Introduction to Linux for bioinformatics

Getting software

Joachim Jacob
5 and 12 May 2014

Figure: http://worldofdtcmarketing.com/website-content-now-critical-for-search-engine-performance-pharma/health-information-online/

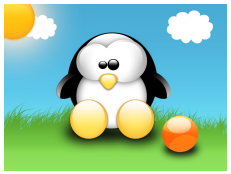
Software for Linux

- Just like Linux, most software written for Linux is **open source and free** (e.g. GNU software).
- Depending on your chosen **distribution**, it is easier/harder to install packages (=another name for software).

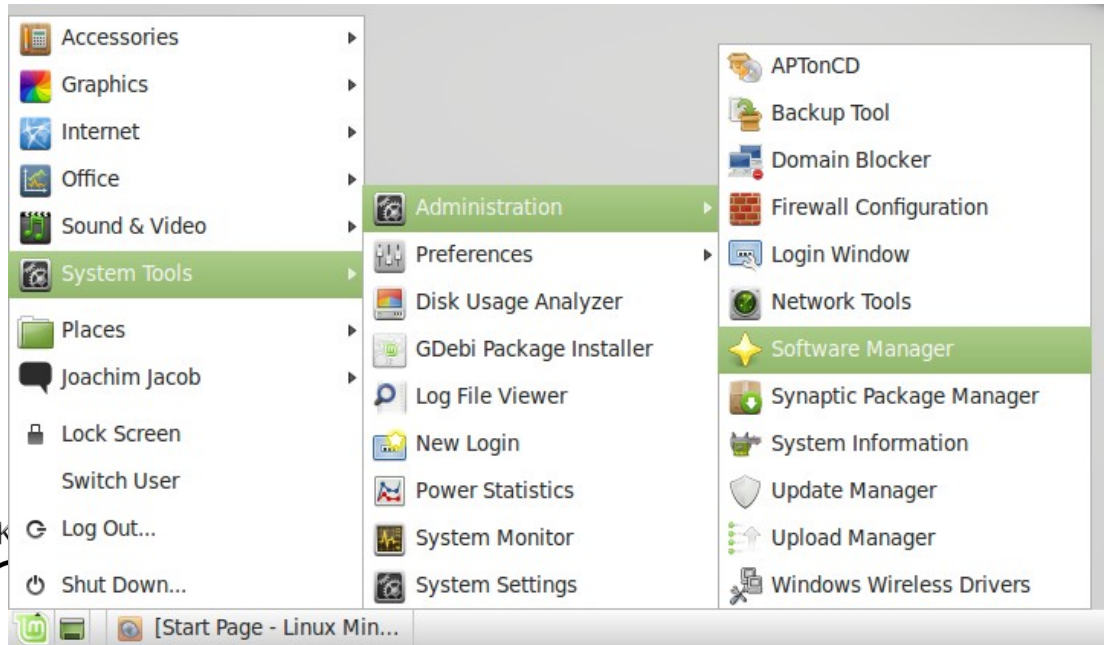
http://www.makeuseof.com/pages/best-linux-software

Installing: use the software center

PREFERRED WAY to install



Software center (= 'app' or 'application' store) every distribution as some kind of *software manager* : search for software and click to install. The software is automatically updated by the **update manager.** Depending on the distro, a lot of bioinformatics packages are available.



Example in Linux Mint

E.g Bowtie via software center

Bowtie

An ultrafast memory-efficient short read aligner

Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate
per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for
paired-end).

- Recent news
- Bowtie on GitHub - 4/11/13
  - Bowtie source now lives in a p

Terminal

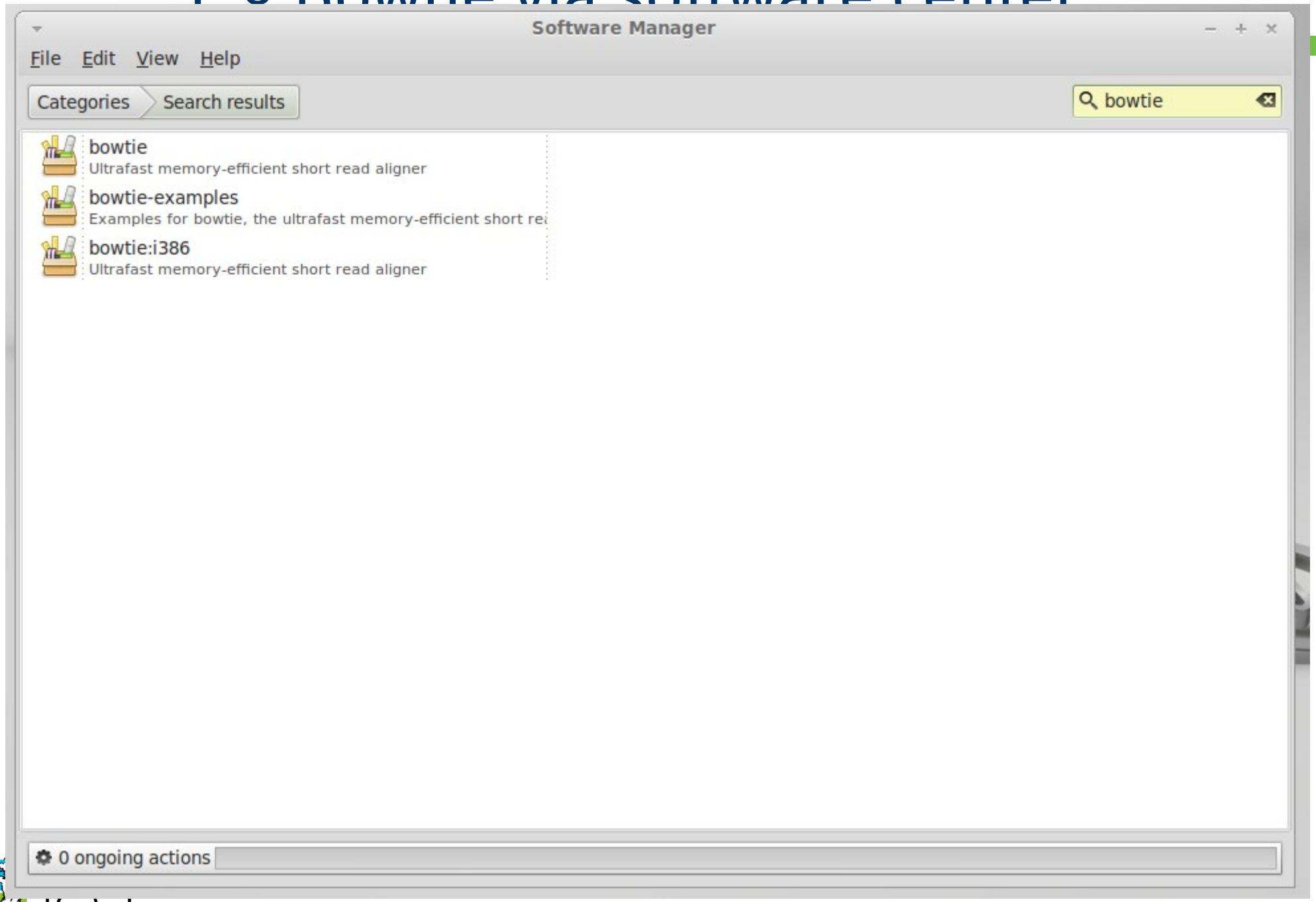
File Edit View Search Terminal Help

```
joachim@mint13 ~ $ bowtie
The program 'bowtie' is currently not installed.  You can install it by typing:
sudo apt-get install bowtie
joachim@mint13 ~ $
```

VIB JAKONIX IRU

E.g Bowtie via software center

E.g. Bowtie via software center



Software Manager

File Edit View Help

Categories > Search results > bowtie

bowtie

Ultrafast memory-efficient short read aligner

Score:

no reviews

Not installed — Install

This package addresses the problem to interpret the results from the latest (2010) DNA sequencing technologies. Those will yield fairly short stretches and those cannot be interpreted directly. It is the challenge for tools like Bowtie to give a chromosomal location to the short stretches of DNA sequenced per run.

Bowtie aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

http://bowtie-bio.sourceforge.net/

Details

Version: 0.12.7-1

Size: 1MB to download, 3MB of disk space required

Impact on packages: bowtie (installed)

Reviews

Your review: Submit

0 ongoing actions

Software Manager

File Edit View Help

Categories > Search results > bowtie

bowtie

Ultrafast memory-efficient short read aligner

Score:

no reviews

Not installed — Install

This package addresses the problem to interpret the results from the latest (2010) DNA sequencing technologies. Those will yield fairly short stretches and those cannot be interpreted directly. It is the challenge for tools like Bowtie to give a chromosomal location to the short stretches of DNA sequenced per run.

Bowtie aligns short ... wtie indexes the genome with a ... enome (2.9 GB for paired-end).

http://bowtie-bio.sc

Details

Version: 0.12.7-1

Size: 1MB to downlo

Impact on packages: bowtie (installed)

Reviews

Your review: Submit

0 ongoing actions

Authenticate

To install or remove software, you need to authenticate.

An application is attempting to perform an action that requires privileges. Authentication is required to perform this action.

Password:

Details

Cancel Authenticate

Software Manager

File Edit View Help

Categories > Search results > bowtie

bowtie

Ultrafast memory-efficient short read aligner

Score:

no reviews

Not installed

This package addresses the problem to interpret the results from the latest (2010) DNA sequencing technologies. Those will yield fairly short stretches and those cannot be interpreted directly. It is the challenge for tools like Bowtie to give a chromosomal location to the short stretches of DNA sequenced per run.

Bowtie aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

http://bowtie-bio.sourceforge.net/

Details

Version: 0.12.7-1

Size: 1MB to download, 3MB of disk space required

Impact on packages: bowtie (installed)

Reviews

Your review: Submit

1 ongoing actions 55%

Software Manager

File Edit View Help

Categories > Search results > bowtie

bowtie

Ultrafast memory-efficient short read aligner

Score:

no reviews

Installed — Remove

This package addresses the problem to interpret the results from the latest (2010) DNA sequencing technologies. Those will yield fairly short stretches and those cannot be interpreted directly. It is the challenge for tools like Bowtie to give a chromosomal location to the short stretches of DNA sequenced per run.

Bowtie aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

http://bowtie-bio.sourceforge.net/

Details

Version: 0.12.7-1

Size: 3MB of disk space freed

Impact on packages: bowtie (removed)

Reviews

Your review: Submit

0 ongoing actions

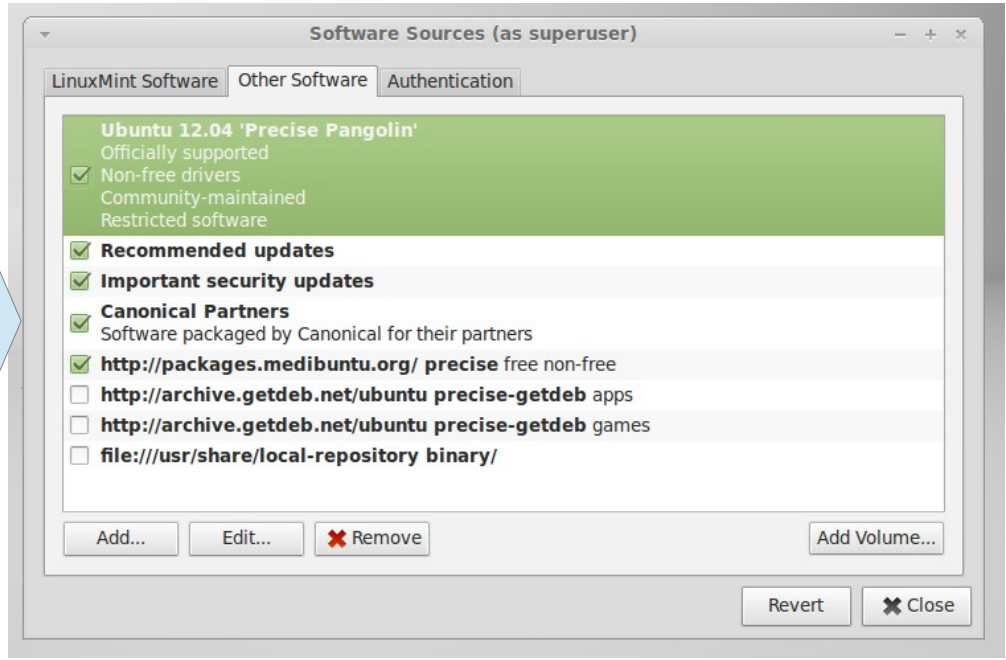
E.g Bowtie via software center

Terminal

File Edit View Search Terminal Help

```
joachim@joachim-VirtualBox ~ $ bowtie --version
bowtie version 0.12.7
64-bit
Built on allspice
Thu May  5 12:19:01 UTC 2011
Compiler: gcc version 4.6.1 20110503 (prerelease) (Ubuntu 4.6.0-6ubuntu2)
Options: -O3   -Wl,--hash-style=both -g -O2  -g -O2
Sizeof {int, long, long long, void*, size_t, off_t}: {4, 8, 8, 8, 8, 8}
joachim@joachim-VirtualBox ~ $
```

Software is fetched from repositories



On the internet, some URLs point to **software repositories** for Linux distributions. You can plug in repo's in the software manager. The repository hosts **installation files** for software. These are typically **.rpm** (Red Hat alike) or **.deb** (Debian alike distro's) files.

Official repositories: secure, high-quality, malware-free !

https://help.ubuntu.com/community/Repositories/Ubuntu
http://en.wikipedia.org/wiki/Software_repository

Example: the Debian Med repo

→ exercise at the end of this section

The Debian Med project prepares packages that are associated with medicine, pre-clinical research, and life sciences. Its developments are mostly focused on three areas for the moment: medical practice, imaging and bioinformatics.

DebianMed is a repository containing a lot of bioinformatics packages for Debian-alike distro's, such as Debian, Ubuntu, Mint,...

Debian Med repository is a PPA-type repo: a **Personal Package Archive (PPA)**. The link to such a repo starts with ppa://.

Downloading installation files

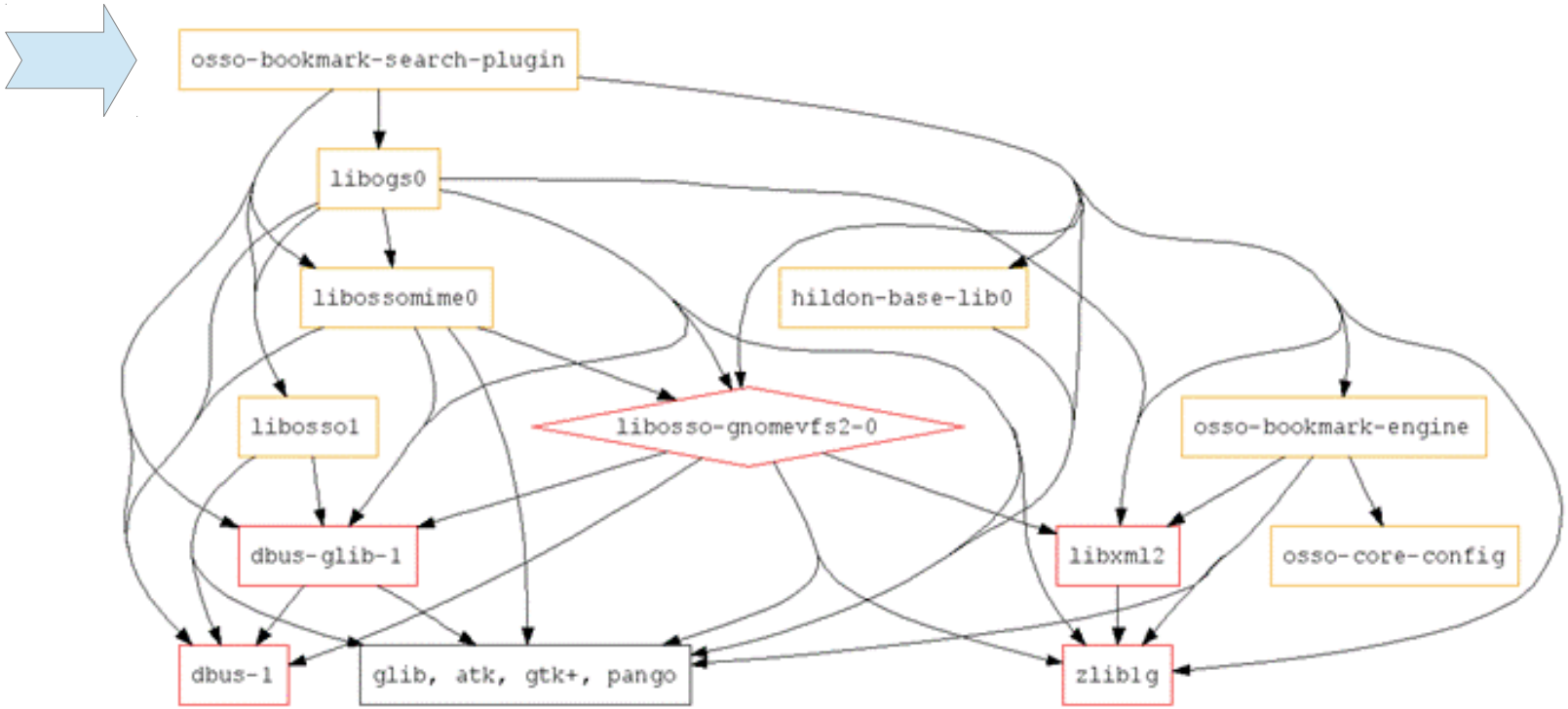
→ PREFERRED WAY number 2

Instead of searching with the Software Manager, the installation files (**.rpm** or **.deb**) can be downloaded from internet separately (e.g. when they're not (yet) in a repository). The Software Manager will install the software contained in these files (usually double-clicking the install file).

Note: **no secure** transfer and **no confirmation** of the package, so be a bit more careful.

Unix philosophy: software interconnects

Software should do one specific task, avoiding redundant code by reusing other software code. This creates **dependencies** between packages. Below a graphical representation of dependencies of a package.



Dependencies **need to be co-installed** with the software if not present. **.rpm/.deb files take care of this**! (and hence also the software manager)

Dependency example

PACKAGES

debian

About Debian Getting Debian Support Developers' Corner

/ packages / squeeze (oldstable) / science / altree

Search package names all options

[Source: altree]

[squeeze] [wheezy] [sid]

Package: altree (1.0.1-3)

program to perform phylogeny based analyses

ALTree was designed to perform phylogeny based analysis: first, it allows the detection of an association between a candidate gene and a disease, and second, it enables to make hypothesis about the susceptibility loci.

Tags: Field: Biology, Bioinformatics, Implemented in: C, Perl, User Interface: Command Line, Role: Program, Shared Library, Scope: Utility, Purpose: Analysing, Comparing, Supports Format: Plain Text

Other Packages Related to altree

● depends ■ recommends ◆ suggests ▲ enhances

- **dep: libc0.1 (>= 2.3) [kfreebsd-amd64, kfreebsd-i386]**
  Embedded GNU C Library: Shared libraries
  also a virtual package provided by libc0.1-udeb
- **dep: libc6 (>= 2.7-1) [not ia64, kfreebsd-amd64, kfreebsd-i386]**
  Embedded GNU C Library: Shared libraries
  also a virtual package provided by libc6-udeb
- **dep: libc6.1 (>= 2.7-1) [ia64]**
  Embedded GNU C Library: Shared libraries
  also a virtual package provided by libc6.1-udeb
- **dep: libgcc1 (>= 1:4.3) [armel]**
  GCC support library
- **dep: perl (>= 5.10.0-10) [not amd64, kfreebsd-amd64, kfreebsd-i386]**
  Larry Wall's Practical Extraction and Report Language

Links for altree

No screenshot available

Debian Resources:
- Bug Reports
- Developer Information (PTS)
- Debian Changelog
- Copyright File
- Debian Patch Tracker

Download Source Package altree:
- [altree_1.0.1-3.dsc]
- [altree_1.0.1.orig.tar.gz]
- [altree_1.0.1-3.diff.gz]

Maintainers:
- Debian-Med Packaging Team (QA Page, Mail Archive)
- Charles Plessy (QA Page)
- Vincent Danjean (QA Page)
- David Paleino (QA Page)

External Resources:
- Homepage [claire.bardel.free.fr]

Similar packages:
- r-other-mott-happy
- plink
- r-cran-genabel
- fastlink
- fastlink-doc
- fcitx-module-cloudpinyin
- r-cran-epir

VIB JAKONIX IRU

http://packages.debian.org/squeeze/altree

Exercise: example of installing a .deb file

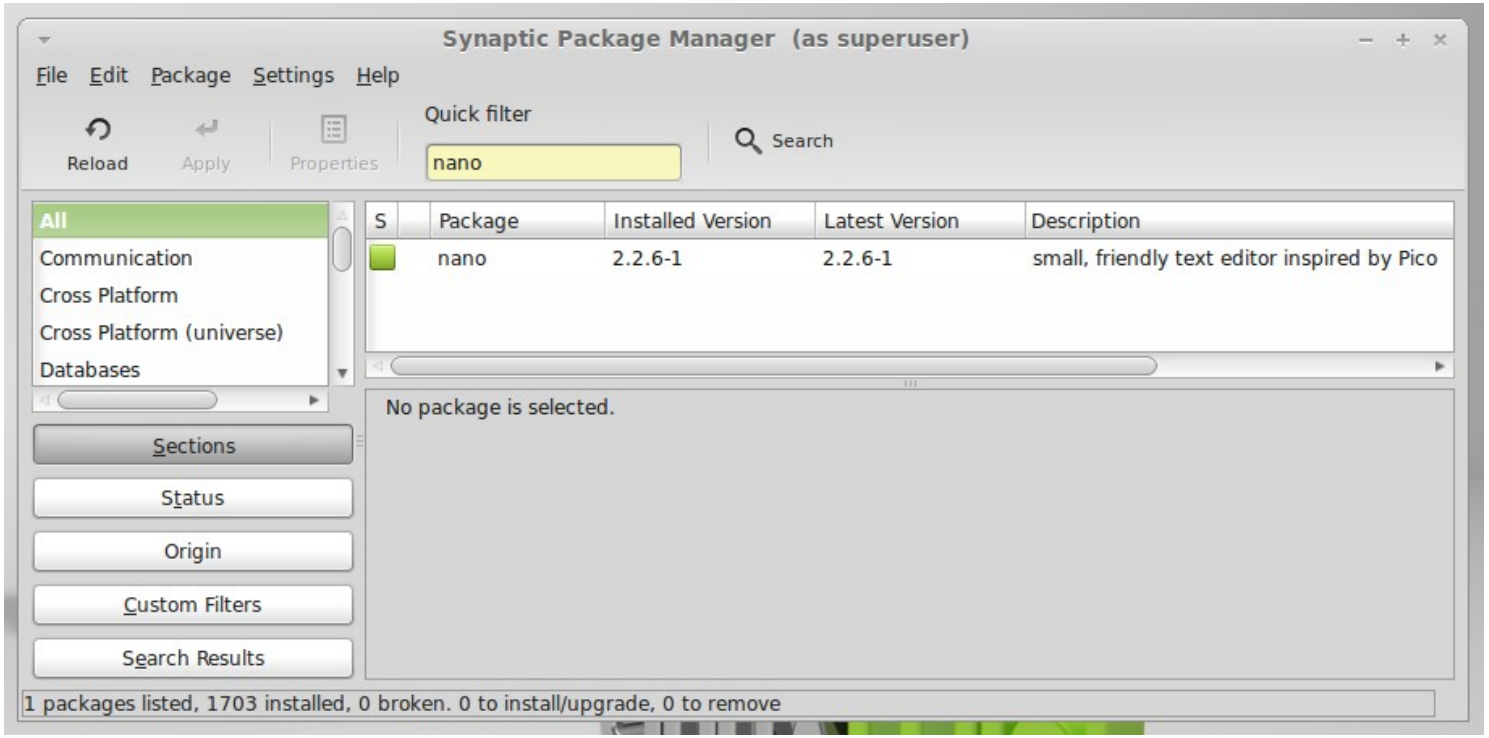
→ Exercise link: 2 exercises!

1. install the multiple sequence alignment tool ClustalW.
2. install Gk-arrays

http://packages.debian.org/squeeze/altree

Software center versus package manager

A **package manager** allows more fine-tuned package installs and more info (e.g. repo info, libraries with code to be shared between programs)

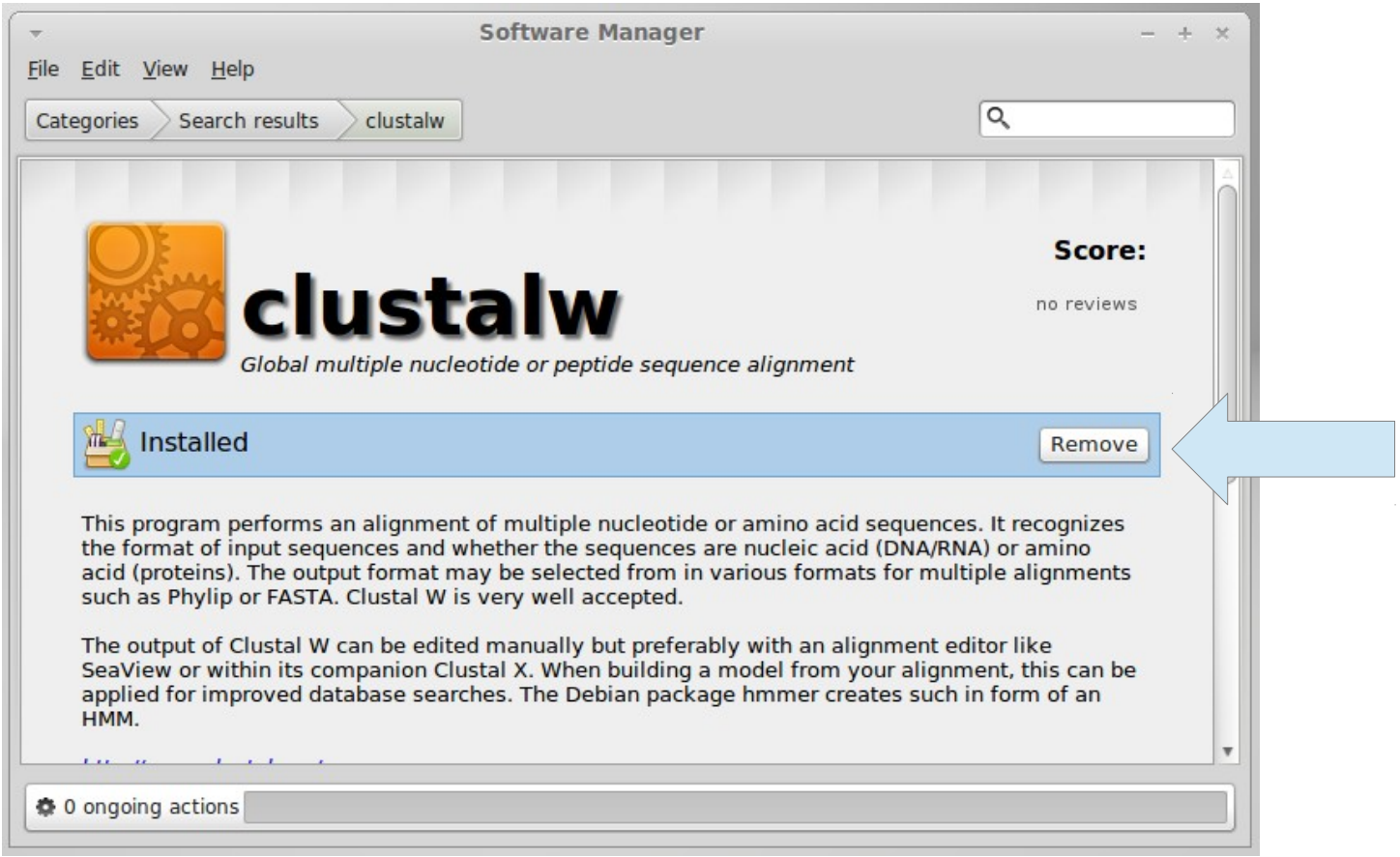


A **Software Manager** offers installation of complete sets of packages constituting one (often GUI) tool. It has also a user rating functionality.

Uninstalling software

If your software manager has installed software, you can delete software from within the interface.

The software manager knows exactly where files have been installed.



Software that is not packed

→ LESS PREFERRED way number 1

Software can come as a **compressed file,** which contains **source code**. This 'human readable' source code needs to be **compiled** first before being usable. Compiling creates executable machine code (a binary) which you can execute. Scientific papers often distribute code in this form, (before eventually being incorporated in a repository).

transpose-2.0.zip

tsne_0.1-2.tar.gz

mcl-latest.tar.gz

- Usually, the source code comes as a **.tar.gz** or **.tar.bz2** compressed file.
- Compiling: a process that is carried out via the terminal (see later)

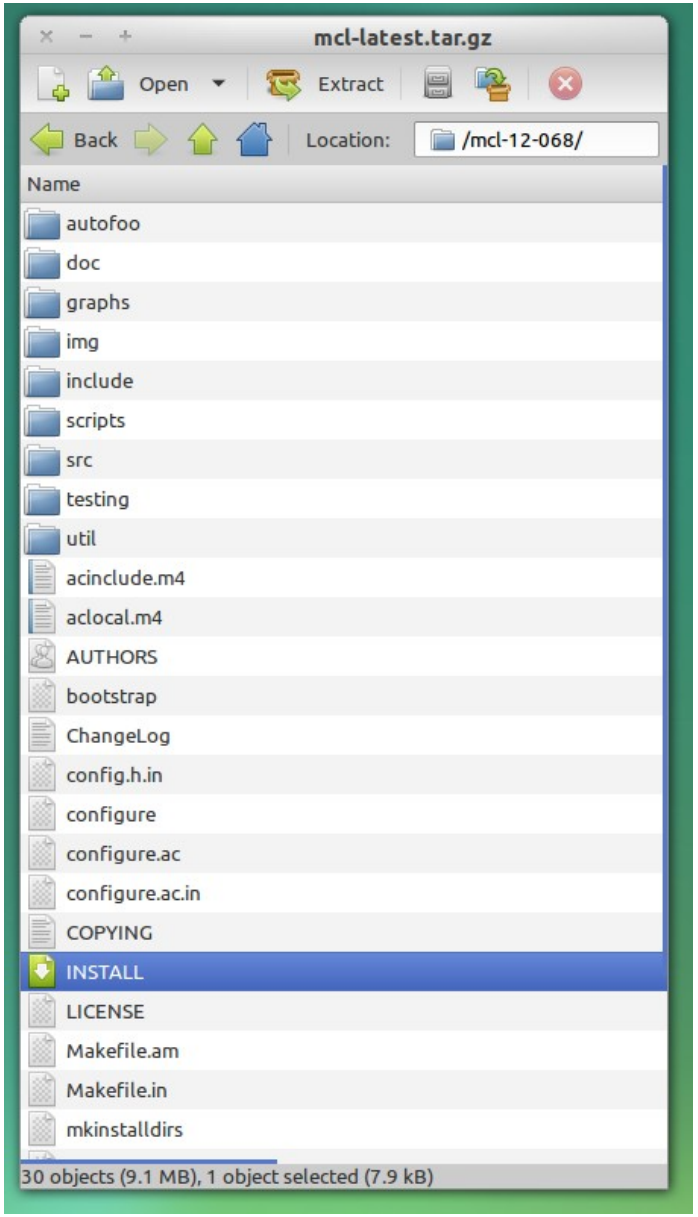
http://www.micans.org/mcl/ and http://sourceforge.net/projects/transpose/

Software that is not packed

Issues:

- You need to install the **dependencies** yourself (best via the software manager).
- Be **organised**: in which folder will you put the software?
- After the software is compiled, you need to **make it available** in your system. (see later)
- It is not easy to **delete** the software: you need to manually remove the files everywhere you have put them.

Summary: a lot of hassle!

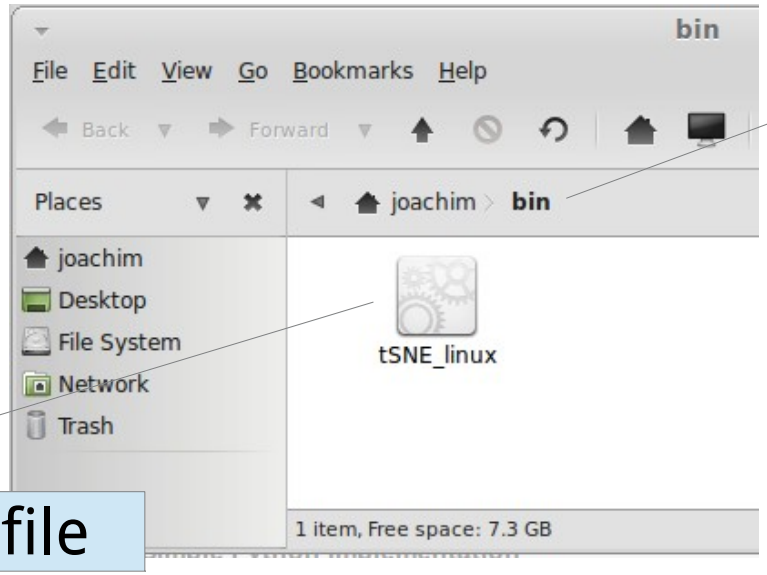


VIB JAKONIX

Software that is not packed

→ LESS PREFERRED way number 2

Sometimes, software is compiled for you, and the resulting **binary file** can be downloaded.
Attention: the binary that is build needs to match your machine *architecture* (usually 64 bit).



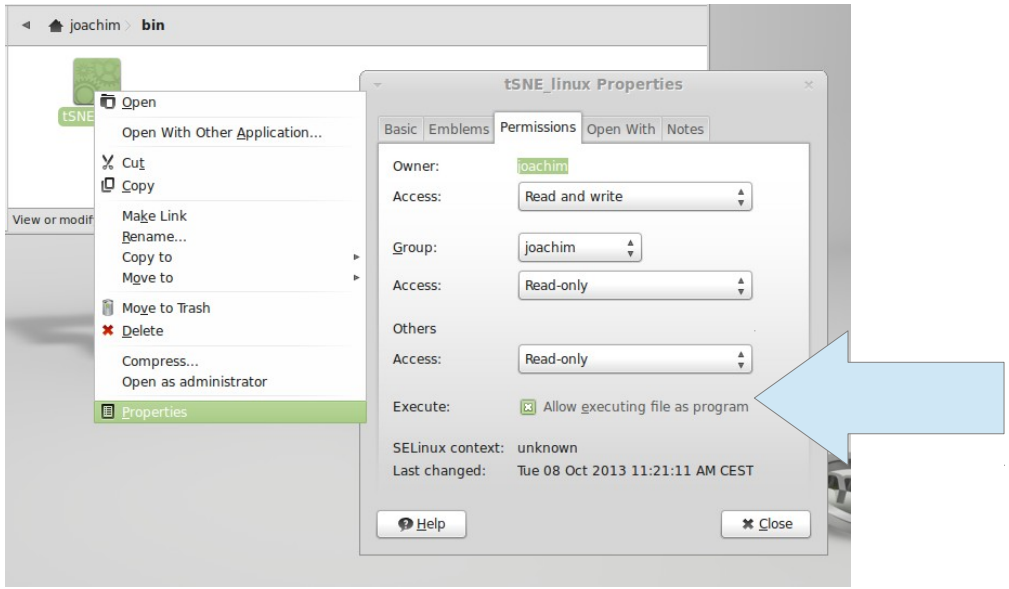
A folder called 'bin' contains executable binary files (the program files)

The binary file

Example: http://homepage.tudelft.nl/19j49/t-SNE.html clustering tool

How to run binaries?

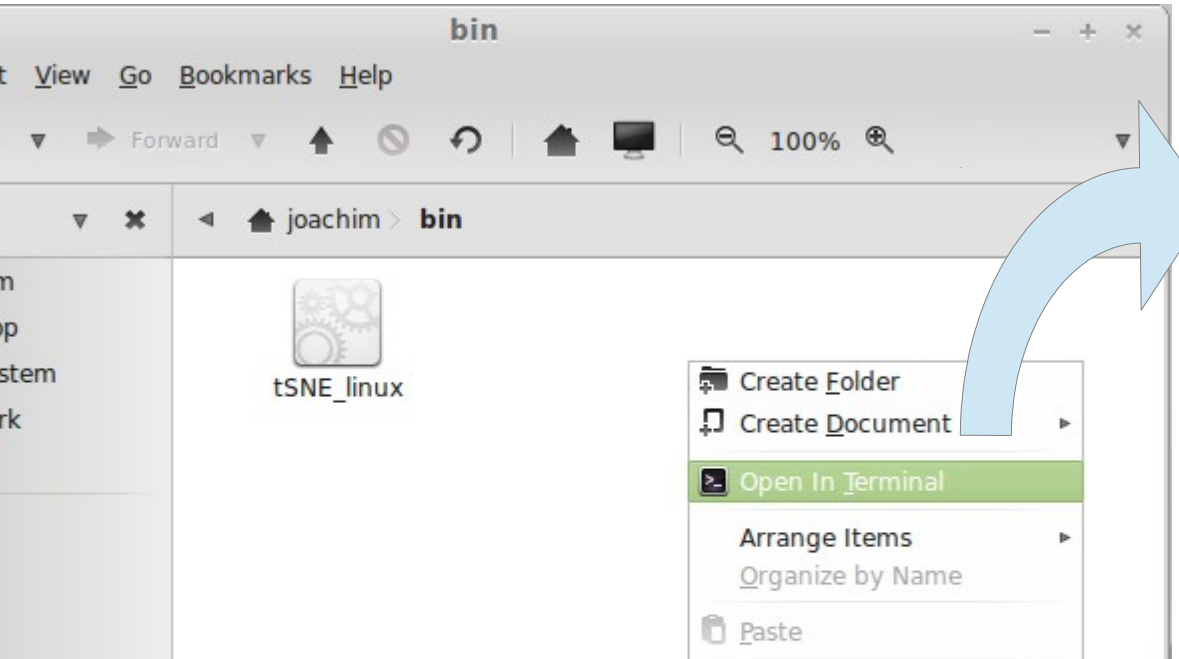
- Normally, you cannot run downloaded binaries as a safety measure. You need to set the permission for this file to '**executable**'. Easily do this by: right-click → properties

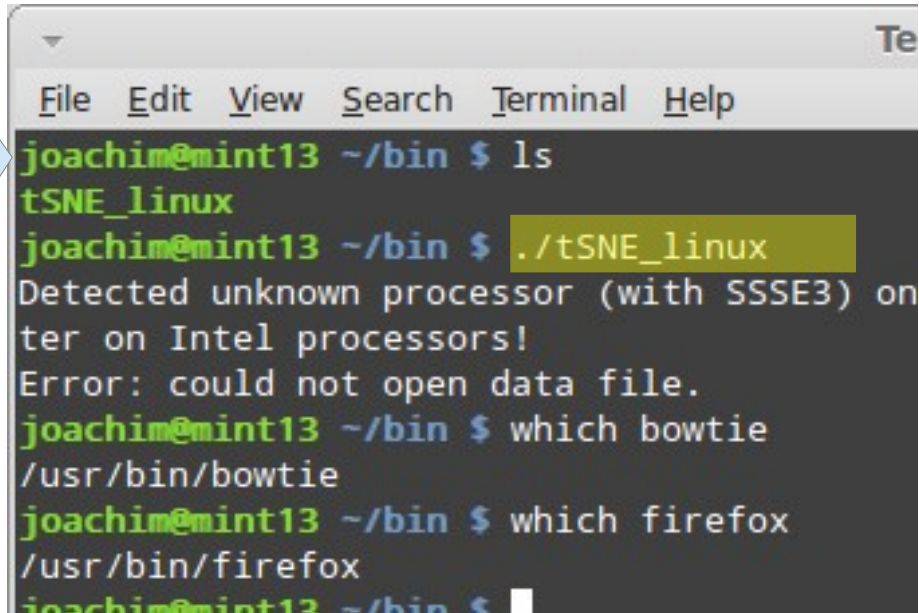


- If your program is designed to work in a graphical environment, double-clicking the binary launches the program.

A program is a binary file

- Every program,installed through packages, or installed manually, also on your desktop, can be started on the command line in the terminal.
- Some binaries, and most bioinformatics programs can **ONLY** be run from the command line (covered later in detail).





Java programs are binary files

- Some bioinformatics programs come as Java code: **program.jar**
- To run this, double-click, or type:

```
$ java -jar program.jar
```

- Example: Picard tools (http://picard.sourceforge.net)

Overview of Picard command-line tools

The Picard command-line tools are packaged as executable jar files. They require Java 1.6. They can be invoked as follows:

```
java jvm-args -jar PicardCommand.jar OPTION1=value1 OPTION2=value2...
```

Most of the commands are designed to run in 2GB of JVM, so the JVM argument -Xmx2g is recommended.

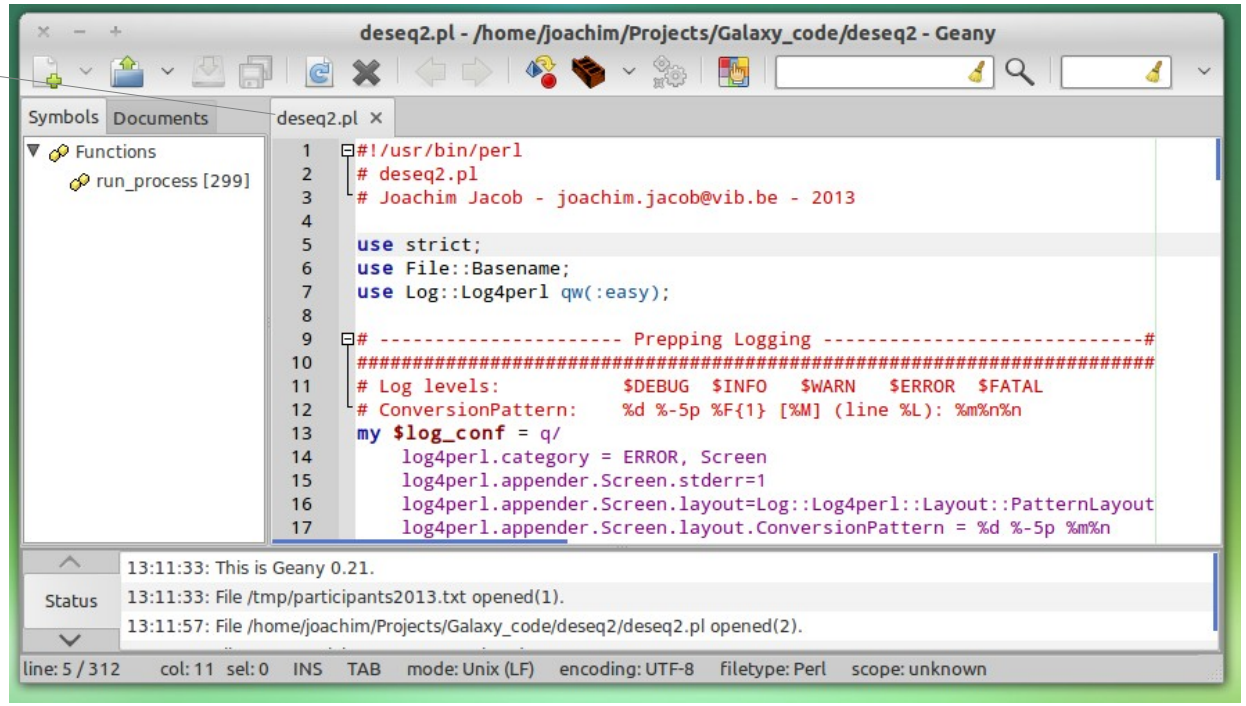
Why different methods of distributing?

- *Commercial software* is usually **binary** only (protection of the source code) – installation instructions are provided by the vendor.
- *Free (Open Source) Software* is usually distributed in **source packages**.
- *Packaging*, the process of creating .rpm or .deb packages takes a lot of time, therefor often the **source code** to compile for yourself is provided.

Scripts are human readable programs

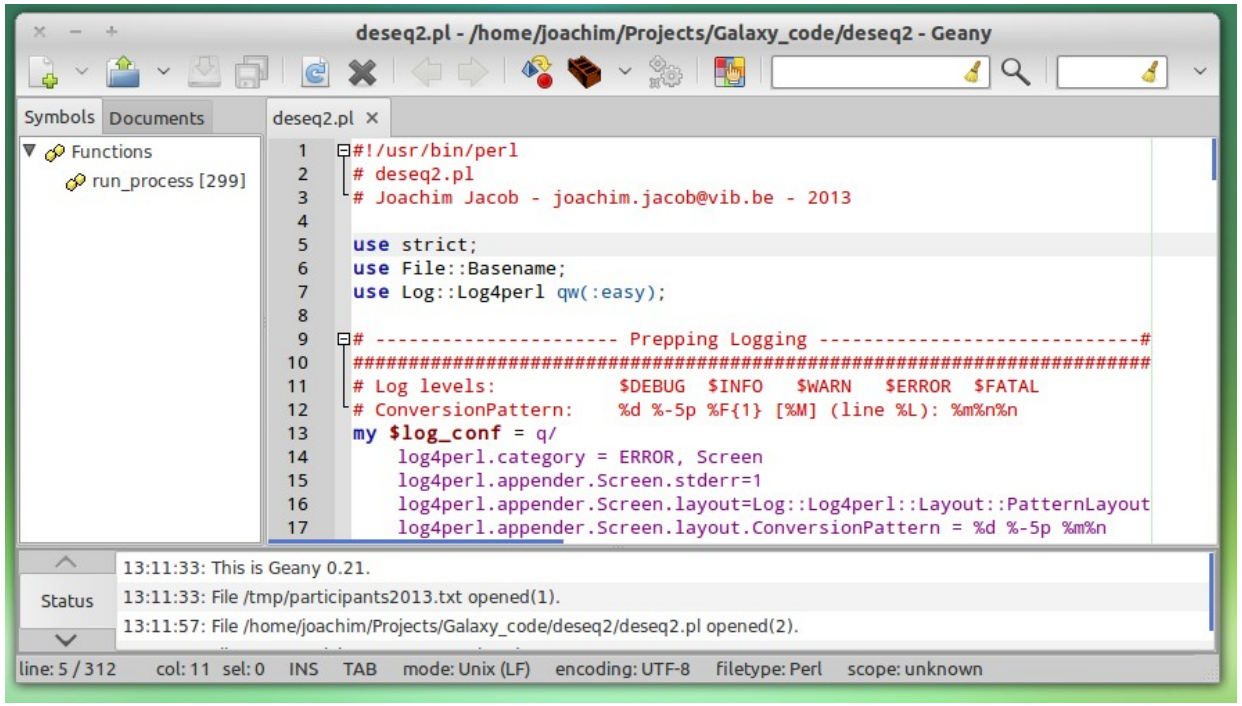
- *Software* exists that reads in **text files** containing instructions, to be executed by the computer. These text files are called **scripts**. They are not binary files. But they are **executable**.
- E.g. perl, python, R, bash

A **script** which just contains text, and can be **interpreted** by perl



Scripts are human readable programs

Scripting languages are very popular in bioinformatics, because of their relatively low barrier to get starting, their **platform independence**, and quick and dirty approach (in the case of Perl), and **easy sharing**: just download the script and execute it (permissions: read and execute).



Summary of this section

- Software **manager**…
- … taps software from different places, called **repositories**.
- You can add repositories by adding the **URL**
- Software is **compiled** and installed on your machine, either by the manager, or manually.
- … or it consists of **scripts**, which are interpreted by an interpreter real-time.

Keywords of this section

software center
package manager
Repositories / software sources
dependencies
RPM and deb files
compiling of source code
binaries
packaging
executable
scripts

Write in your own words what the terms mean

Exercise: Getting software

→ Adding software sources containing bioinformatics packages
→ Install software from a package file graphically.

→ Install the very good text editor Geany (PPA exercise)
→ Install the very good terminal program **Terminator** (Software center)

→ Good bioinformatics packages: Ugene – interface to many algorithms.