Bioinformatics Algorithms & Next-Generation-Sequencing (NGS) Data Analysis

T W Lam, David W L Cheung, R Luo

Department of Computer Science

April 2, 2018

Summary of the Impact

- Bioinformatics software technologies helping the genomics industry to resolve the major computational bottleneck in analyzing high-throughput sequencing data
- Technology integration & transfer via ITF projects — efficient, accurate & easy-to-use NGS data analysis for biomedical applications (e.g., Department of Health's Clinical Genetic Service, HKSH Pathology)
- A Spin-off Company (L3 Bioinformatics, 2014), with 60M of funding, to boost the Mainland's BioIT (e.g., BGI Online, UEC Helicube)

Underpinning Research (or Teaching & Learning)

- Explain the quality of the knowledge embodied in your KE project

1. An outline of what the underpinning research was

- Algorithmic research on constructing compressed data structures for text indexing and practical schemes for indexing human genome
- Bioinformatics research on NGS analysis for biological and clinical applications (e.g., metagenome assembly, DNA variation detection)
- Related bioinformatics software: BWT-SW, SOAP2, SOAP3, SOAP3-dp, BALSA, database.bio, MegaHit.

Underpinning Research (or Teaching & Learning)

2. When your research (or teaching & learning) was undertaken, and your role (and your team members' roles, if applicable) in the creation of such knowledge

- TW Lam: PI of algorithm & bioinformatics research, chair of L3B Bioinformatics
- DWL Cheung: Deputy Project Coordinator of ITF projects; executive director of L3 Bioinformatics
- R Luo: Co-PI of bioinformatics research, director and (former) CEO of L3B Bioinformatics

Underpinning Research (or Teaching & Learning)

3. Any relevant key contextual information about this area of research (or teaching & learning), e.g. where it is a wider body of research in collaboration with other institutions

- The work of NGS alignment algorithms and software was in collaboration with BGI.
- The work on other NGS analysis tools was in collaboration with BGI, Novogene (Beijing) and Hong Kong Sanatorium & Hospital.
- The work of metagenome assembly software was motivated by the Great Prairie Soil Metagenome Grand Challenge (JGI, US)

Underpinning Research (or Teaching & Learning)

4. Innovativeness of the knowledge arising from your research (or teaching & learning) at HKU

- We merge the CS theoretical work on compressed text indexing into bioinformatics applications.
- Skillful indexing not only can drastically speed up many computational-intensive bioinformatics tasks but also improves the quality of results.
- Easy-to-use NGS analysis for biomedical users (with minimal bioinformatics background) can realize a bigger value of NGS

Underpinning Research (or Teaching & Learning)

5. Significance of the key insights or findings from the research (or teaching & learning) that relate to the impact achieved by the KE project

- We were the first to shorten the alignment time for a typical NGS dataset (WGS, ~100Gb) from days or weeks to hours. Our solution is based on algorithm and software advancement instead of using more hardware, making it scalable to handle a large number of datasets.
- We were the first to be able to assemble complex metagenomic dataset (typically a few hundred Gb) entirely, and with good quality.

Engagement

1. Explain the engagement process through which the knowledge arising from the research (or teaching & learning) described above was shared with or transferred to the target beneficiaries

- Pilot projects (with ITF funding support) to integrate our bioinformatics technologies into the business
- Joint research on user requirement & methodologies
- Business partnership via spinoff company; open-source versus licensing
- Joint workshop and support to other institution's research projects

Engagement

2. External partners, if any

BGI (Shenzhen 300766)

UEC (Shenzhen 002642)

Novogene (Beijing)

Hong Kong Sanatorium & Hospital

Department of Health (Hong Kong)

Joint Genome Institute (US)

Engagement

3. Innovativeness of the engagement approach

A deep understanding of how to elevate the business of the partners (genomics/BioIT/clinical) with new bioinformatics technologies

Impacts Achieved

1. Beneficiaries - the non-academic sector(s) or organizations that have benefitted or been impacted on

Genomics sector

BioIT sector

Clinical sector

Impacts Achieved

2. Nature and extent of the impact

- Improving the efficiency of the genomics industry: Since 2008, SOAP2 and its successors have been heavily used by BGI for WGS analysis, notably in a number of clients' projects leading to nature series/science publications.
- Our NGS analysis solution has been the primary solution of the Department of Health for analyzing genetic diseases since 2015.
- Via our spin-off company, we've built the first bioinformatics cloud for the Mainland (BGI Online) to support NGS analysis, which was a key initiative when BGI went public in 2016.
- UEC has acquired our NGS analysis solution in 2016 to boost their BioIT business in the Mainland.
- Indirect impact: The source code of our compressed indexing software has inspired other researchers in developing their bioinformatics software. A typical example is Sanger Institute's BWA, which is perhaps the most popular alignment software.