEP: ensure plugin directory included in library. biobambam: ensure w...	22 days ago
 cramtools.rb	Update cramtools jar sha1 to match latest upstream update with exclud...	27 days ago
 freebayes.rb	Ensure FreeBayes version set prior to build, for systems with differe...	2 months ago
 gatk-framework.rb	Update gatk-framework to latest version with fix for quoted shell arg...	a month ago
 gla.rb	Revert gla back to pre-gssw version to avoid compile errors. Fixes #4	4 months ago
 hall-lab-sv-tools.rb	Provide recipe for structural variation scripts from the Hall lab	a month ago
 libmaus.rb	Ensure pkg-config installed as part of libmaus install so biobambam c...	22 days ago
 platypus-variant.rb	Avoid naming conflict for platypus with OSX build tool	4 months ago
 ma-star.rb	Update ma-star to fixed z4 version.	2 months ago

Code

Issues 0

Pull Requests 0

Wiki


Pulse

Graphs

Network

Settings

HTTPS clone URL

https://github.com/ 

You can clone with HTTPS, SSH, or Subversion

Download ZIP

<https://github.com/chapmanb/homebrew-cbl>

Simple recipe

```
1  require 'formula'
2
3  class Vt < Formula
4    homepage 'https://github.com/atks/vt'
5    version '2014-04-23'
6    url 'https://github.com/atks/vt.git', :revision => '22894f949a'
7
8    def install
9      system 'make'
10     bin.install 'vt'
11   end
12
13   test do
14     system 'vt'
15   end
16 end
```

<https://github.com/chapmanb/homebrew-cbl/blob/master/vt.rb>

Complex recipe

```
1 require 'formula'
2
3 class Vep < Formula
4   homepage 'http://ensembl.org/info/docs/variation/vep/index.html'
5   version '75_2014-06-12'
6   url 'https://github.com/Ensembl/ensembl-tools/archive/771dfa1016c357145be7016c91e1155ae7c021f2.zip'
7   sha1 '141a8c639c442bf02d15846fab534dbe58dec3e4'
8
9   resource "plugins" do
10     url "https://github.com/ensembl-variation/VEP_plugins/archive/2c123aff2deef07ee094984fc44e19c48975af4.zip"
11     sha1 "0569239ed8255d277db034d838d6ec51b90481a8"
12   end
13
14   resource 'loftee' do
15     url 'https://github.com/konradjk/loftee/archive/545cf9ac5f25b6a6872984dd1a3197a7e7caf000.zip'
16     sha1 'a61c6196964526becfb7afa9838d0495d7996e9'
17   end
18
19   def install
20     # VEP
21     inreplace 'scripts/variant_effect_predictor/variant_effect_predictor.pl' do |s|
22       s.sub! 'use lib $Bin;', "use lib $Bin;\nuse lib '#{prefix}/lib';\nuse lib '#{prefix}/lib/Plugins';\n";
23       s.sub! "my $default_dir = join '/', ($ENV{'HOME'}, '.vep')", "my $default_dir = '#{prefix}/lib';"
24     end
25     inreplace 'scripts/variant_effect_predictor/INSTALL.pl' do |s|
26       s.sub! "$DEST_DIR |]=" '.', " $DEST_DIR |]=" '#{prefix}/lib';"
27       s.sub! "$CACHE_DIR |]=" ENV{'HOME'} ? ENV{'HOME'}. '/.vep' : 'cache';", "$CACHE_DIR |]=" '#{share}/data';"
28     end
29     inreplace 'scripts/variant_effect_predictor/convert_cache.pl' do |s|
30       s.sub! 'use strict;', "use strict;\nuse lib '#{prefix}/lib';"
31     end
32     prefix.install Dir['scripts/variant_effect_predictor/*.pl']
33     bin.install_symlink prefix / 'variant_effect_predictor.pl'
34     bin.install_symlink prefix / 'filter_vep.pl'
35     bin.install_symlink prefix / 'INSTALL.pl' => 'vep_install.pl'
```

<https://github.com/chapmanb/homebrew-cbl/blob/master/vep.rb>

Recap

- Overview of CloudBioLinux
- Create custom set of tools
- Install from custom flavor
- Add a new tool