Research Assessment Exercise 2020
Impact Case Study

University: The Hong Kong University of Science and Technology
Unit of Assessment (UoA): 11-Mathematics & Statistics
Title of case study: Statistical and computational methods for analyzing genetic data of East Asian populations

(1) Summary of the impact

The knowledge transfer and commercial adoption of Can Yang's HKUST-developed statistical and computational methods for large-scale genetic data analysis, in particular that of East Asian populations, has directly contributed to hi-tech economic impact and skills upgrading in leading-edge fields such as machine learning, genotyping and sequencing in the Greater Bay area. This has been achieved through the novel methods' evidenced role in the exponential growth of a direct-to-consumer DNA ancestry testing platform and personalized healthcare testing provider WeGene (2016: 8,000 customers; July 2019: about 300,000 customers). The new analytical methods have enabled company scientists and collaborators to publish China's first genotyping studies of the country's different ethnic groups. The company's work has been shared at major international conferences and attracted global media attention for its focus on Asian heritage and health issues. The research has also had a direct effect on global data analysis practitioners, through the inclusion of the HKUST mathematics team's BOOST method into the 1.9 version of PLINK, a standard computer toolkit used to handle genetic data by researchers globally.

(2) Underpinning research

The Human Genome Project, completed in 2003, opened new opportunities for each individual to know their DNA make-up, with benefits ranging from medical care and managing disease risks, to satisfying age-old questions about who we are and where we come from. As a result, there is now a rapidly growing market for saliva-based DNA test kits, valued at US$525.5 million in 2018, according to Transparency Market Research, and forecast to grow 23.7% a year between 2019 and 2027. However, the great majority of this is based on datasets from Caucasian ethnic groups in western countries, and the technology could still be greatly improved in its power, efficiency and range of capabilities.

Can Yang (PhD Electronic and Computer Engineering, HKUST, 2011; re-joined as Assistant Professor, Department of Mathematics, 2017) and his group have used their statistical and machine learning expertise to make significant advances in DNA-testing technology, and equip it to better cater for genetic differences found in East Asian populations. Collaborating with the Shenzhen-based genomics platform WeGene, the team has developed statistically accurate and computationally efficient methods facilitating direct-to-consumer genetic data analysis and other services, such as "From DNA Variations to Human Faces", enabling customers to see what their ancestors looked like.

Yang initially worked on computational architectures to accelerate genetic analysis. The BOOST method [**See Section 3, R1**] was designed jointly by Yang and Xiang Wan (Visiting Assistant Professor at HKUST) to detect gene-gene interactions in genome-wide association studies (GWAS). Their new data structure and associated operations introduced Boolean representations (used in algebra to create true and false statements) and Boolean operations to handle genotype data, accelerating genotype matrix computations to 60 times faster than traditional approaches. GBOOST **[R2]** further enhanced the BOOST method by exploring the computational power of graphics processing units (GPU), yielding 40 times faster results. The BOOST method was incorporated into the most popular GWAS data analysis toolkit PLINK in 2013, released as part of the 1.90 version in 2015, and incorporated since then (https://www.cog-genomics.org/plink2/epistasis). When Yang returned to HKUST in 2017, this novel data structure (Boolean representation and operations) was adapted for and adopted by WeGene in 2018 for calculating whether a DNA segment is identical by state (IBS) in two or more individuals.

Starting from 2017,Yang and his team also developed a statistical method named "LEP" for genetic risk prediction **[R3]**. Nearly 80% of publicly available genotype data are from European ancestry and less than 10% from East Asia, making risk prediction in East Asian populations less accurate due to the smaller sample size. The LEP method advanced the field of risk prediction by constructing probabilistic models to take genetic correlation into account. Computational challenges of statistical inference when dealing with large-scale genomic data in ultra-high dimensions were overcome via tractable approximations using mean-field variational Bayes, providing 5% improvement of risk prediction compared to other methods, which usually focused on one phenotype at a time. The LEP method has also been adopted by WeGene for genetic risk prediction.

In the field of artificial intelligence, Yang's team developed a novel method named "VGrow" for learning deep generative models – generative adversarial networks (GAN) via variational gradient flow **[R4]**. The VGrow method can generate realistic photos in a stable and efficient manner, with theoretical guarantees. This has been adopted by WeGene for its new "From DNA Variations to Human Faces" service. The success of the HKUST-WeGene collaboration resulted in it being awarded a HK$1 million partnership research program grant by the Hong Kong government's Innovation Technology Fund to further develop large-scale genomic data analysis and commercialize the technology, paving the way for increasing impact. Yang was made director of HKUST's Health Data Analytics Lab in 2018.

(3) References to the research

[R1] X. Wan*, **C. Yang***, Q. Yang, H. Xue, X. Fan, N. Tang and W. Yu. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325-340. *Joint first author. 2010.

[R2] L.S. Yung, **C. Yang**, X. Wan, and W. Yu. GBOOST: A GPU-based tool for detecting gene-gene interactions in genome-wide case control studies, *Bioinformatics*, 27(9):1309-1310, 2011. https://doi.org/10.1093/bioinformatics/btr114.

[R3] M. Dai, X. Wan, P. Hao, Y. Wang, Y. Liu, J. Liu*, Z. Xu*, and **C. Yang***. Joint analysis of individual-level and summary-level GWAS data by leveraging pleiotropy. *Bioinformatics*. 15;35(10):1729-1736. 2019. https://doi.org/10.1093/bioinformatics/bty870.

[R4] Y. Gao, Y. Jiao*, Y. Wang, Y. Wang, **C. Yang***, S. Zhang. Deep Generative Learning via Variational Gradient Flow. International Conference on Machine Learning. 2093-2101. 2019. *Corresponding authors.

Funding: The research team was supported in part by the Hong Kong Research Grants Council [No.12301417 (HK$472,351) and No.16307818 (HK$456,452)] and National Science Foundation of China [61501389 (RMB210,000)].

(4) Details of the impact

Yang and his team's research is having significant ***economic impact*** by supporting corporate growth and success; ***impact on human health*** by providing new tools for personal health protection and medicine; ***societal impact*** by satisfying individuals' curiosity about their ancestry and public awareness of DNA testing; and impact on ***practitioners*** making use of the team's advances in understanding gene technology.

The Yang team's research has had significant ***economic impact*** on the development of WeGene, a wholly-owned subsidiary of WeGene (Shenzhen Zaozhidao Technology Co, Ltd). After Yang rejoined HKUST as a faculty member in 2017, he was approached by Dr Gang Chen, Chief Executive Officer of WeGene, to collaborate with the company. An agreement for scientific co-operation was signed between the Department of Mathematics at HKUST and WeGene to develop and transfer their statistical and machine learning advances for DNA testing **[See Section 5, S1, S2]**. WeGene, launched in 2014, began its genomic testing services using genotyping arrays optimized for Chinese people in November 2015, but initially only provided genotyping and whole-genome sequencing testing and analysis services to consumers. It collected large amounts of genomics and phenotypic data (clinical

information) but lacked the tools to analyze them. Working with the HKUST math team facilitated a rapid expansion of its services and market by leveraging its knowledge and skills in statistical methods and massively parallel algorithms. These were used to develop novel and high-performance bioinformatic tools based on data structure specialized for genotype data (i.e., Boolean representation and operation [**See Section 3, R1, Section 5, S1**]). In addition, the GPU-accelerated genomic data computing **[R2**] developed at HKUST proved 25 times faster than traditional CPU-based algorithms, reducing cloud-associated business costs by 50%-60% per annum [**S2**].

This new capacity also allowed WeGene to handle complex and high-throughput genomic and phenotypic data, as applied in the company's "From DNA Variations to Human Faces" service. WeGene used the Yang team's deep-learning technique VGrow to generate faces of ancestors, based on customer images and information, and those of thousands of faces gathered for training purposes. This has satisfied interest among about 15,000 consumers so far. These customers have purchased the service for RMB499 per unit, generating income of RMB7.48 million since February 2019. **[S1]**

WeGene CEO Chen affirmed the significant ***economic impact*** that Yang's research had generated, as reflected in the rapid expansion of the company's services, market and income. He writes, in testimony: "*WeGene has been expanding rapidly over the past few years, with our user base growing from 8,000 customers in 2016 to more than 300,000 as of July 2019, and I keenly appreciate that Prof Yang's research findings have directly contributed to this exponential growth."* As its market expanded, the company had increased its staff from 24 in 2016 to 82 by July 2019, while the successful collaboration prompted its decision to open a branch and laboratory in Hong Kong **[S1**], which is due to extend its service to Hong Kong and Southeast Asian markets from 2020. The Hong Kong laboratory is also collaborating with Illumina, a leading US-headquartered developer, manufacturer and marketer of life science tools and integrated systems for large-scale analysis of genetic variation and function, and has received a HK$1 million grant from the Hong Kong Innovation Technology Fund **[S3].** The company's success and potential has also been recognized by being named among the best companies in technology and healthcare innovation (4th China Healthcare Industry Summit) **[S4].** The HKUST-WeGene collaboration and company's growth has thus supported Hong Kong and Greater Bay Area policy to develop as a leading innovative hi-tech science and technology hub, first in Shenzhen and now Hong Kong. In March 2019, WeGene formally invited Yang to serve as a scientific consultant to the company.

WeGene's brochure spells out the benefits of the technology for *human health* and *well-being* and personalized medicine from its services, including helping individuals identify more than 90 health risks related to genes (e.g. some forms of cancer, Type 2 diabetes) and potential hereditary diseases that can be mitigated when known **[S5]**, facilitated by the more accurate LEP technology. By 30 September 2019, WeGene's user forum had received over 10,000 comments from its customers. Most comments reflect with interest on their health risks related to the genetic information they had gained from the service, as well as ancestry and kinship networks. Some indicated greater awareness of the need to maintain healthy lifestyles, and positive behavior changes that would increase their well-being. For example, on discovering a greater risk for diabetes, one reflected: "*I bought a sport band and started to do more exercise right after the test*." Another wrote of the need to "*maintain a disciplined life, and do more exercise. Otherwise, I will definitely have diabetes.*" **[S6]**

The research has had ***impacts on professional practice and society***, with the new technical tools enabling WeGene scientists and collaborators to undertake and publish genotyping studies on different ethnic groups in China, contributing to precision health in China. One example focuses on the Jing people, an ethnic group in Guangxi, Southwest China, published in the *American Journal of Physical Anthropology (AJPA*), 2018. Another example, also published in *AJPA,* investigated the genotype of individuals from Tibetan-Yi Corridor and provided insight into migration to the Tibetan plateau since Neolithic times from surrounding lowlands.

The Yang team's BOOST advance has also had wide-reaching ***impact on practitioners*** in the global genetic data analysis field through the incorporation of BOOST into the PLINK (1.9 version). First-generation PLINK (1.07 version) was developed in 2007. Researchers worldwide, including leading universities such as Harvard and Oxford, use PLINK as a standard toolkit to handle genetic

data. Second-generation PLINK (1.9 version) was re-designed to handle large-scale genetic data much more efficiently and released in 2015. The HKUST mathematics team's BOOST method was incorporated into that version to serve as a standard modular in the field of gene-gene interaction detection, "which has greatly facilitated data analysis at WeGene", while BOOST has been cited more than 400 times since its publication **[S1]**.

WeGene's innovative services have been shared by the company in dozens of conferences and seminars to inform the public and commercial and academic practitioners of the latest DNA scientific advances, including: (i) Gu Dafu Salon Phase II: Consumer Genetic Testing and Health Consulting and Management, Aug 2018; (ii) The 10th National Congress of the Chinese Society of Genetics and Academic Symposium, Nov 2018; (iii) European Conference on Human Genetics, June 2019. The new services have attracted extensive national and global media interest **[**e.g. **S7, S8]**. They are also included in global databases and reviews of DNA test kits **[**e.g. **S9, S10]**. Top10DNAtests.com describes WeGene as "one of the rare companies that specializes in the genetic exploration of Asian heritage" and highlights benefits including "numerous fitness and health-related results"; "reputable partners"; "science behind the testing based on thousands of peer-reviewed scientific studies"; "simple testing procedure"; "quick turnaround time" and "accurate and reliable results" **[S9]**.

Sources to corroborate the impact

[S1] Letter, WeGene CEO. [On file]
[S2] "WeGene signed a strategic cooperation agreement with HKUST, promoting the development of genetic big data and Artificial Intelligence", 27 March 2019. https://www.iyiou.com/p/96384.htm
[S3] Letter of support for ITF Partnership Research Program (PRP) [On file]
[S4] 4th China Healthcare Industry Summit (CHIS), Beijing, 25-27 July 2019. https://www.sohu.com/a/330103019_115035
[S5] Corporate brochure. [On file]
[S6] Website user feedback. [English translation on file]
[S7] "Chinese DTC genomics firm WeGene plans to go international, expand cohort research", 20 September 2018. https://www.genomeweb.com/business-news/chinese-dtc-genomics-firm-wegene-plans-go-international-expand-cohort-research
[S8] "Is China the wild west of consumer DNA testing?" Caixin CX Daily Briefing, 25 August 2017. [On file]
[S9] Top10DNAtests.com. https://www.top10dnatests.com/reviews/wegene-review/
[S10] dnatestingchoice.com. https://dnatestingchoice.com