The New Zealand Institute for Plant & Food Research Limited

FLOSSing in the Lab

Plant and Food's use of Free/Libre Open Source technologies

Zane Gilmore, Ben Warren, (Eric Burgueno, Roy Storey)

FLOSSing in the Lab

What you are in for:

- Who is Plant & Food?
- What do we do?
- Why do we need software?
- Why we use OSS
- Some examples
- Genetic science
- Genetics and FLOSS

Crown Research Institutes

- AgResearch
- ESR
- Scion
- GNS
- Landcare Research
- NIWA
- Plant & Food Research

Who we are

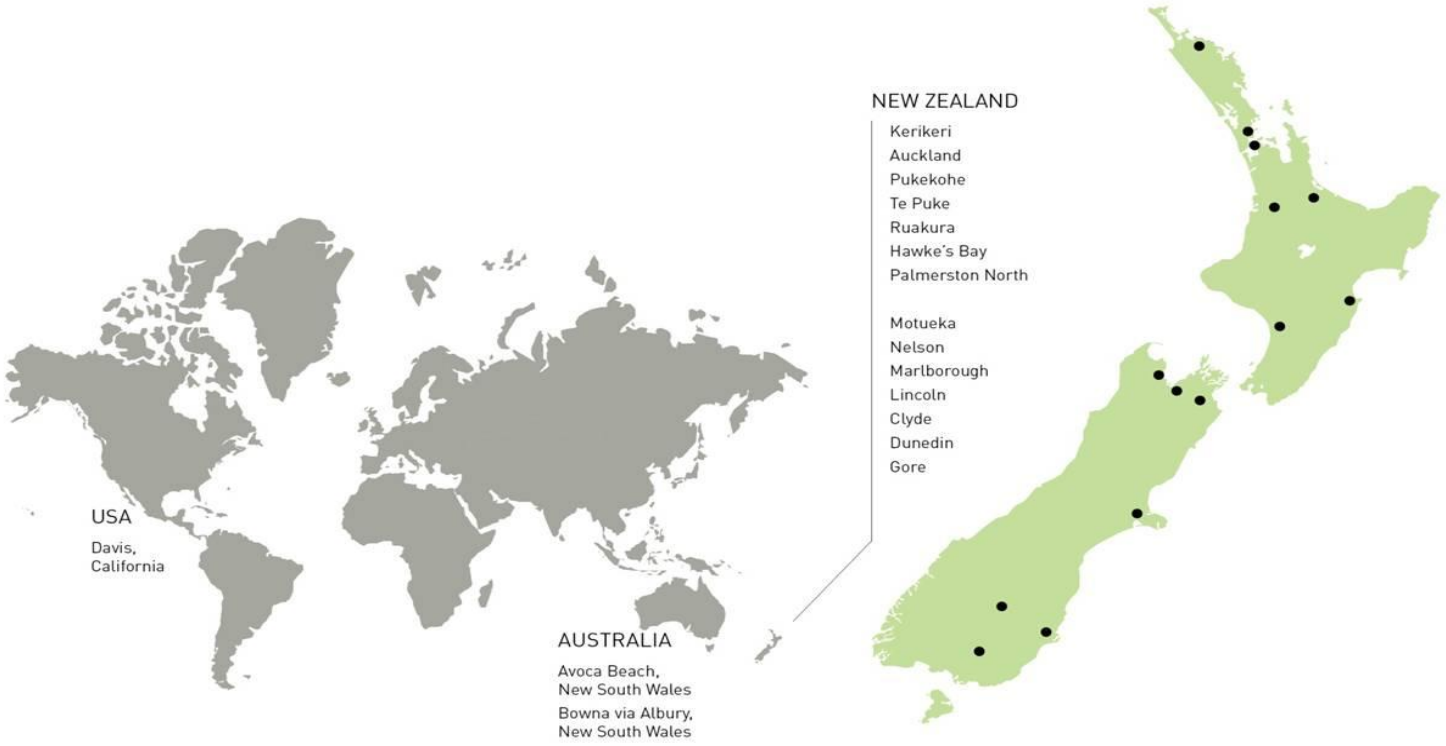
- Based in New Zealand
- Government-owned Crown Research Institute
- Revenue NZ$119.6 million (2013/14)

A mix of private contracts and royalties, and NZ Government contracts

Over 900 employees

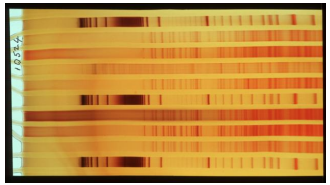
- 650 research staff
- 2 dedicated programmers
- 15 sites in New Zealand
- Representatives in USA, Australia

Our Locations

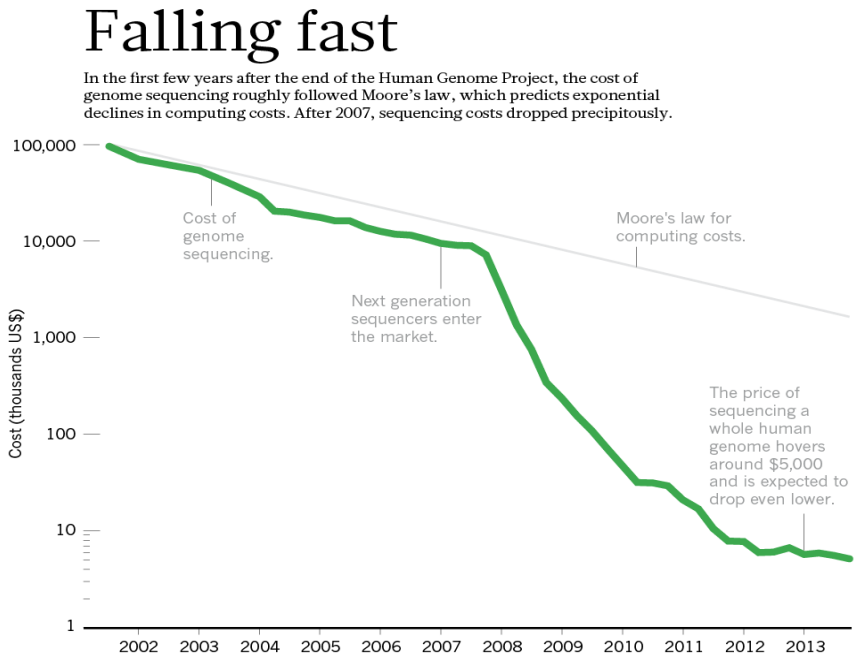


What PFR does

- Plants
  - Breed new cultivars
  - Cultivation
  - Diseases
  - Insect pests
- Food
  - Nutritional health
  - Nutrient analysis
  - Food manufacturing
- Seafood and fishing
- Other stuff but mainly in the service of, or related to the above e.g. soil science and electro-spinning



Computing problems we face



http://www.nature.com/news/technology-the-1-000-genome-1.14901

Reproducible research

IP[y]: Notebook Seaborn TSPlot and ggplot Last Checkpoint: Jan 08 15:19 (autosaved)

File Edit View Insert Cell Kernel Help

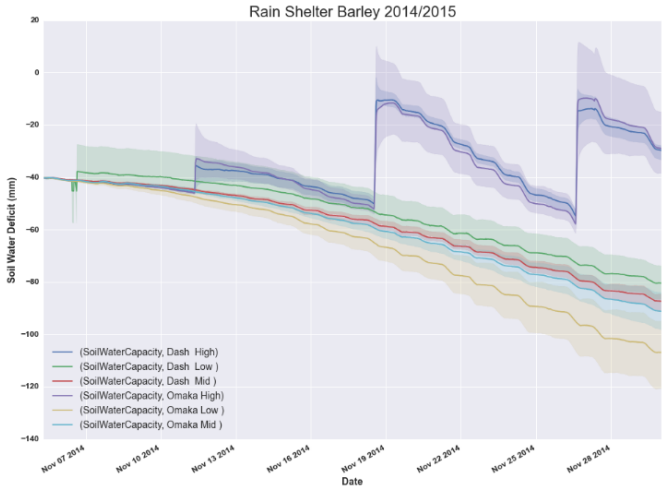
Code Cell Toolbar: None

```
    ax.fill_between(x, low, high, alpha=.2, color=palette.pop(0))

plt.title("Rain Shelter Barley 2014/2015", fontsize=22, fontweight="roman")
plt.ylabel("Soil Water Deficit (mm)", fontsize=14, fontweight="bold")
plt.xlabel("Date", fontsize=14, fontweight="bold")
plt.xticks(fontsize= 12, fontweight="semibold")
plt.yticks(fontsize= 12, fontweight="semibold")

ax.legend(loc=3,prop={'size':14})
```

Out[12]: <matplotlib.legend.Legend at 0x7bf3c630>



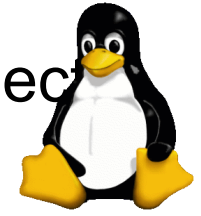
In [14]:
```
import mpld3
```

In [35]:
```
data=SoilWaterDeficitData.groupby(["time", "Group"]).mean().unstack("Group")
#fig1, ax=plt.subplots(111); data.plot()
fig1=plt.figure(figsize=(16,12)); data.plot()
#ax=plt.subplot(111)
#SoilWaterDeficitData.groupby(["time", "Group"]).mean().unstack("Group")

x=SoilWaterDeficitData.groupby(["Group", "time"]).mean().unstack("Group").index
```

FLOSS issues

- Biologists often aren't at home in the world of computing
- Managers (who are often biologists) don't unders FLOSS concepts
- CRI funding model
- Geneticists ARE good informaticians
- Battle is not futile as scientists are clever and respect data



Food Composition (FCDB)

- > 2600 Foods
- > 300 Nutrients/Components/Attributes
- > 400 recipes
- Produce Food Files for Ministry of Health
- Present system is old and creaky
- Data has high “coolness coefficient”
- www.foodcomposition.co.nz
- We are going to rebuild it

MINISTRY OF HEALTH
MANATŪ HAUORA

More FCDB

- Attribute calculator
- Recipe calculator
- Recipes of Recipes
- Meat pie example
  - Recipe for pastry
  - Recipe for meat stew filling

Kea

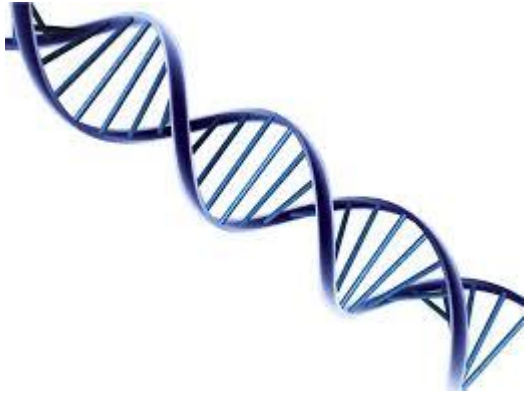
- Plant breeding needs to be done faster
- We use genetic and chemical analysis for breeding decisions
- Thousands of plants
- Kea sample tracking (in-house then with help from Encode)
- Linux-Django-Postgres stack with Elastic search
- Just produced alternative provenance system
- Working on getting it Open Sourced

Other stuff

- Data loggers: Lysimeters, rain-shelters
- Chemistry databases
- Continuous requests

Next Guy

Time for Ben



FLOSSing in the Lab

What you are in for:

- Who is Plant and Food?
- What do we do?
- Why do we need software?
- Why we use OSS
- Some examples
- Genetic science
- Genetics and FLOSS

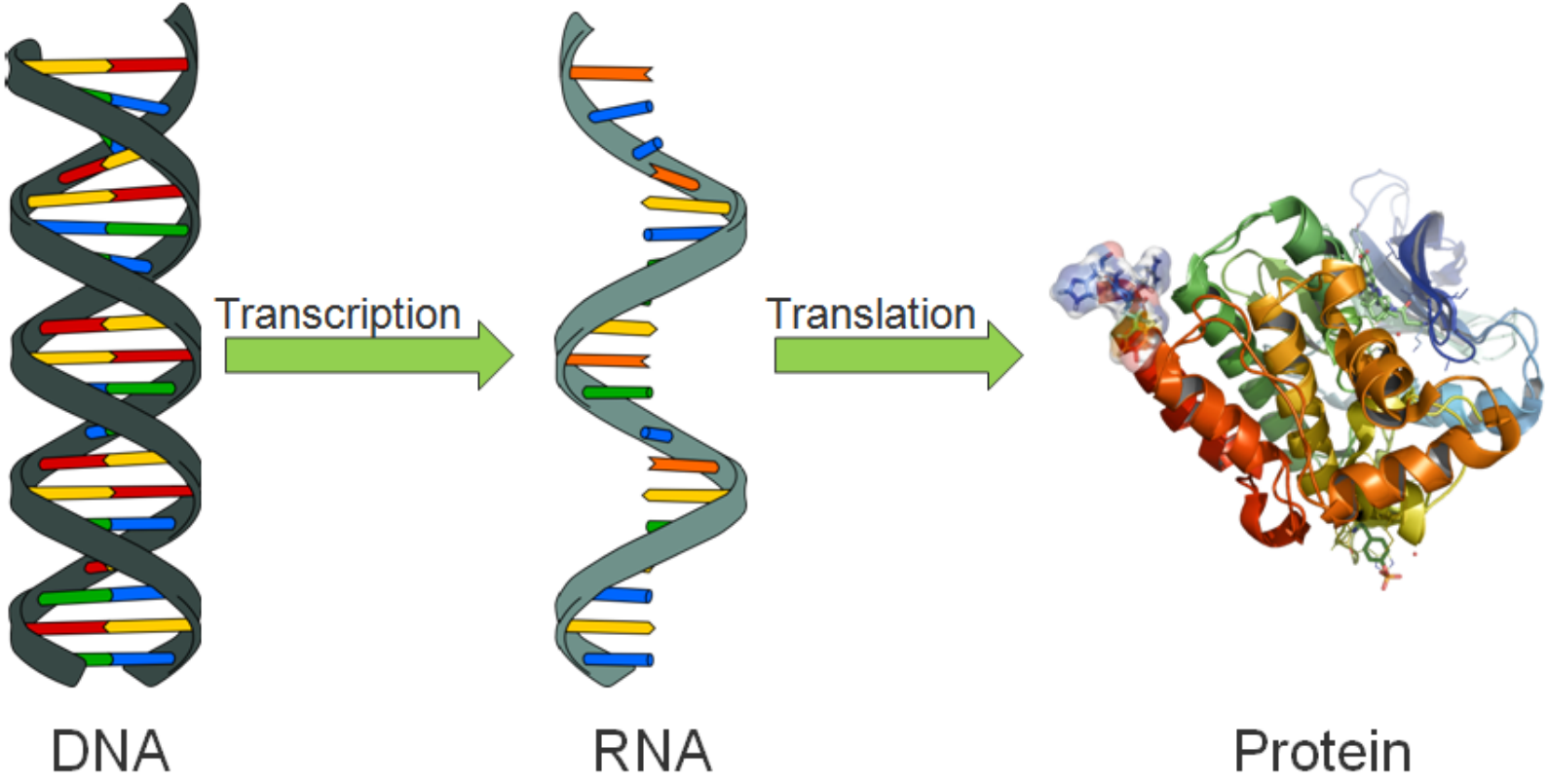
We Do *omics

What is an *omics?

There are many species of *omics.

In the bioinformatics department at PFR we mainly do **gen**omics and **transcript**omics. This is the study of the genome (DNA) and the transcriptome(RNA) respectively.

The Central Dogma



Genome Assembly - A Computational Problem

The assembly problem:

Given **N** of the same textbooks (possibly differing editions) cut into strips and put in a pile, reconstruct the **N** original texts.

We Need Software for Computation

Assembly and other *omics tasks often require large computations.

- openLava[1] - Job scheduler Software
  - Assign jobs to appropriate nodes
  - Priority queues
- powerPlant - Compute cluster
  - Shared data store (~1PB)
  - Virtual compute nodes
  - Physical compute nodes (e.g. 2TB of memory)

We Need Software for Visualisation

Visual representations of data enhance understanding and spark new ideas about data.

Ensembl[2] allows us to visualise genomic data.

- Can incorporate user data easily
- Extendable and customisable

Ensembl - Wine Grape Genome

We Need Software for Reproducible Research

- A workflow is a recipe describing how to get from input data to results
- A well-documented workflow allows the process to be reproduced exactly
- This is necessary for;
  - transparency
  - verification
  - sanity

We Need Software for Reproducible Research

Moa[5] provides extendable templates based on common workflows.

“Moa hopes to make meticulous organization of a command line project much less of a burden - leaving you to focus on the fun parts.” - Mark Fiers, http://moa.readthedocs.org/en/latest/

- Integration with Git
- Integration with openLava

We Need Software for Reproducible Research

We can use Git[3] to store workflows, allowing reproduction of the workflow at any version.

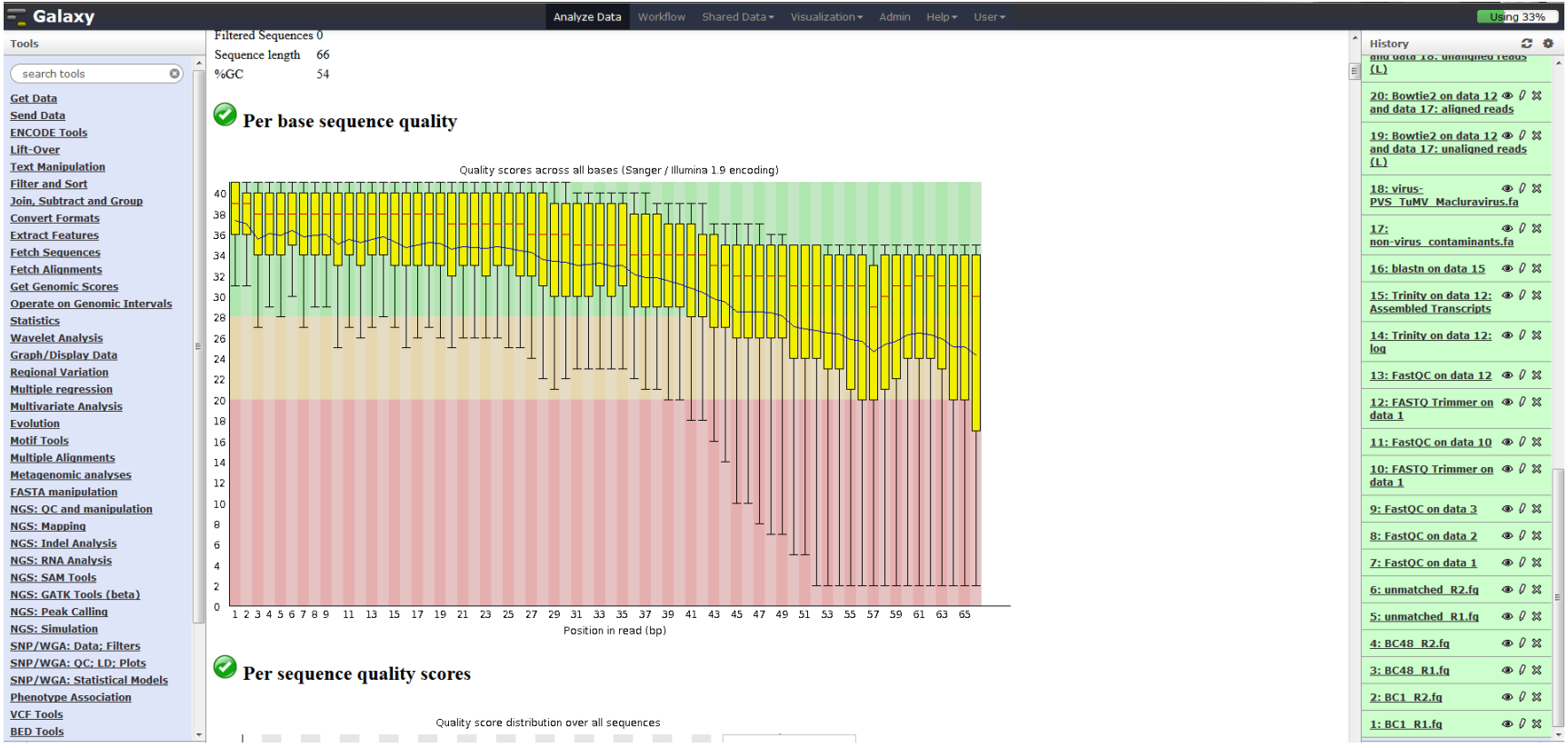
- Branches can store specific instances of a workflow
- Github[4] allows easy workflow sharing and collaboration on development

We Need Software for Scientists

Galaxy[6] delivers:

- A GUI to command line tools
- History of processes
- Construction of workflows
- Running workflows
- Integration with job schedulers
- Per-user management
- Extendable tool suites

Galaxy Example



Why FLOSS?

- **Open:** Similar philosophy to scientific research
- **Current:** Keeps up with the scientific community
- **Community:** Collaboration, knowledge sharing
- **Flexible:** Adaptation to related problems
- **Trust:** Scientists do not trust what they cannot read/understand

References

1. www.openlava.org

2. www.ensembl.org

3. git-scm.com

4. github.com

5. https://github.com/mfiers/Moa

6. galaxyproject.org