Bioinformatics core facilities

Small bioinformatics core on a (modest) budget

Wellcome Trust
Centre for Cell Biology
WTCCB @ EDINBURGH
...let's clear up some stereotypes

Canny

/kani/ adjective

1. having or showing shrewdness and good judgement, especially in money or business matters.

synonyms:

shrewd, astute, sharp, sharp-witted, discerning, acute, penetrating, discriminating, perceptive, perspicacious, clever, intelligent, wise, sagacious, sensible, judicious, circumspect, careful, prudent, cautious;
My background and prejudices

1993: Started my PhD in Bioinformatics at Nottingham, UK

Postdoc MBL, Cape Cod, MA, USA

Had to learn some Unix sys-admin

Teaching reduced frequent trivial requests (BLAST jobs & alignments)

Corporate sys-admin policies vs rapid evaluation & prototyping

Stung by closed data in commercial software

2 Industry positions, Biotech (Cambridge MA, USA), & Pharma (Newhouse, Scotland)
Our Bioinf-core

Part of the 5 year Wellcome grant for the centre
No grant writing
Petition for core funds
Advise PIs to add costs to their grants

Extra core person in 2009 (Shaun Webb)
Help from some advanced users

~20 groups with ~190 students & staff.
Support some of the advanced users

2004: Sequence analysis, Microarrays and some processed mass-spec and microscopy data

2008: as before + *-Seq
Core Facility: Deliverables

- Teaching
- Collaborations
- Hardware
- Software
- Data Management

Infrastructure
• Open software and data just makes sense
  – Many mature software projects, well supported by the community
  – I attend BOSC most years for ideas for new installs

• Commercial disasters: painful data recovery..
  – [2003] Incyte suite: dropped after cost increase
  – [2008] VectorNTI: free version went commercial

• But we still use some
  – LaserGene from DNASTar
Reproducible Research

- GEO: Gene Expression Omnibus
- ENA: European Nucleotide Archive
- Short Read Archive
- Galaxy Tool Shed
- GitHub
Biologists (usually) need GUIs

2004

2005

2007

2013
Galaxy

• Adopted in 2006 and replaced taverna. Still in use and is our main environment for non CLI users.

• Reasons for adoption
  – Easy custom GUIs
  – Rapid prototyping of code via user input
  – Workflows
  – Data libraries
Data sharing

Libraries with group and user permissions

For viewing in Ensembl, IGV, IGB, UCSC etc:

URL's for Bigwig and bed files
Sharing results of data analysis from collaborations

Distributed annotation system [DAS]
Data Sharing and Visualisation

• R Shiny
  – Very fast way to create applications

• Shiny Tables
  – Removes the need for excel
  – Custom searching allow for a better use experience

• Shiny Graph
  – Powerful tool for data exploration, especially when using ggplot2 with faceting or colouring
Rack Server Hardware

• Disk:
  – ~500TB split over backup and primary storage
  – Using ZFS on newer servers

• Compute: 3 key servers
  – 64 GB RAM, 24 logical cores
  – 256GB RAM, 64 logical cores
  – 512GB RAM, 64 logical cores

• Total Cost new ~£70,000 (UK pounds)