Leveraging High Performance Computing for Genotyping and Phenotyping on the Cloud

Olivier Salvado, Group Leader, CSIRO Biomedical Informatics
August 2015
CSIRO Biomedical Informatics group

4 Teams
~20 staff
15 PhD students
50 journals in 2014

Medical Image Analysis (Jurgen Fripp, Brisbane)
Develop software to extract clinical information from medical images to reduce costs, provide new insights, improve diagnostics

Clinical Imaging (Steve Rose, Brisbane)
Optimise identification of injury, pathology, and treatment response using advanced medical imaging technologies

Transformational Bioinformatics (Denis Bauer, Sydney)
Develop and apply advanced computational and statistical methodologies to large datasets in the health/life sciences space

Biodata Visualisation (Sean O'Donoghue, Sydney)
Develops novel visualisation methods that enable new insights from biological and health data
DNA | Proteins | Phenotype | Treatment

Visualisation

Genomics

Biostatistics

Image analysis

BIG DATA

- Large clinical studies
- Population wide analysis
- Precision medicine
Aquaria: simplifying insight & discovery from protein structures

O'Donoghue et al., Nature Methods, 2015
Minardo: Layout strategy for proteomics time-series data


Turning the hairball into knowledge

http://minardo.org
Reflect, Compartments & Tissues: 3 tools from text mining

O'Donoghue et al., J. Web Semantics (2010); Binder et al., Methods (2014); Santos et al., PeerJ (in press)

Download Reflect Plug-in

Your download should begin in a few seconds, but if not, click here.

The plug-in adds a Reflect button to your browser: clicking this button tags genes, proteins, chemicals, and Wikipedia terms in any web page.

To learn more about Reflect, go to the About page.
Cost of genome sequencing
Exactly how large is “large volumes”?  

We can sequence the entire earth genome in 10 days.  

10^7 species x 10^8 bp genome size = earth genome 10^{15} bps  
1000 HiSeqsa can sequence 10^{11} bps per day  

Adapted from Prof. Knut Reinert (FU, Berlin)  
Inspired by anzska information booklet
Revolutionizing clinical practice through genomics

<table>
<thead>
<tr>
<th>Disease Burden</th>
<th>Rare variant detection</th>
<th>Efficient patient selection for research</th>
<th>Population screening for ranges of markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Rare genetic diseases</td>
<td>complex genetic diseases</td>
<td>disease with genetic &amp; environmental component</td>
</tr>
<tr>
<td>Most Benefit</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Impact on Health System

- Conventional biomarkers
- Genome sequencing analysis
- Electronic health records
- 'Omics technology integration

Technological investment

Research@CSIRO: Omics Project

Full Omics Project Cohort: 500 (178 to date) individuals from colorectal resection at the John Hunter Hospital, Newcastle Private Hospital and Royal Newcastle Centre (surgeons Dr Brian Draganic, Dr Peter Pockney & Dr Steve Smith) organized by Dr Desma Grice and Prof Rodney Scott (University of Newcastle)
AIM: accredited pipeline and compute environment for self-execution and extension

Data analysis pipeline

Data processing infrastructure

Augmented data for clinical use

Pathlabs

Scientists

CSIRO Compute Infrastructure

Clinician

Denis Bauer
NGSANE and Cloud-based genomic services
Goal: Provide cloud-based analytics for genomic data

Services

- Loci Diagnostics (interrogate single disease loci)
- Clinical Diagnostic (interrogate known Disease genes)
- De-novo associations (prioritize genome-wide variants)
- Individual Health Report (Food/Lifestyle recommendations based on risk loci Reported In literature)
- Genealogy/Ancestry

Customer

- Clinician
- Researcher (non expert)
- Researcher (genomics expert)
- Individual (e.g. RSL care customer)

Raw genomic information
Sequencing facility
Sample

(Virtual machine, Docker)

NGSANE Production informatics

Smart analytics

(Infrastructure, Security, Visualization)

http://bioinformatics.oxfordjournals.org/content/early/2014/01/25/bioinformatics.btu036
https://github.com/BauerLab/ngsane
Richness of MRI

Anatomy

Angiogram Connections

lesions Spectroscopy Functional
The Australian Imaging Biomarkers and Lifestyle study
http://aibl.csiro.au/

Multi institutions consortium
Industry Involvement

Clinical biomarkers

- Cognitive functions
- Blood biomarkers
- Genomic
- Demographic And Lifestyle

2006
baseline 18 months 36 months 54 months 72 months
1000+ subjects

2014

MRI, PET, Ultrasound, and retinal Imaging

- T1W Anatomy
- T2W CSF and structures
- SWI Venous tree
- FLAIR White matter
- DWI Diffusion imaging
- PET Amyloid $^{11}$C and $^{18}$F markers
- Retina

- Biomarkers from several streams: blood, cognition, lifestyle, imaging, genomics...
- Biostatistics and genomics done from central database
- Image analysis performed automatically
- Web based reporting and querying

Multi institutions consortium
Industry Involvement
Cloud based Image analysis and reporting

**MilxView/SMILI**
Design standalone application for processing images locally
*Open source*

**MilxCloud**
Run standard processing pipelines remotely
*Open access*

**MilxXplore**
Create reports that can be accessed through internet
*Open source*
SMILI (Simple Medical Imaging Library Interface)

- light weight wrapper around ITK and VTK
- model/surface processing
- visualisation
- user-interface independent library, so that it easily utilised in your own command line applications.
MIXCloud, image analysis pipeline from a web portal
Individual report: MilxXplore

http://milxview.csiro.au/MilxXplore/?q=Tour

Bourgeat et al. JAMIA 2013
Cloud based clinical reporting

3D scans

Remote processing

Web interface
Change in Aβ burden over time
Florbetaben

Baseline MCI

12 months MCI

24 months AD

Images generated through CapAIBL® (capaibl-milxcloud.csiro.au)
CSIRO Biomedical Imaging Group
Frontotemporal Lobar Degeneration

**Behavioural FTD**

![Brain images](image1)

**Progressive non-fluent aphasia**

![Brain images](image2)

**Semantic Dementia**

![Brain images](image3)

**Logopenic aphasia**

![Brain images](image4)
AD neuroimaging signature

- Neurodegeneration
  - Grey matter volume

- Neuronal injury
  - FDG

- Pathology
  - Aβ imaging

Images generated through CapAIBL® (capaibl-mixcloud.csiro.au)
CSIRO Biomedical Imaging Group

Adapted from La Jolie et al, Neurobiol Dis, 2012
Acknowledgments

CSIRO
Denis Bauer (Team leader)  Steve Rose (Team leader)
Pierrick Bourgeat  Olivier Salvado (Group leader)
Samantha Burnham  Neil Saunders
David Conlan  Kaikai Shen
James Doecke  Christian Stolte
Vincent Dore  William Wilson
Jason Dowling
Nicholas Dowson
Firoz Anwar
Jurgen Fripp (Team leader)
Soumya Ghose
Julian Heinrich
Jhimli Mitra
Sean O’Donoghue (Team leader)
Aidan O'brien
David Rivest-Henault

Contact:
olivier.salvado@csiro.au
The Australian eHealth Research Centre
aehrc.com