Research Assessment Exercise 2020
Impact Case Study

University: The University of Hong Kong (HKU)
Unit of Assessment (UoA): UoA 13, computer studies/science (incl. information technology)

Title of case study: Bioinformatics Algorithms that Revolutionized Next-Generation-Sequencing (NGS) Data Analysis for the Genomics Industry

(1) Summary of the impact

In the early 2000s, the Department of Computer Science at HKU pioneered the algorithmic research in indexing the human genome with compressed data structures. The department's technologies subsequently resolved the genomics industry's computational bottleneck in aligning DNA next-generation-sequencing (NGS) data, reducing the time taken from days to hours per sample and markedly improving accuracy. The research team then extended its work to NGS bioinformatics for medical diagnosis. In 2015, the HK Department of Health adopted the department's NGS bioinformatics system to diagnose genetic disorders, improving the throughput from around a hundred cases per year to 1,900 cases in the first 1.5 years, and more importantly, resolving previously unknown cases with unprecedented accuracy.

In 2014, a spin-off company was founded, which has built the first NGS bioinformatics cloud in China (BGI Online) and received 60M HKD of investment/revenue.

(2) Underpinning research

Algorithm research. The department, led by Professor TW Lam, initiated the algorithmic study of constructing compressed text indexing in 2001 and published the first such algorithm in 2002 [3.1]. Since then, the department has worked on a series of theoretical problems related to compressed text indexing, publishing ten conference papers and seven journal papers between 2003 and 2015. At the same time, envisioning the practical value of its theoretical work, the department also researched on the application of its algorithms, in particular for indexing the human genome in the main memory (first publication appeared in the Sixth Workshop on Algorithm Engineering and Experiments, 2004). This work led to the first practical solution for indexing the human genome in a personal computer to support efficient pattern matching. Furthermore, the department engineered its DNA-indexing software to suit the bioinformatics requirement for a higher error rate, and in 2007 released the most efficient software library in the world for indexing the human genome (followed by a paper in Bioinformatics 08 [3.2]). This work has been well received by the algorithm and bioinformatics communities.

Research team members include HKU-CS staff TW Lam (PI) & SM Yiu, former PhD students WK Sung (now professor at National U of Singapore), WK Hon (now professor at National Tsinghua U), CK Wong & SL Tam, and former postdoc K Sadakane (now professor at Tokyo U).

Bioinformatics research. In the late 2000s, the emergence of high-throughput and low-cost sequencing (i.e., NGS) technologies significantly eased the generation of DNA data. Yet to analyze the huge amount of NGS data was a big computational challenge (a single whole-genome-sequencing (WGS) dataset contains over a billion short DNA fragments). In 2008, the

department started to collaborate with BGI, a private genomics company which then had the largest sequencing throughput in the world. By adapting the department's indexing technology to tolerate the sequencing errors in short NGS fragments, the researchers developed a totally new core engine for NGS alignment, which was adopted into the BGI analysis pipeline SOAP2 (published in Bioinformatics 09 [3.3]). SOAP2 was more accurate than its predecessor and ten times as fast. Subsequently the department extended its indexing technology to the GPU architecture and developed different scheduling strategies to support GPU-based parallel alignment. The resulting software packages SOAP3 (Bioinformatics 12 [3.4]]) and SOAP3-dp (PLOS One 13) prompted another tenfold improvement (enabling one million DNA fragments to be processed in ten seconds on a single computer). Based on the department's indexing technologies, various NGS analysis software programs have subsequently been developed. The latest example is MegaHIT for metagenome assembly (Bioinformatics 15 [3.5]), which received the HKU Faculty Best Research Output Prize in 2016.

Members of the research team include HKU-CS staff TW Lam (PI), SM Yiu, D Cheung & HF Ting, former PhD students CM Liu, R Luo, D Li, T Wong and E Wu, BGI/Novogene collaborators RQ Li, C Yu, YR Li and B Wang.

(3) References to the research

Algorithm research

[3.1] A Space and Time Efficient Algorithm for Constructing Compressed Suffix Arrays. Algorithmica 48(1): 23-36 (2007), WK Hon, TW Lam, K Sadakane, WK Sung, SM Yiu (preliminary versions appeared in proceedings of COCOON 02 and ISAAC 03).

[3.2] Compressed indexing and local alignment of DNA. TW Lam, WK Sung, SL Tam, CK Wong, SM Yiu. Bioinformatics 24(6): 791-797 (2008)

Bioinformatics research

[3.3] SOAP2: an improved ultrafast tool for short read alignment. R Li, C Yu, Y Li, TW Lam, SM Yiu, K Kristiansen, J Wang. Bioinformatics 25(15): 1966-1967 (2009). (Google Scholar: 2,967 citations)

[3.4] SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. CM Liu, T Wong, E Wu, R Luo, SM Yiu, Y Li, B Wang, C Yu, X Chu, K Zhao, R Li, TW Lam. Bioinformatics 28(6): 878-879 (2012) (Google Scholar: 224 citations)

[3.5] MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. D Li, CM Liu, R Luo, K Sadakane, TW Lam. Bioinformatics 31(10), 1674-1676 (2015) (Google Scholar: 631 citations)

(4) Details of the impact

[4.1] Adoption by the Genomics Industry

Two notable examples are SOAP2 for BGI (Shenzhen.300766) and MegaHIT for JGI (US Department of Energy). (1) BGI was the largest sequencing service provider in the world in the late 2000s. Since 2008, the alignment software SOAP2 has been BGI's primary alignment tool for its whole genome sequencing (WGS) service, supporting hundreds of biological and medical

projects under a business model of contract research. SOAP2 was able to boost the efficiency of BGI, contributing to BGI's success in the early 2010s. During the assessment period (i.e., from Oct 2013), Google Scholar [5.1] has recorded at least 100 BGI projects using SOAP2. (2) JGI of the US Department of Energy maintains perhaps the largest platform of microbial genomes in the world. Soon after the metagenome assembly software MegaHIT was released, JGI provided thorough benchmarking and technical feedback. In 2015, JGI announced MegaHIT as its standard metagenome assembly pipeline for 580 metagenome projects [5.2]. MegaHIT is able to assemble metagenomes (especially complex ones) which were not previously possible to assemble, and provides novel information about different microbial environments (e.g., soil) which have an important impact on human wellness.

[4.2] Medical applications

The department's success in the basic analysis of NGS data has attracted support from the HKSAR Government for the development and trial of bioinformatics software for clinical diagnosis (ITF grants of HKD 9M [5.3]). The department has developed ultra-fast software tools for detecting, annotating and visualizing variants with superior accuracy (collectively referred to as BALSA). In 2014 the department launched a pilot study of its integrated NGS bioinformatics system at the Clinical Genetic Service (CGS) of the Department of Health. Prior to this pilot study, CGS used to outsource bioinformatics analysis. The turnaround time was often over a month, and the analysis did not always meet CGS's clinical requirements. BALSA was set up in CGS in late 2015. Within 1.5 years, CGS analyzed 1,900 cases (for Mendelian Diseases) [5.4]. The time required for analysis was markedly shortened to tens of minutes per case, and over 99% of detected variants were confirmed. Some previous undiagnosed cases were resolved. More importantly, our bioinformatics system enabled CGS's doctors to conduct NGS analysis themselves, i.e., without the support of bioinformaticians. The Hong Kong Sanatorium & Hospital (HKSH) was another pilot user [5.5]. The impact of the department's work is also reflected in the appointment of TW Lam as a member of the Government's Steering Committee of Genomic Medicine in 2018. In the same year, he also received a HKU Faculty KE Award.

[4.3] Technology spin-off and Licensing

In 2014, TW Lam, together with Professor D Cheung and Dr R Luo (HKU), co-founded L3 Bioinformatics Limited (L3B). BGI invested USD 2,000,000 for 40% of L3B's shares [5.6]. At that period, large-scale NGS analysis demanded extensive resources and expertise, preventing many users from conducting their analysis. L3B developed a public cloud platform for BGI (at a cost of USD 1,000,000) to deliver and manage clients' sequencing data and to provide easy-to-use NGS analysis on a demand basis. This platform, BGI Online, was launched on Amazon AWS in 2015, primarily to serve BGI's clients in the US. It was cited as a key initiative when BGI went public in 2016 (as reported in the BGI IPO Prospectus [5.7]). BGI Online is now managed by Aliyun (of Alibaba) to serve mainland users [5.8]. L3B at its peak had 16 employees, and received over 60 million HKD of investment and revenue between 2014 and 2018. Ruibang Luo, the CEO of L3B, was included among the 30 entrepreneurs in the Forbes 30 under 30 list in 2017, in the field Healthcare & Science [5.9].

HKU (via Verstitech) granted a non-exclusive source-code license of BALSA to L3B for a fee of USD 150,000 in 2014, and later to United Electronics Limited (Shenzhen.002642) for USD

150,000. United Electronics has packaged the bioinformatics tools to serve medical and genomics users in China (e.g., 浙江省肿瘤科医院, 香港大学深圳医院, 金准基因).

[4.4] Indirect impact

Besides its involvement with SOAP2, the department's work on indexing and alignment (specifically, open-source software BWT-SW) has inspired other research teams in their development of alignment software. A typical example is BWA, which was developed by the UK's Wellcome Sanger Institute and has adapted the source code of BWT-SW (as reported in BWA's publications [5.10]). BWA has an even wider take-up than SOAP2. For example, it was the primary analysis tool used for the 1000 Genomes Project to discover genetic variations.

(5) Sources to corroborate the impact

[5.1] The following link shows the citations to SOAP2 from 2014 to 2019, which include at least 100 BGI projects using SOAP2:
https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&sciodt=0%2C5&cites=11842085765588565880&scipsc=&as_ylo=2014&as_yhi=2019

[5.2] JGI & MegaHIT:
https://3q8i7m48ig9en9v121qx83t166i-wpengine.netdna-ssl.com/wp-content/uploads/sites/2/2015/04/04_Alicia_2015.03.23.JGI-user-meeting-GT-workshop-Assembly-talk-Clum-FINAL.pdf

[5.3] ITF (Innovation Technology Fund) grant records: ITF HKD 5.85M (2013 - 2015) + ITF PSTS HKD 3.2M (2015-16): A Genomic and Pharmaceutical Knowledge-based System for Clinical Diagnosis and Case Repository, PI: T W Lam

[5.4] Department of Health-CGS report (two pages)

[5.5] HK Sanatorium & Hospital support letters (two pages)

[5.6] [……]

[5.7] http://www.csrc.gov.cn/pub/zjhpublic/G00306202/201512/P020151218527195532976.pdf

[5.8] "Aliyun Partners with BGI to Launch Cloud-based Genome Analytics Engine", https://thetechrevolutionist.com/2016/02/aliyun-partners-with-bgi-to-launch.html

[5.9] Forbes 30 under 30 2017: Healthcare & Science:
https://www.forbes.com/30-under-30-asia/2017/healthcare-science/#3f97c0ce1722

[5.10] The following two references of BWA both mentioned using our open-source software BWT-SW: doi:10.1093/bioinformatics/btp324; 10.1093/bioinformatics/btp698