世界首創支援完整次世代定序識別基因變體之系統

文／翁健棋

自 1975 年，諾貝爾獎得主桑格發明了雙脫氧鏈終止法，也就是第一代的 DNA 定序技術，為 DNA 定序研究領域打下紮實基礎後，讓人類得以跳脫巨觀角度，轉而從微觀的 DNA 序列切入，分析生物體的化學表現。然而桑格測序法測序成本高、通量低與耗時長等缺點，使得相關應用難以普及；在各國龐大的資金與資源投入下，催生了次世代定序（Next Generation Sequencing, NGS）的出現，不單降低定序所需的成本，也讓定序檢測不再受限於基因大小與多寡。

本院資訊工程學系洪瑞鴻教授，專業研究領域即為與之相關的次世代定序演算法、生物資訊分析等範疇。去年九月，洪瑞鴻教授與指導團隊以可應用在高速疾病檢測、生物醫療診斷、生物資訊分析、物種偵測等多元領域的「適用於次世代定序識別基因變體之系統晶片」作品，榮獲以「科學突破性」及「產業應用性」兩大指標作為評選重點，科技部所授予之 2021 未來科技獎。該獲獎作品為全世界第一個支援完整次世代定序識別基因變體之系統，搭配團隊所設計與下線之系統單晶片與客製化電路板、周邊電路，可以達到全世界最快的運算速度，相較於高階顯卡有 66 倍之加速幅度。同時，系統單晶片可支援四種運算，包括：資料前處理、短片段回貼、半倍體搜尋與變體識別。本成果後續在今年一月舉辦的 2022 年消費性電子展 (CES) 展出。

透過設計團隊所開發之 sBWT 演算法，搭配上現有的基因組分析套件，獲獎作品之精準度可與軟體平台一致。此外，該系統晶片透過臺積電 28nm 製程下線，可操作在最高 400MHz，功耗為 0.975W，在 37 分鐘內便可以完成完整基因資料分析。與高階顯卡相比，不僅速度有顯著提升，於能量效率與面積效率上，亦有數個量級以上的增益。同時配合高度平行、硬體共享、複雜度化簡等硬體優化技巧，使設計作品得以達到高效能、低功耗之特性。此晶片亦設計了「多工排序引擎」與「動態規劃處理引擎」兩個主要運算單元，用來支援整個基因定序資料分析的複雜運算。不單如此，設計系統內整合一顆 Synopsys ARC 處理器，可用於檔案傳輸介面、記憶體資料與 IP 控制等等，以增加系統彈性。此系統除經由 FDA 之標準測資完成驗證，可達到 99.6% 精確度外，所設計之客製化 GUI 亦可滿足即時判讀之需求。

原先受限於運算能力上限，使用一般之 DNA 資料分析工具，就算搭配上高端的 GPU，也需要超過三天的時間才能分析完整的人類 DNA 序列。獲獎作品「適用於次世代定序識別基因變體之系統晶片」的出現，透過短序列回貼、串連各組回貼好的 DNA、與 DNA 資料庫對比三步驟，識別出變異位置，有效率地在 40 分鐘內提供完整分析，保持高準確性的同時，大幅降低測序時間。快速辨識出的基因變體不單可應用於疾病診斷，亦可用於病毒基因演化追蹤、胎兒基因檢測等各層面，可謂基因工程發展之重要里程碑，也再次恭賀獲獎的洪瑞鴻教授與指導團隊！

World's First Complete Next-generation Sequencing System to Identify Genetic Variants

In1975, Frederic Sanger, a Nobel Laureate, invented the Dideoxy termination method, which was adopted as a primary technique in the “first generation” of DNA sequencing applications. Laying a solid foundation for DNA sequencing research, Sanger’s discovery leads human beings to move beyond macroscopic perspective to microscopic DNA sequence to analyze chemical interactions of organisms. However, the disadvantages of Sanger sequencing, such as high cost, low throughput, and time consuming, make related applications difficult to popularize. Therefore, with huge contributions in various countries, the emergence of Next Generation Sequencing (NGS) not only reduces sequencing cost, but also allows multiple genes to be analyzed at once and can detect all types of variants.

Professor Jui-Hung Hung of the Department of Computer Science at NYCU specialized in next-generation sequencing algorithms and Bioinformatics. Last September Professor Hung and his team won the 2021 FUTEX Future Tech Award from the Ministry of Science and Technology, which was evaluated by two criteria: scientific breakthrough and industrial practicability, for their work "Genetic variant discovery SoC for analyzing Next-generation sequencing data'' that can be widely used in high-speed disease detection, biomedical diagnosis, bioinformatics analysis, and species detection, etc. The award-winning work is the first complete NGS data analysis system in the world. Integrating with a SoC, a customized circuit board and peripheral circuits which were designed by the team, the system can reach the fastest computing speed in the world. Their work achieved 66 times speed-up compared to existing high-level GPU platforms. Meanwhile, the SoC supports four kinds of operations: data preprocessing, short reads mapping, haplotype search, and genome variations detection. This work was also exhibited in the Consumer Electronics Show (CES) in January 2022.

Integrating the sBWT algorithm developed by the team with the existing genome analysis suite, the winning work achieved an accuracy as consistent as the software platform. In addition, the SoC, manufactured by TSMC's 28nm process, can run at a maximum of 400MHz with a power consumption of 0.975W, and completes an entire genetic data analysis in 37 minutes. Compared to existing high-level GPU platforms, the system not only significantly accelerated the analysis, but also increased energy efficiency and area efficiency by several orders of magnitude. Meanwhile, combining high parallelism, hardware sharing, complexity reduction, and other hardware optimization techniques, the work achieved high performance and low power consumption. The SoC comprised two main computing units, "multiplex sequencing engine" and "dynamic programming processing engine", to handle the complicated operations of the entire sequencing data analysis. Furthermore, the system integrated a Synopsys ARC processor, which could be used for file transfer interface, memory data and IP control, etc., to increase system flexibility. In addition to the system accuracy of 99.6 validated on standard test data of FDA, the customized GUI of the system could also offer real-time interpretation.

Because of the limits of computing power in the past, it would take more than three days to analyze a complete sequence of the human genome using common DNA data analysis tools even with a high-end GPU. The award-winning work "Genetic variant discovery SoC for analyzing Next-generation sequencing data" adopts a workflow with short reads mapping, haplotypes reconstruction using de Bruijn graph for sequence assembly, and comparison with the DNA database to efficiently identify the variant position and deliver a complete analysis in 40 minutes. It significantly reduces the time of genome sequencing while maintaining high accuracy. The technique of rapidly identifying gene variants can be used not only for disease diagnosis, but also for virus gene evolution tracking and fetal genetic testing, etc. Therefore, it can be seen as an important milestone in the development of genetic engineering. Once again, congratulate Professor Jui-Hung Hung and his team on their success!