QIMR Berghofer
Medical Research Institute

Cancer and Computers

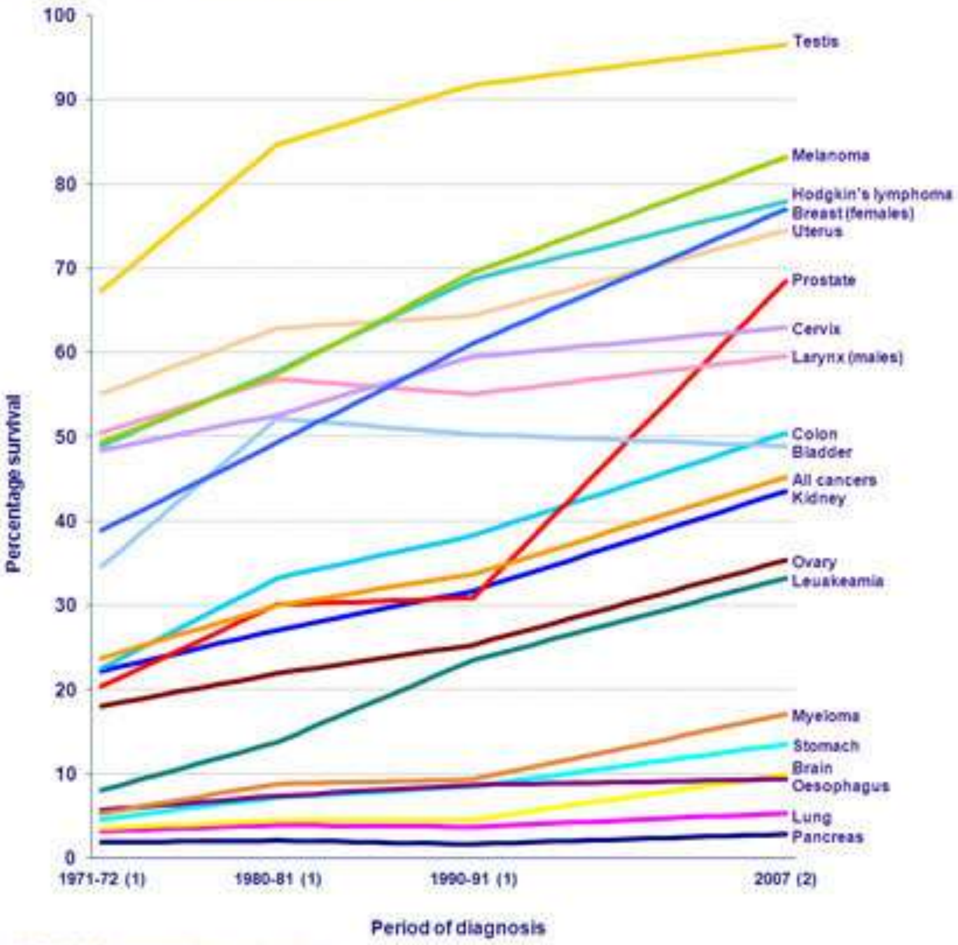
John Pearson

QUESTnet 2016, RACV Royal Pines Resort, Gold Coast

6 July 2016

Cancer - cancer types we work on:

Figure 1.2: Relative survival (%), adults (15-99 years), selected cancers, England and Wales: survival trends for selected cancers 1971-2007



(1) 1971-1991 Cohort analysis - actual survival
(2) 2007 Hybrid analysis - predicted survival

Pancreatic

Brain metastases

Oesophageal

Mesothelioma

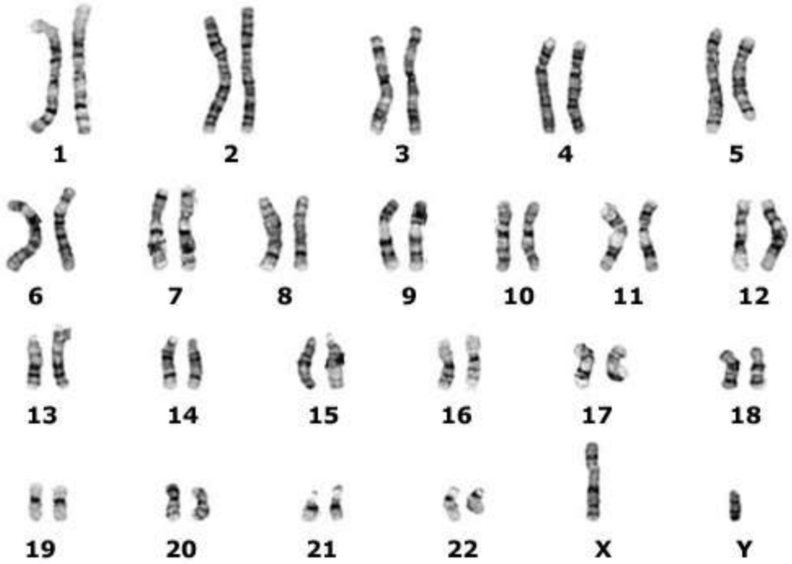
Ovarian

Melanoma

Breast

The Human Genome

- Every cell in the human body starts with a copy of the human genome
- The human genome is a set of 23 chromosomes (2 copies of each)
- Each chromosome is a long molecule of a type called DNA
- A DNA molecule is a string of chemicals called nucleic acids
- There are 4 nucleic acids – A, C, G, T
- The 23 human chromosomes contain 3 billion nucleic acids



Next-generation Sequencing (NGS)

- Smash DNA genome into small pieces

- Sequence the fragments and create lists of base-sequences as strings

- Use HPC to align strings to genome to recreate genome

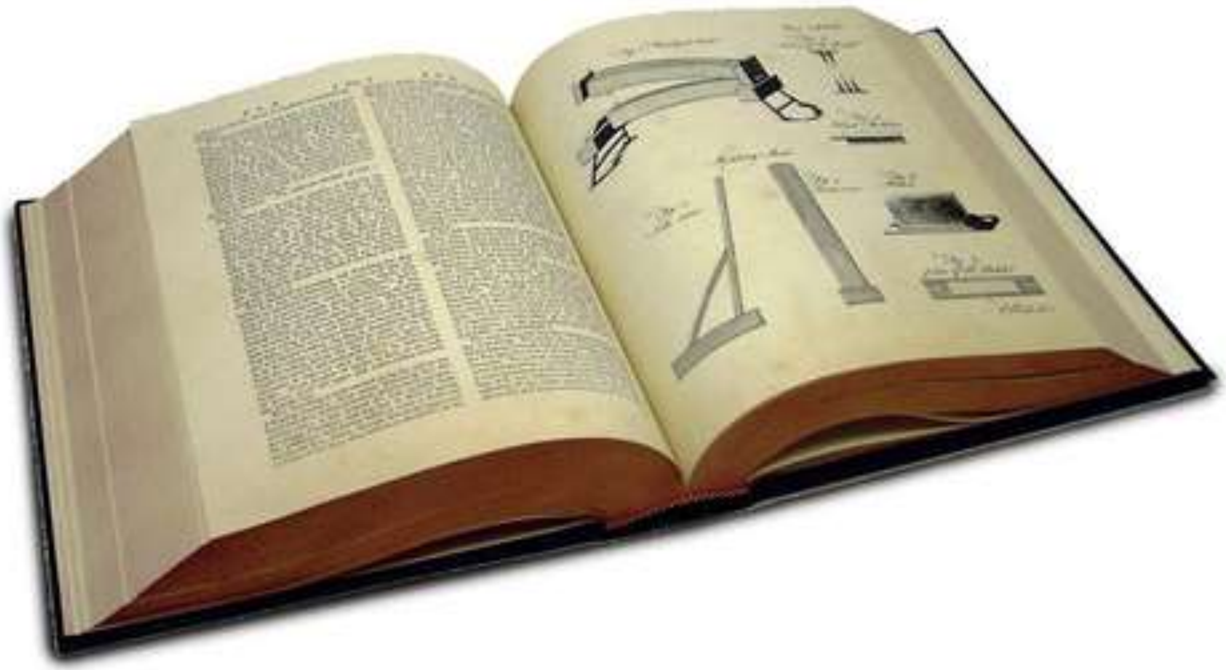
Genomics Data Sets are BIG

A volume of the Encyclopedia Britannica:
500 double-sided pages, 8 million characters

x 1

8

megabytes



QIMR Berghofer
Medical Research Institute

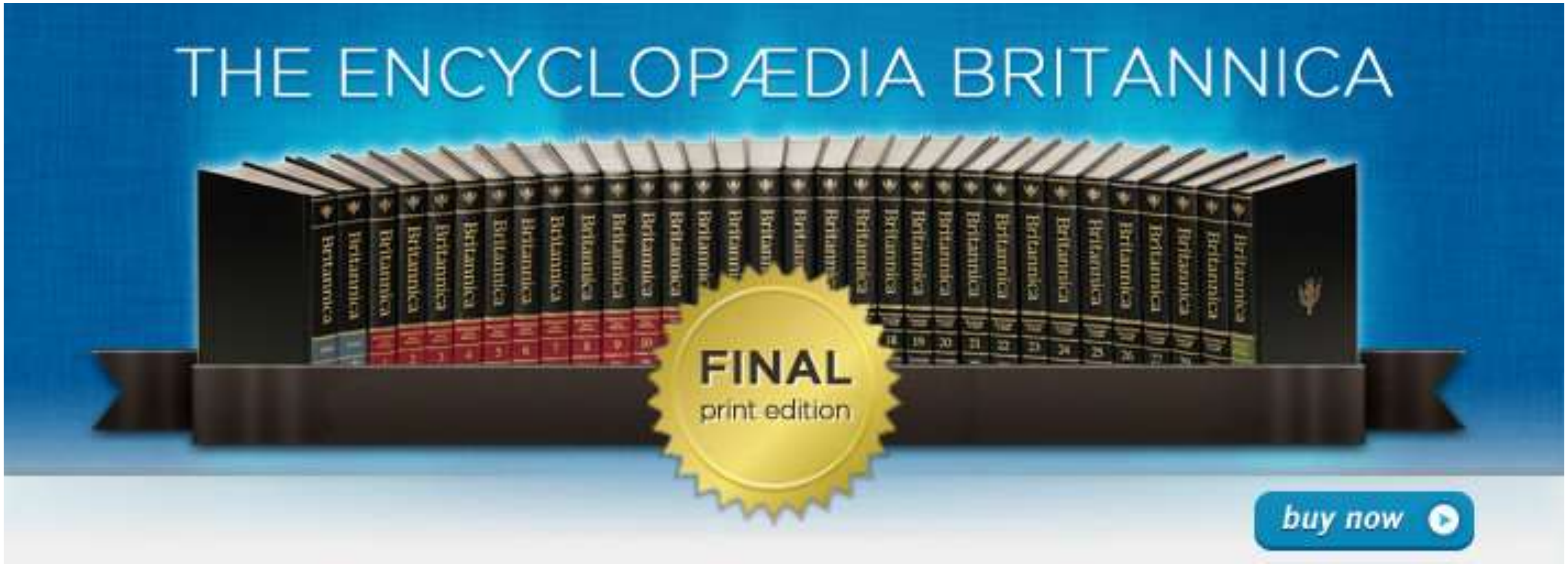
Genomics Data Sets are BIG

A set of the Encyclopedia Britannica:
40,000 articles, 44 million words

x 31

250

megabytes



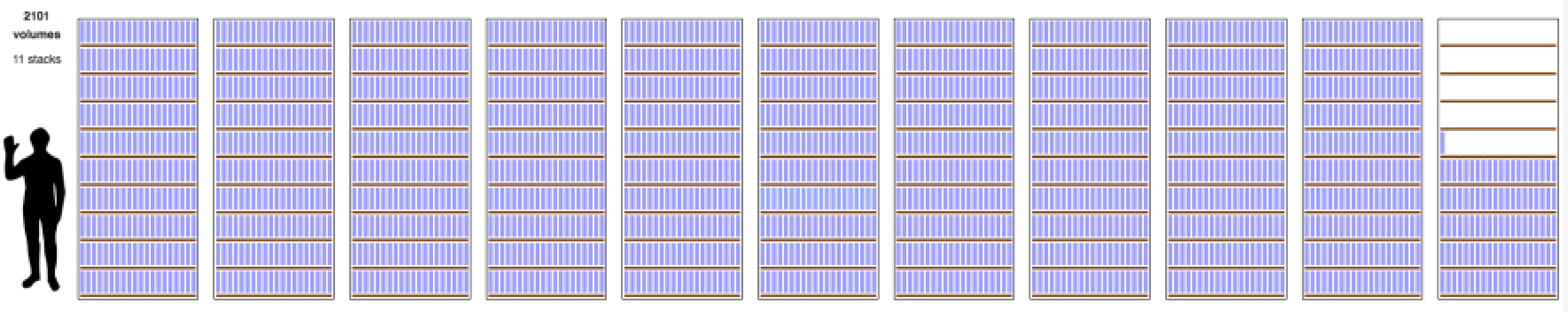
Genomics Data Sets are BIG

A printed copy of the English version of Wikipedia: 4.7 million articles, 2.7 billion words.

x 2,100

17

gigabytes



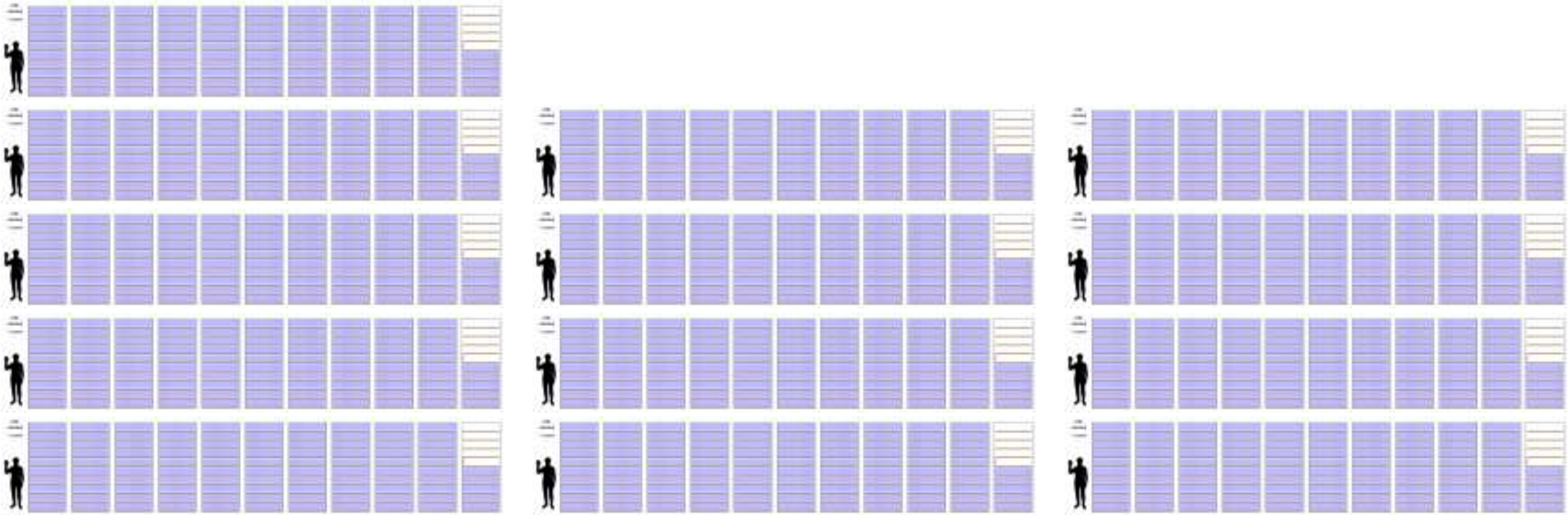
Genomics Data Sets are BIG

A single patient's cancer genome data: tumour sample @60x, normal sample @30x

x 37,500

300

gigabytes



QIMR Berghofer
Medical Research Institute

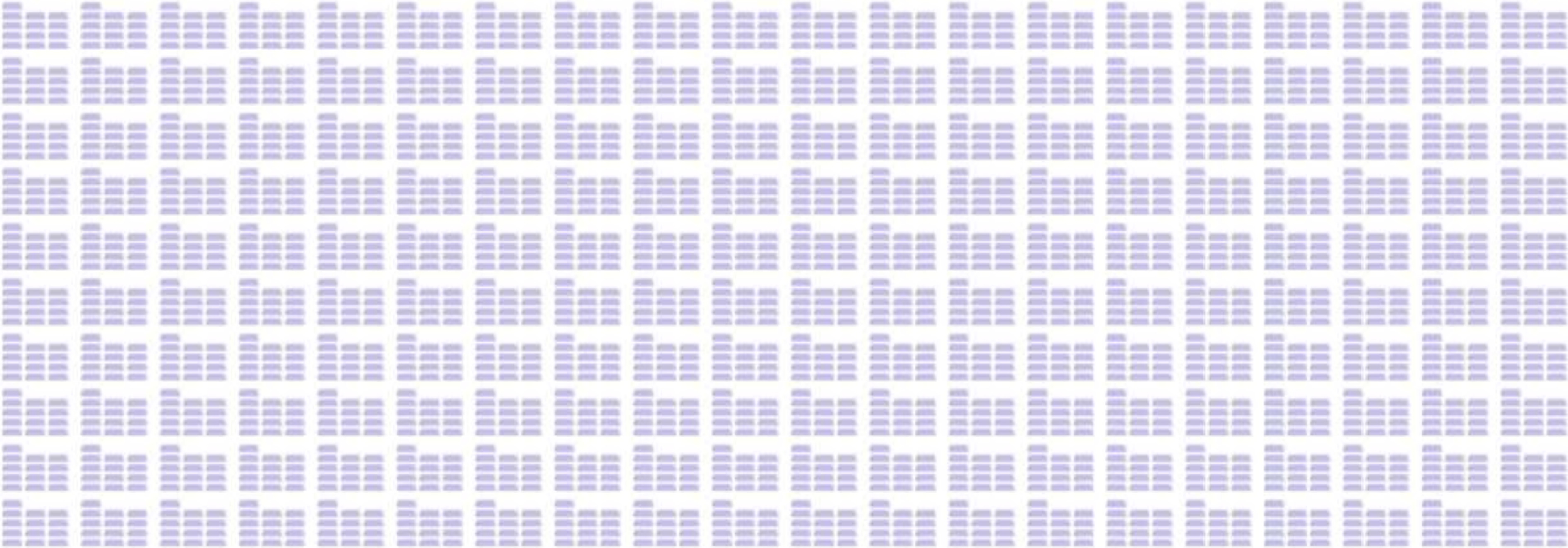
Genomics Data Sets are BIG

A study of 200 cancer patients using whole genome sequencing : tumour sample @60x, normal sample @30x

x 750,000

60

terabytes



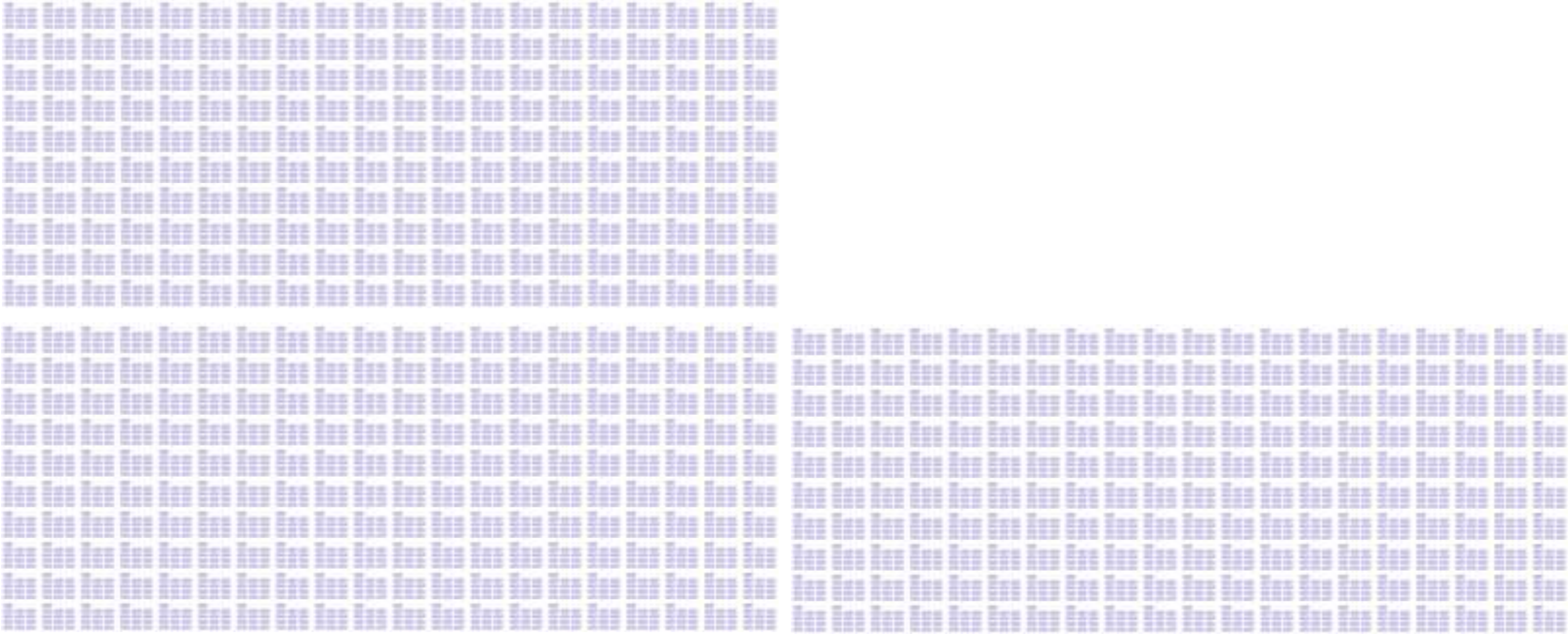
Genomics Data Sets are BIG

During analysis we need triple the space so we have room for intermediate and temporary files

x 2,250,000

180

terabytes



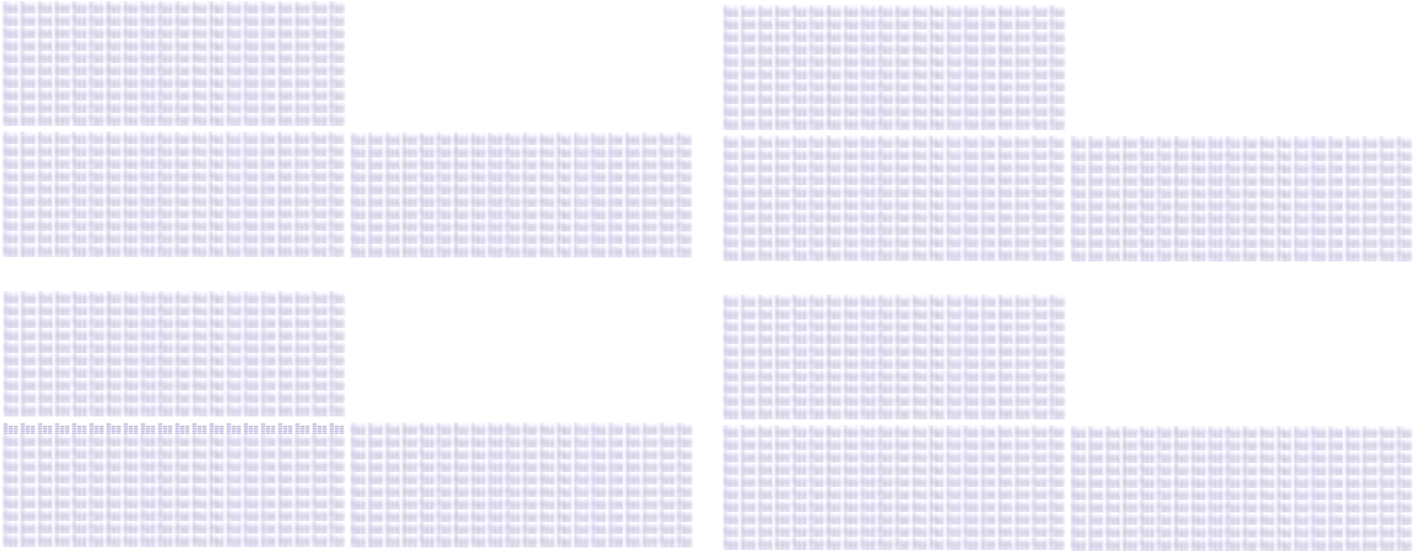
Genomics Data Sets are BIG

Ideally we'd do 1000 patients per cancer instead of 200 and what happens if we decide to do more cancer types …

x ???????

?

petabytes



Cancer is a story of good cells gone bad

Skin (good):

- Keratin armour – waterproof, airtight, puncture resistant, self healing
- Continually being renewed – complete replacement every fortnight
- DNA copy fidelity – “10 9’s quality”
- Repair injuries – cells copy themselves but just enough

Cancer (bad):

- Moves, makes blood vessels, stops listening to cellular signals, copies
- DNA repair is broken so DNA copy mistakes get made
- Mistakes accumulate
- To spot a mistake you need to be able to compare the “broken” cancer genome AND the “normal” genome

Detecting broken genes in a cancer sample

Human Reference Genome (Pete)

A C G T A G T C T C A A T T T A A T G C A C T A G A A C G G

Bob’s Cancer Genome

A G G T A G C C T C A T A T T A A T G C A C T A A A A G G G

Bob’s Normal Genome

A G G T A G C C T C A A A T T A A T G C A C T A A A A G G G

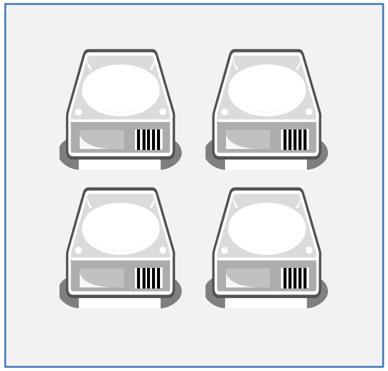
Hardware:

Compute

40 nodes

- 2 x 12 core Intel
- 256 GB RAM
- 4 x 1TB disks striped
- 1 x 10 GigE

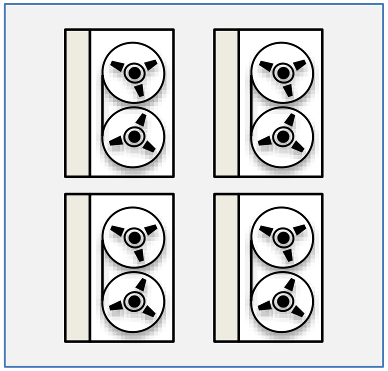
Disk



1.4 PB Lustre

- 4 x OSS pairs
- 8 x 10 GigE per pair

Tape

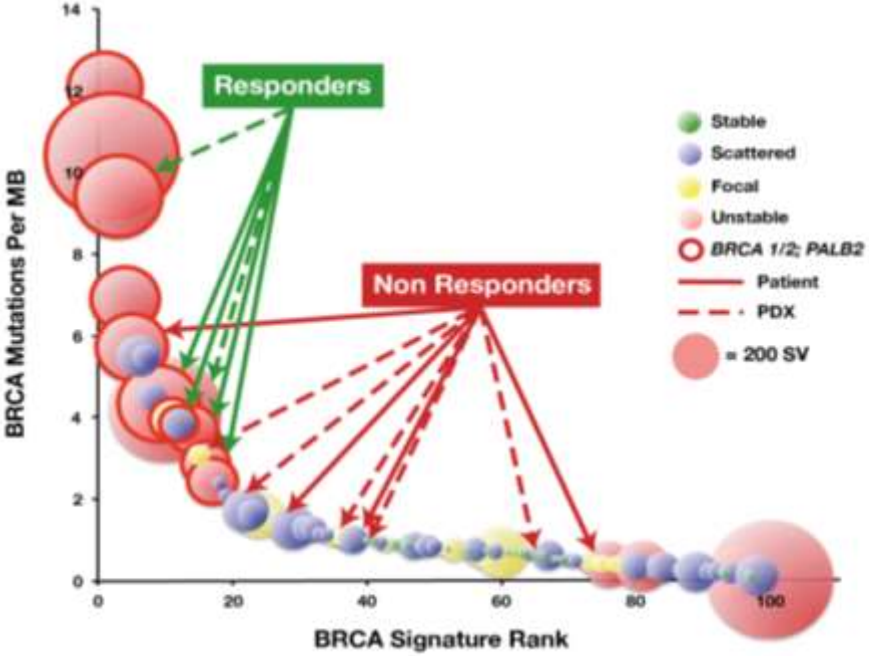


2 redundant sites

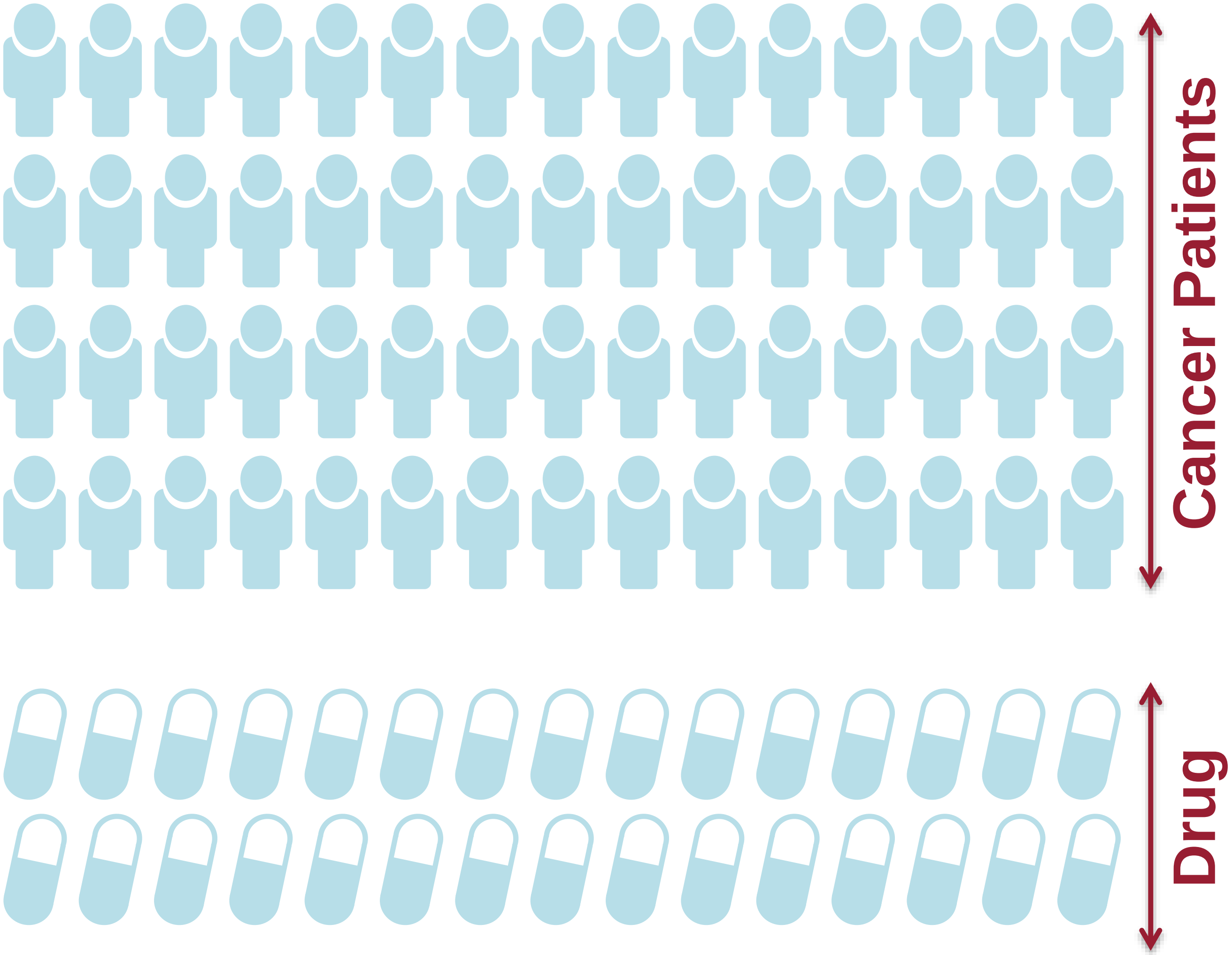
- 30 kms apart
- 2.5 PB per site
- 10 GigE connection
- Nightly updates

What can we do with cancer genome sequencing:

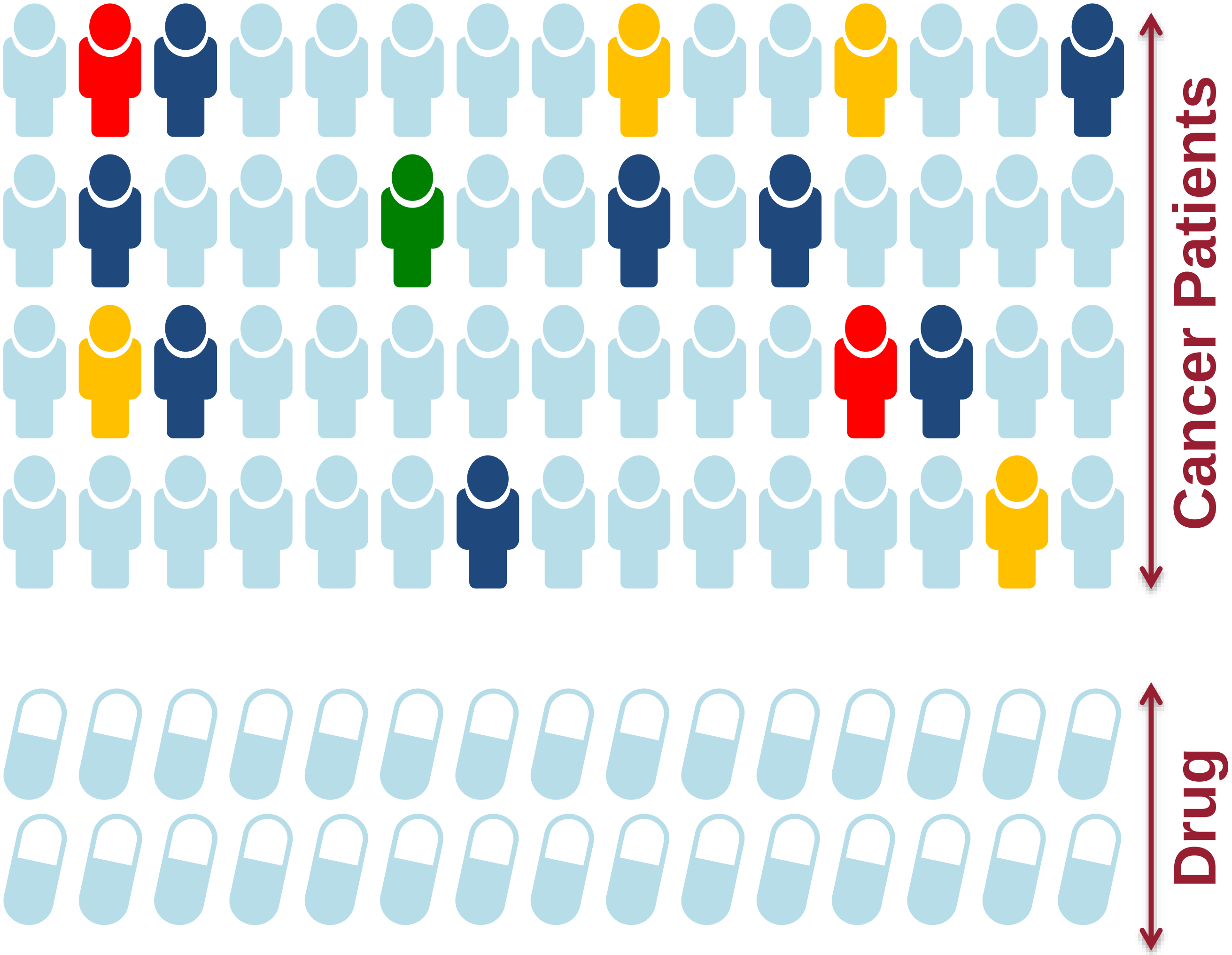
- Increase our understanding of what cancer is
- Better classification of patients into cancer subtypes
- Identification of potential new targets for drugs
- Can improve diagnosis for patients



The current way we treat cancer:

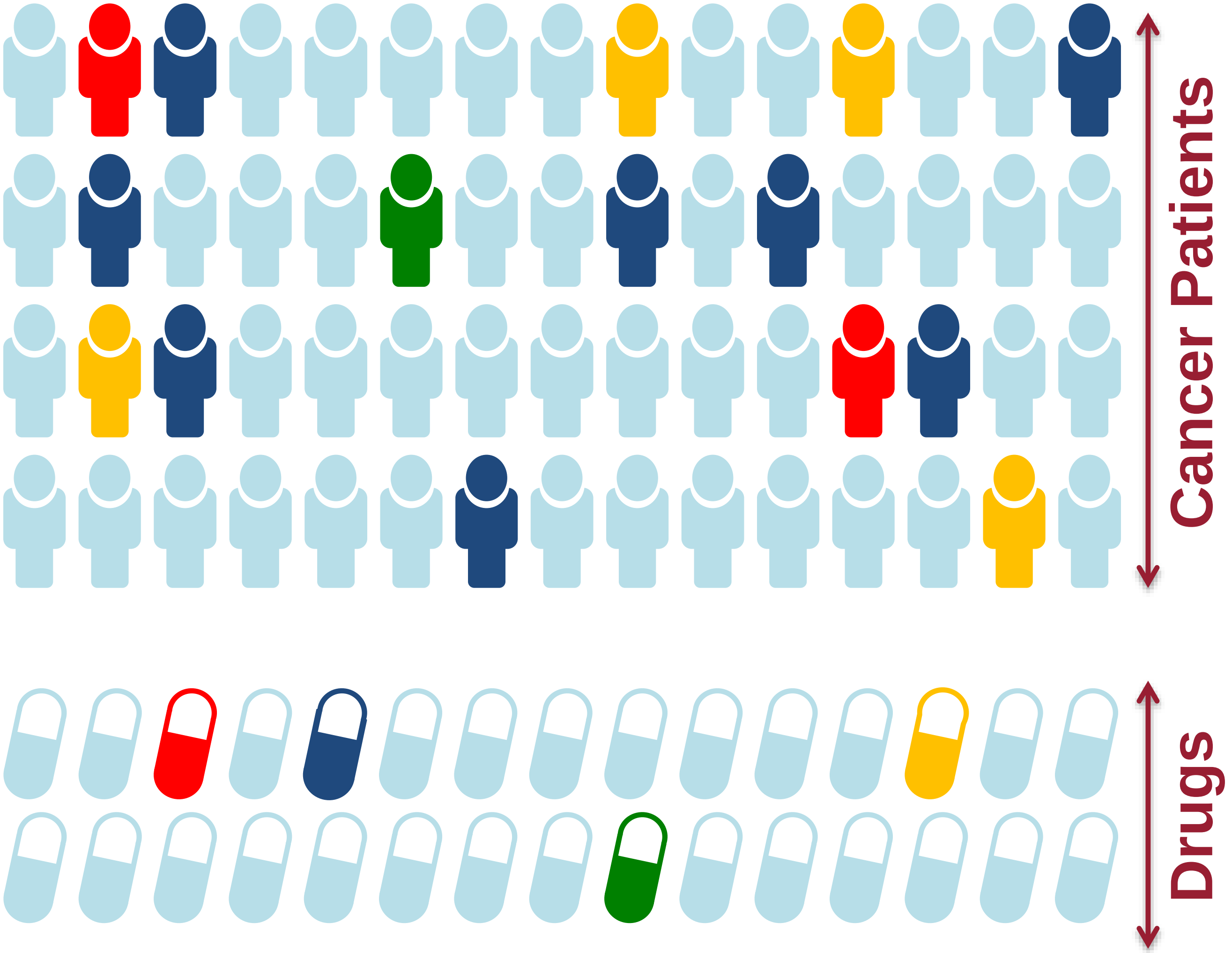


But we know cancer patients are different:



QIMR Berghofer
Medical Research Institute

We need to match patients to drugs – Precision Medicine:



QIMR Berghofer
Medical Research Institute

Cancer & Computers - Summary

- Cancer research is hard and expensive
- New technologies are reshaping our knowledge of cancer
- Progress is being made in diagnosis and treatment
- Modern cancer research needs:
  - labcoats *and* laptops
  - bunsen burners *and* broadband
  - freezers *and* fiber-optics

Acknowledgements:

Genome Informatics:
John Pearson
Conrad Leonard
Oliver Holmes
Christina Xu
Scott Wood
Xiaping Lin

Medical Genomics:
Nic Waddell
Ann-Marie Patch
Katia Nones
Stephen Kazakoff
Martha Zakrzewski

The University of Queensland Australia
Cancer Council NSW
Queensland Government
Australian Government National Health and Medical Research Council
Australian Government Australian Research Council

QIMR Berghofer Medical Research Institute