LCA 2016 Open Source and Bioinformatics

Clinical Genomics

A Computational Perspective

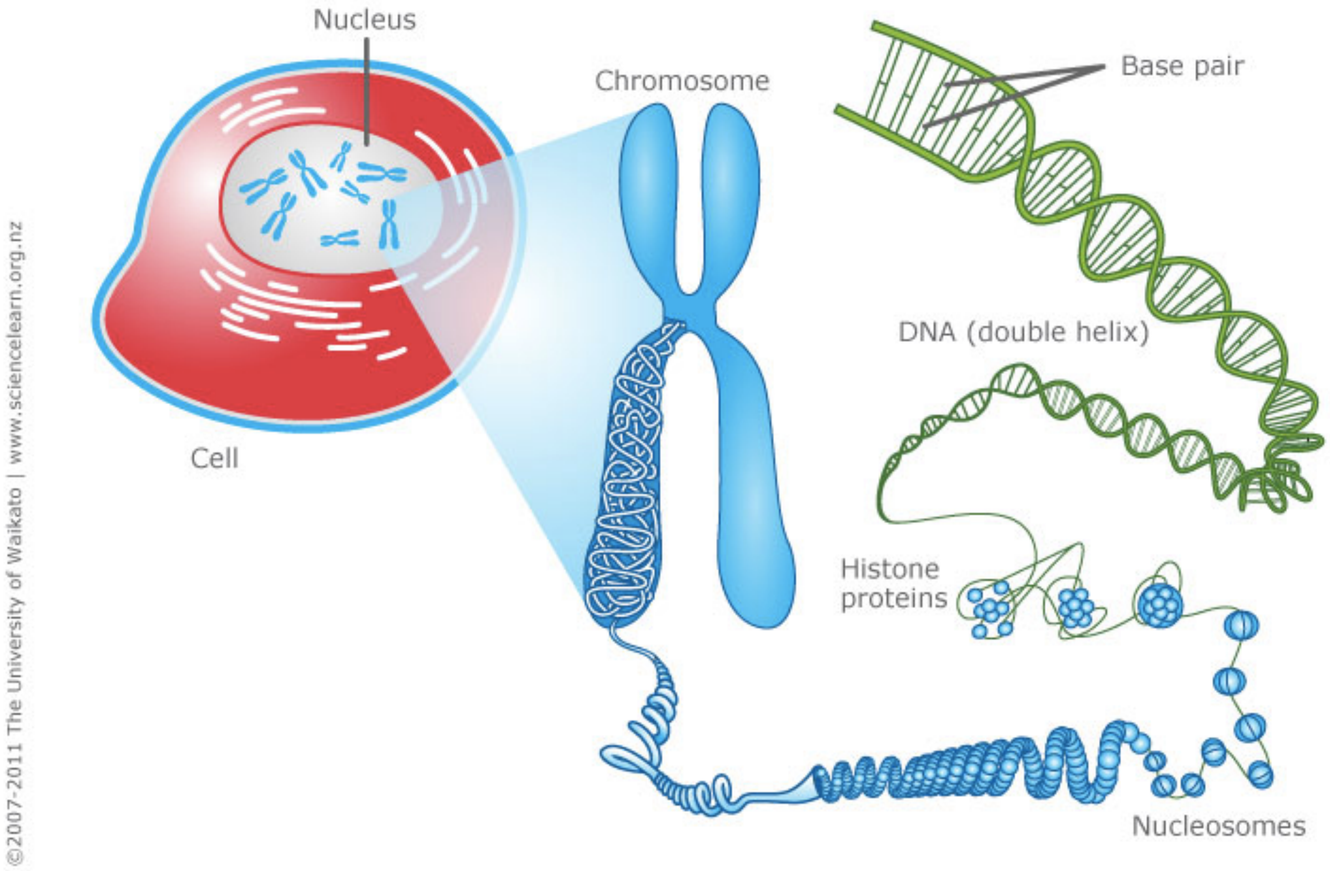
Bernie Pope, bjpope@unimelb.edu.au
Lead Bioinformatician
Cancer Genomics
Clinical Genomics

Definition

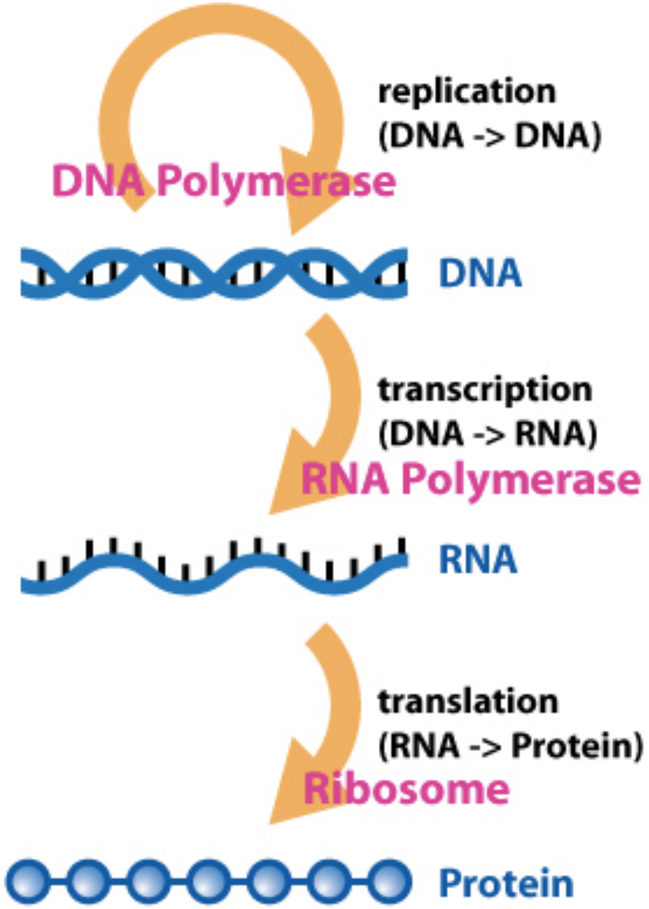
- **Genomics**: the study of the structure and function of DNA in a *holistic* manner.

- **Clinical Genomics**: the diagnosis and treatment of *medical* conditions with reference to a patient's DNA.

Basic Cell Biology



Basic Cell Biology



https://commons.wikimedia.org/wiki/File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg

Genotype Versus Phenotype

- **Genotype**: the DNA contained in a cell, interpreted as a biological code.
- **Phenotype**: the observable traits of an organism, such as shape, colour, size, behaviour *etcetera*.

genotype + environment ⇒ phenotype

High Throughput DNA Sequencing

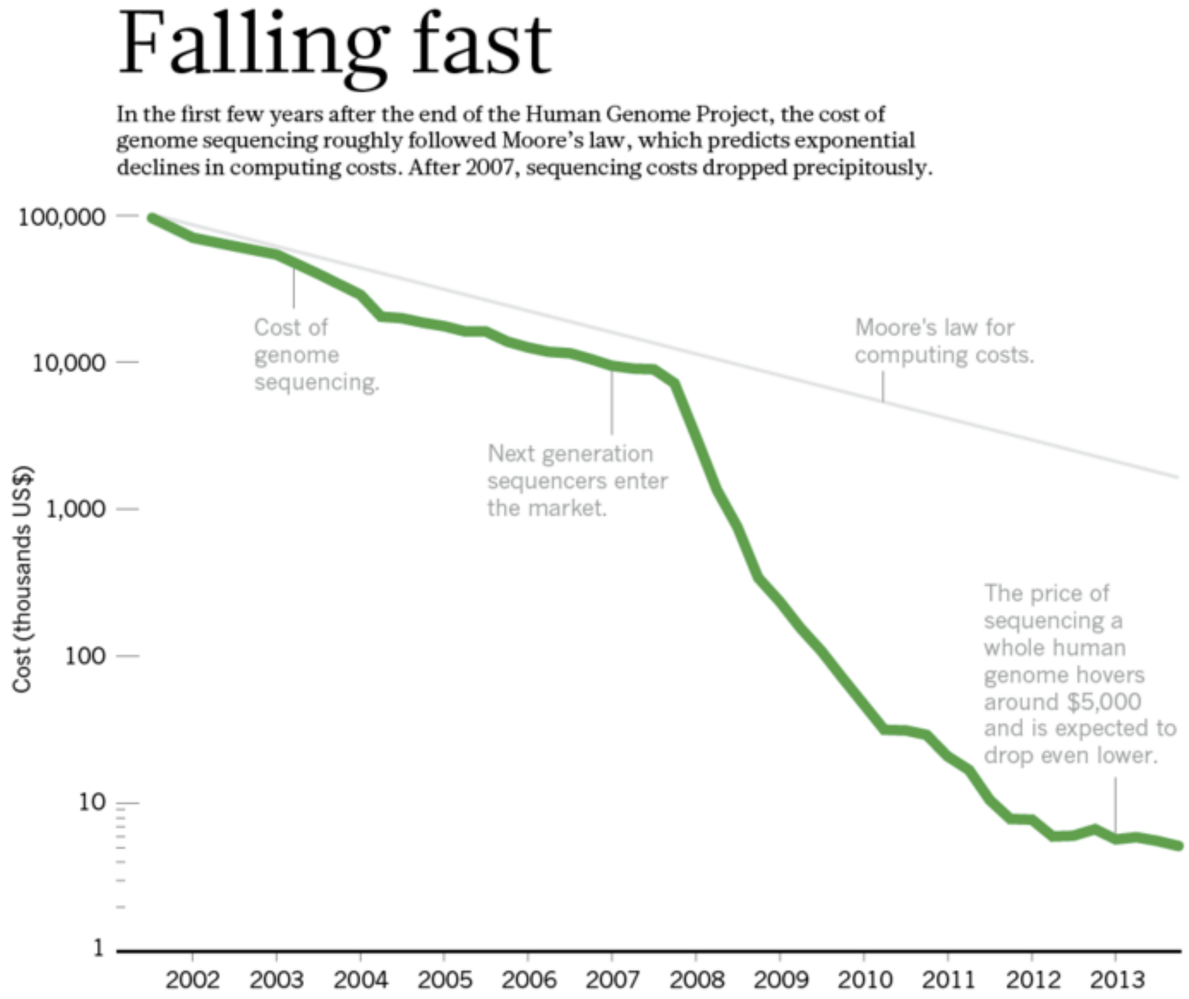
DNA Sample
(biological)

DNA Sequences
(digital)

Example: 1 human genome
≃ 70 GB

Image courtesy of Illumina (https://www.illumina.com/)

High Throughput DNA Sequencing



E. Hayden, The $1,000 genome, Nature, 19 March 2014

A New Tool for Treatment and Diagnosis

- Measuring genotypes is *cheap* and *accurate*.

- Accumulated (anonymised) data over a population provides insight into the connection between genotype and phenotype.

Melbourne Genomics Health Alliance

The Royal Melbourne Hospital

The Royal Children's Hospital Melbourne

The University of Melbourne

Walter+Eliza Hall Institute of Medical Research

Murdoch Childrens Research Institute

CSIRO

agrf

Peter Mac Peter MacCallum Cancer Centre Victoria Australia

Austin Health

MonashHealth

Major funder

Victoria State Government

http://www.melbournegenomics.org.au/

Example Benefits

Childhood syndromes, 80 patients

	Genomics	Standard Care
Diagnosis rate	47/80	10/80
Cost per diagnosis	$6,003	$27,040

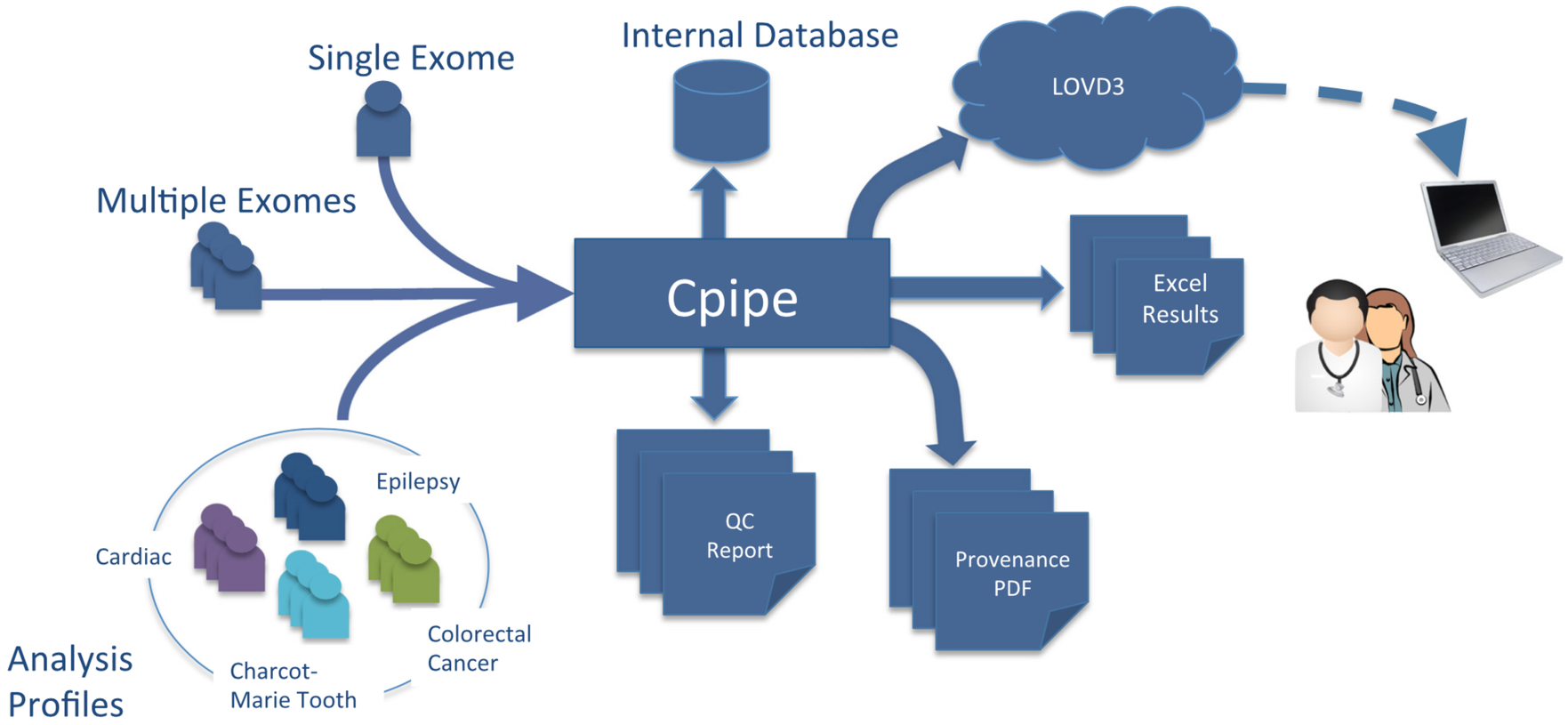
Technical Challenges

- Current DNA sequencing technology produces millions of short fragments of DNA.
- Any two human individuals differ in about 3 million DNA bases.
- Most DNA variation is benign.

Technical Challenges

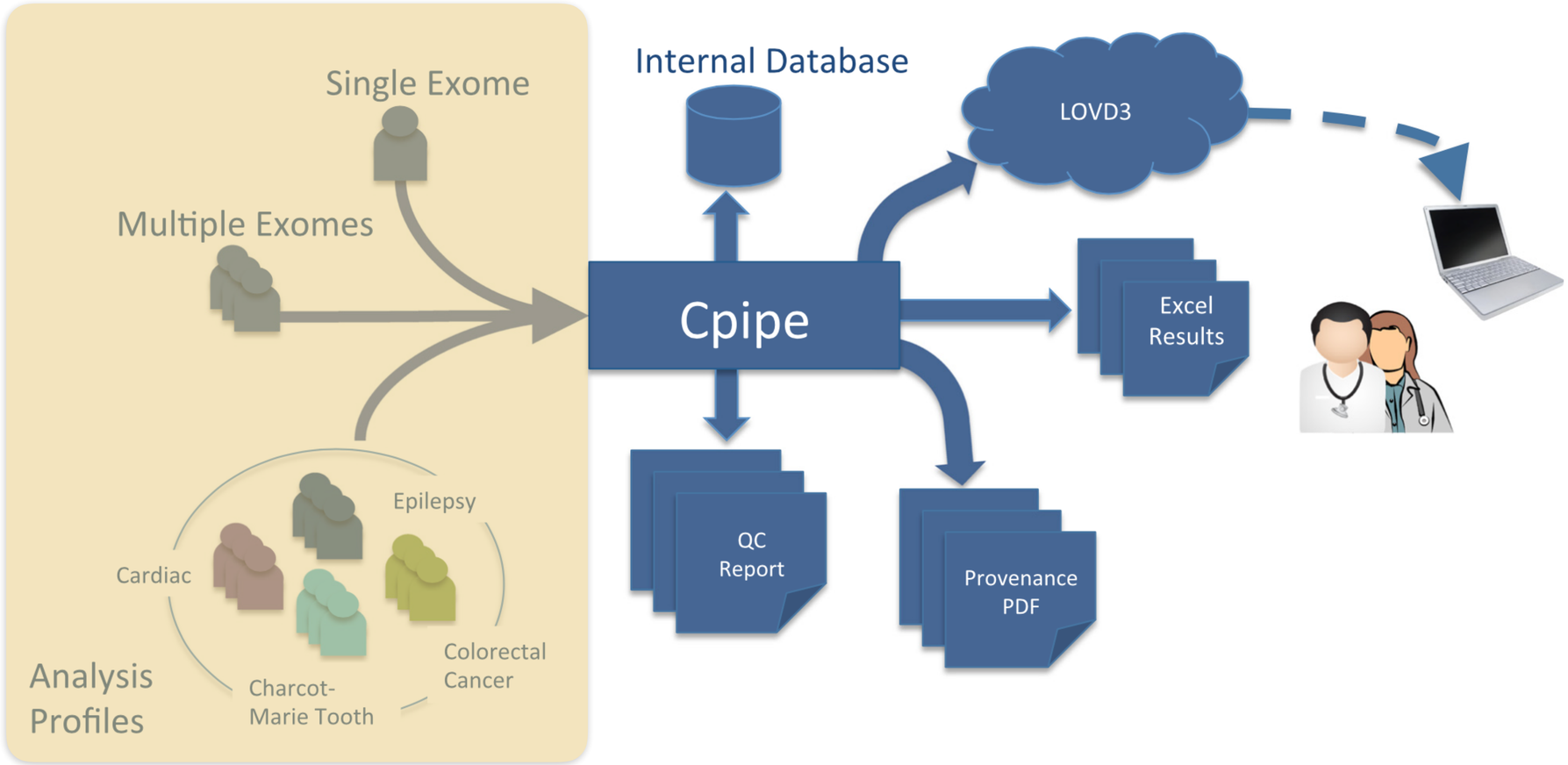
- The DNA fragments must be pieced back together.
- The large number of variants must be filtered for significance.
- The entire process, from sample collection to diagnosis must be robust. **Quality control is crucial!**

Bird's-Eye View of Analysis



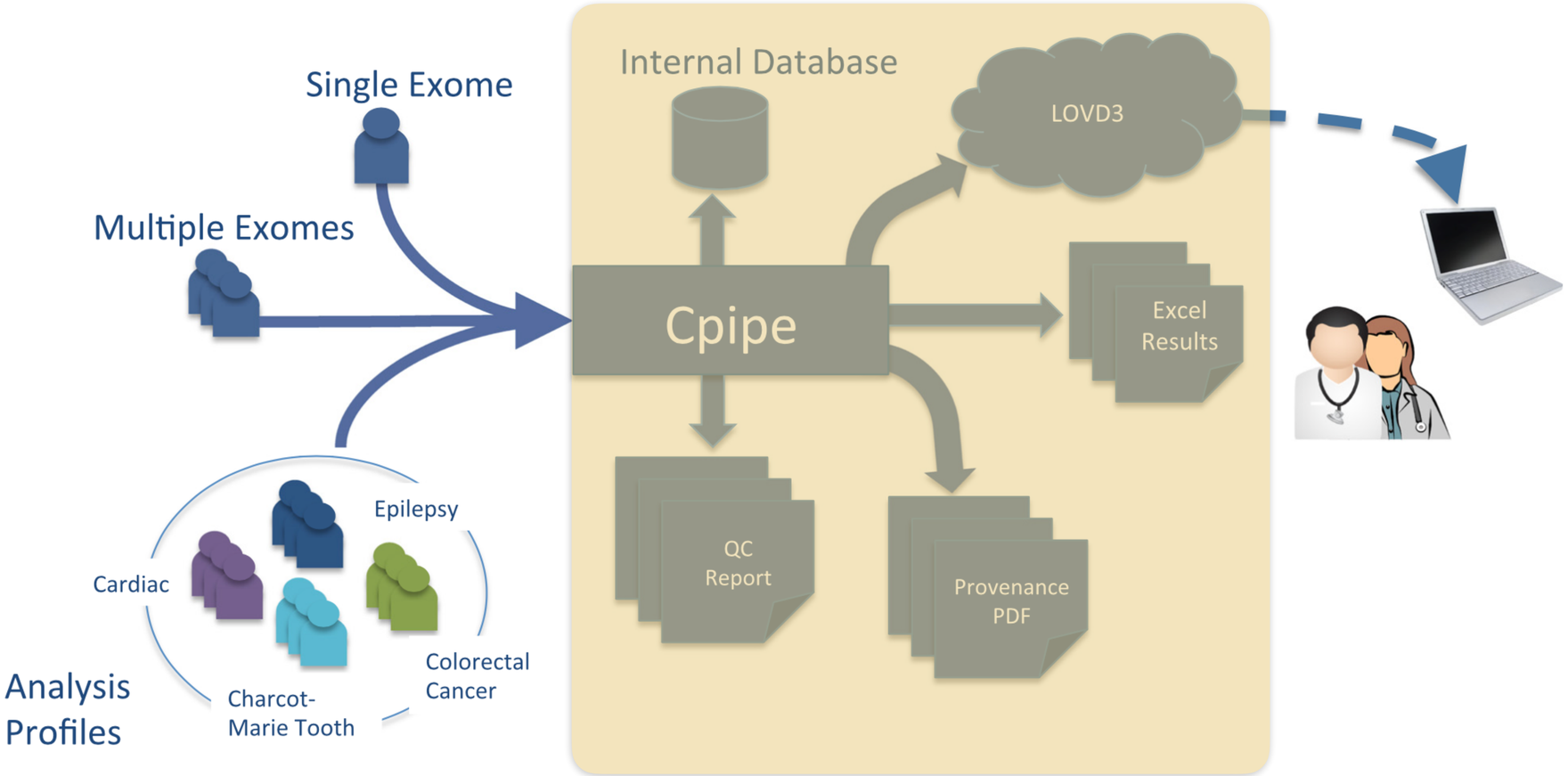
Cpipe: a shared variant detection pipeline designed for diagnostic settings, Genome Medicine, 2015

Bird's-Eye View of Analysis



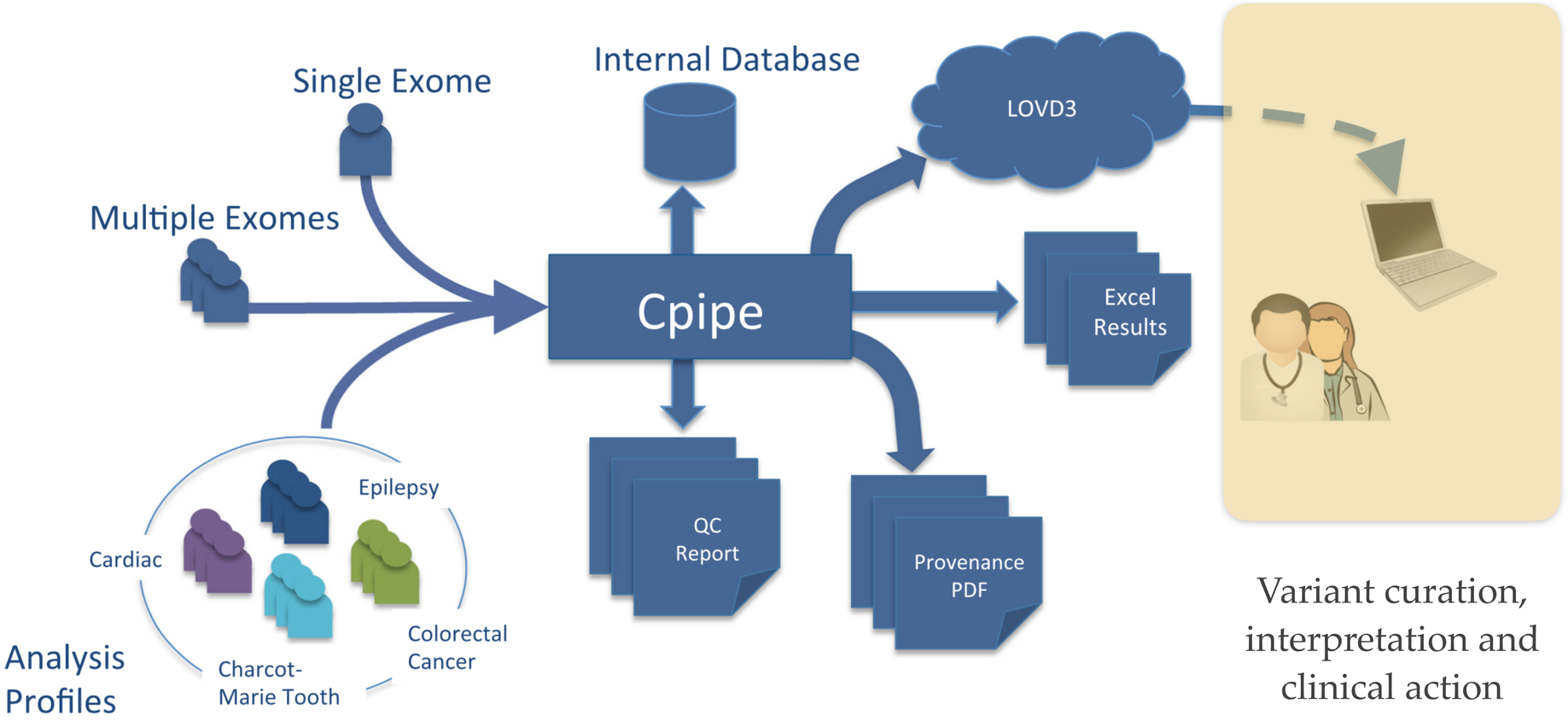
Sample collection and DNA sequencing

Bird's-Eye View of Analysis



DNA variant detection and collection

Bird's-Eye View of Analysis



Variant curation, interpretation and clinical action

Major Software Components

- **LOVD**: a database and web application for the collection and curation of DNA variants.

- **Cpipe**: a dataflow system for orchestrating the analysis.

- Leiden Open Variation Database
- Leiden University Medical Center
- PHP, MySQL
- GPL v3
- https://github.com/LOVDnl/LOVD3/
- We have extended it significantly

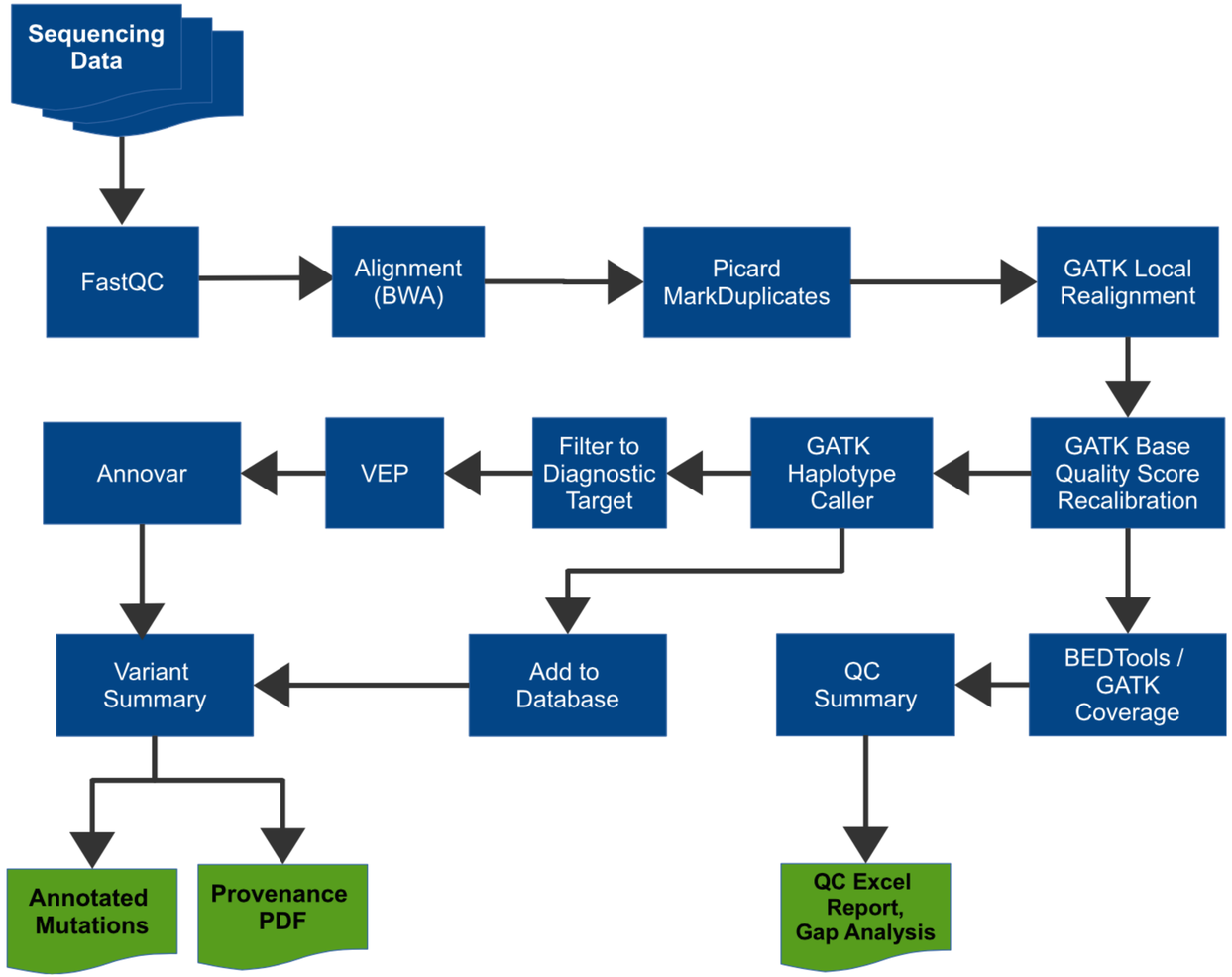
Cpipe

- Created and maintained by MGHA
- Groovy (Java), Python and Bash
- GPL v3
- https://github.com/MelbourneGenomics/cpipe

Cpipe is based on Bpipe

```
align_reads = {
    exec “bwa aln -t 8 $input > $output”
}
…
run {
    align_reads + [ dedup, calc_stats ] +
        call_variants
}
```

Cpipe Computational Workflow



Tools used within Cpipe

- Annovar (academic license)
- Bedtools (GPL v2)
- bpipe (BSD)
- BWA (GPL v3)
- FastQC (GPL v3)
- GATK (academic license)
- IGV (LGPL)
- igvtools (LGPL)
- Picard (MIT)
- R (GPL)
- Samtools (MIT)
- Snpeff (LGPL)
- Sqlite (custom, public domain)
- Trimmomatic (GPL v3)
- VEP (Apache)

Computational Infrastructure

- Pipeline execution infrastructure, provided by VLSCI:
  - 5 cluster nodes (x86 IBM iDataPlex)
  - 16 cores per node
  - 1TB of RAM per node
  - RHEL 6
  - Infiniband
  - IBM GPFS filesystem with Hierarchical Storage Management
  - SLURM

Image courtesy of IBM (http://www.ibm.com/)

Computational Infrastructure

- NeCTAR cloud for LOVD.
- VicNode storage for backup.
- Jira and Confluence for project management.
- Git and Github for revision control.

http://vicnode.org.au/

https://nectar.org.au/

Conclusion

- Clinical Genomics represents a nexus between computing and health care.

- Open source software systems and software engineering have a big impact in medical diagnosis and treatment.

We are hiring

- Two software development positions will be opening soon.

Acknowledgements

Melbourne Genomics Health Alliance

- Clara Gaff
- Natalie Thorne
- Tim Bakker
- Karen Meehan

VLSCI

- Andrew Lonie
- Peter Georgeson
- Anthony Marty
- Gayle Philip
- Candice McGregor