How can HPC and Bioinformatics help to cure infections and cancer?

Associate Professor  
Fabio Luciani  
Faculty of Medicine  
School of Medical Sciences, Kirby Institute
Right now, one of your cell has

- 1 Genome: 3.2 Tera bases of DNA
- 2000-10,000 genes expressed at any time
- ~100,000 proteins

Right now, there are 10-30 trillions cells per human body
How do we study this immense set of molecules?

• Next generation sequencing has revolutionised biomedical research
• We can now sequence 1-10 genomes in a day in a single machine
Cost per Raw Megabase of DNA Sequence

- Second Generation Sequencing
- Moore's Law

NIH National Human Genome Research Institute

genome.gov/sequencingcosts
Oh, my God!
What should I do now?

NGS Machines
Massive amount of sequence data
Next generation sequencing for vaccine research

Population studies
- Vaccine safety
- Host genotyping (HLA, etc.)

Host-pathogen interactions
- Immune response
  - T and B cell repertoire
  - Immune escape
  - HLA typing
  - Epitope discovery
- HLA-peptide complex
- Host transcriptome
- Micro-RNA, si-RNA
- Pathogen particles
- Pathogen genome
- Pathogen replication

Molecular epidemiology
- Metagenomics
- Detection of new pathogens

Transcriptomics

Genomics

Epigenetics

Luciani et al. Trends in Biotechnology 2012
Immunology and immunotherapies
Nobel Prize in Medicine 2018

• Jim Allison and Tasuku Honjo for their discovery of molecules that “put a break” on the immune response (T cells).
• Their work led to the development of new drugs that can unleash immune responses against cancer.
• Science Discovery to Medicine application
Application of genomics in cancer research and immunotherapies

Patient tumor biopsy → Genomic and transcriptomic profiling → Rx: Cancer immunotherapy
- Immune checkpoint blockade
- Adoptive cell transfer
- Therapeutic vaccine
- Oncolytic virotherapy

Clinical decision-making
Application of genomics in cancer research and immunotherapies
Machine learning and bioinformatics

• Integrating different “omics” data set from same patient, sample, or cell.
• Visualisation of these large data sets and of the interactions between different molecules
• Prediction of disease outcome, novel mechanisms
• New therapies
Bioinformatics tools for genomics

Genomics is a set of very large data sets

- NGS are error prone, need QC
- Need alignment of NGS data
- Normalisation and quantification of expression levels
- Identification of rare variants
- Differential expression between samples
- Statistical associations with disease features
What is machine learning?

A class of computational algorithm which iteratively learn an approximation to some function

- Representation
- Evaluation (loss function)
- Optimisation

Supervised learning
- Classification
- Regression

Unsupervised learning
- Clustering
Data Science, Statistics
Machine learning is not glorified statistics
“Cool” is not necessarily correct!
Causation and correlation

1992 - MAD COW DISEASE OUTBREAK AREAS

IT MAY, HOWEVER, BE A MISTAKE TO JUMP TO CONCLUSIONS
Artificial Neural Network -> Deep Learning

- An artificial neural network, initially inspired by neural networks in the brain (McCulloch & Pitts, 1943; Farley & Clark, 1954; Rosenblatt, 1958), consists of layers of interconnected compute units (neurons).
- The depth of a neural network corresponds to the number of hidden layers, and the width to the maximum number of neurons in one of its layers.

![Diagram of Artificial Neural Network](image)

**A** Input layer → Hidden layer → Output layer

**B**

- **Inputs**
- **Weighted sum**
- **Activation function**
- **Output**

**C**

- **FORWARD PROPAGATION**
  - **PREDICTED label**
  - **TRUE label**
- **BACKWARD PROPAGATION**
  - **LOSS**
  - **w' = w + ηΔw**
  - **Local optimum**
  - **Global optimum**
Application of Deep learning to genomics
Application of single cell technology, unsupervised statistical analysis and immunotherapies
What do we do with HPC

• NGS data from single cells
  – RNAseq
  – Immune receptor analysis

• Proteomics
  – Flow cytometry

• Pathogen genomics
  – Viral diversity
Issues with HPC

• Bioinformatics pipeline are not always parallelizable.

• Several steps requires minimum CPU power and maximum memory allocation
  – Normalisation of multiple transcriptomes across a large number of single cells’ data

• Lack of statistical model and algorithms for multi-omics integration (DNA+RNA+Proteins)
1-HITS-p cohort
Prospective cohort of high-risk IDU

2- HCV RNA extraction and full genome amplification

3- Deep Sequencing & epitope prediction

4- Matrix IFN-Y ELISPOT

5- Phenotyping of HCV-specific CD8+ T cells

6- Linking markers to clonotype and transcriptomics
Single cell instead of “Bulk” analysis: why do we need it?
Towards single cell systems immunology

From microscope to genomic lens

OLD VERSUS NEW
Mixture of immune cells.

BULK RNA SEQUENCING
Sequencing a mixture of seemingly identical cells fails to capture the diversity of the immune cells surrounding a tumour.

SCALE UP
In the past decade, biologists have moved from analysing a few genes in a handful of cells, one cell at a time, to surveying thousands of genes in hundreds of thousands of cells, in parallel.

SINGLE-CELL GENOMICS
Using single-cell genomics, biologists can capture the molecular signature of all immune cells found in and around the tumour.

*Data points included represent high-profile studies that introduced technological advances in single-cell RNA sequencing.
Application of scRNAseq to study immune cells (T cells)

• We study T cell responses against viral infection or cancer
• We isolate “one cell at the time” T cells circulating in the blood of patients that are specific for a disease or virus
• We study their protein expression and their gene expression (transcriptome)
• We use statistical and bioinformatic analysis to link this information and understand disease
Linking surface phenotype with scRNAseq using index sorting

VDJPuzzle reconstruct both T- and B-cell receptors from full-length scRNA-seq

[Diagram showing the process of VDJ transcript alignment, homologous recombination, and IgBlast for error correction and repertoire generation]

https://bitbucket.org/kirbyvisp/vdjpuzzle2

Eltahla et al. Immunology and Cell Biology 2016
Rizzetto et al. Bioinformatics 2018
A subset of Tscm are consistent with

**Cluster 2**: OX-PHOS, Mitochondrial electron transport, T cell differentiation enhanced respiratory capacity (NADH), fatty acid oxidation (cytochrome c oxidase genes (e.g. COX3), and mitochondrial biogenesis (ATP, CYTB).

**Cluster 1,4**: Cell adhesion Ribosomal biogenesis Regulation type 1 IFN NFKB signalling
Kirby Institute, UNSW
Simone Rizzetto
David Koppstein
Auda Eltahla
Jerome Samir
Curtis Cai
Mehdi Pirozyan
Elizabeth Keoshkerian
Prof. Andrew Lloyd

Westmead Hospital
Dr Ken Micklethwaite, Dr Emily Blyth, Leighton Clancy, David Gottlieb

University of Western Australia
Prof. Silvana Gaudieri

Garvan Institute
Joanne Reed
Mandeep Singh
Katherine Jackson
Prof. Chris Goodnow

NSW Health

Australian Centre for HIV and Hepatitis Virology Research

Australian Government
National Health and Medical Research Council