Revolutionizing Cancer Genomic Medicine by AI and Supercomputer with Big Data

Satoru Miyano and IMSUT Cancer Clinical Sequence Research Team

Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo, Japan
miyano@ims.u-tokyo.ac.jp

Abstract

We are running a cancer clinical sequence system based on whole genome/exome, RNA sequence and epigenome as research. When focused on hematology/oncology, it takes currently four days for a patient from signing informed consent (IC) to diagnosis. This process consists of IC, specimen collection, next-generation sequencer analysis, data analysis, interpretation/translation of mutations, determining the diagnosis combined with all pathological/clinical data and returning the result to the patient. Therapies are not only drugs but also hematopoietic stem cell transplantation. A pipeline Genomon for analyzing cancer genomes and RNA sequences by next-generation sequencers plays one of the key roles. It is running on the supercomputer system at Human Genome Center. The bottleneck of interpretation/translation was drastically resolved by employing IBM Watson for Genomics in harmony with our in-house human curation pipeline. We report how our system works as a conglomerate of oncologists, cancer biologists, bioinformatics experts augmented with Watson and Genomon.

Introduction

Cancer is a very complex disease caused by a variety of mutations in genomes that evolve spatiotemporally in our body. The knowledge about cancer is also getting deeper and broader. More than 200,000 papers with keywords tumor/cancer were published only in 2017 and rapidly increasing. More than 5,000,000 coding mutations are reported in 25,000 papers. Reported biological mechanisms in the literature are digitalized and enormous. When we sequence one patient’s whole cancer genome, we face with several hundreds to millions of mutations. Pathological diagnosis is often incorrect due to the complexity of cancer. Actually, the efficacy of anti-cancer drugs specified to organs is just around 20%. Cancer genome sequencing is a savior for precision medicine. However, it is beyond human abilities to clinically interpret the patient genome sequence information with big data. The speed also matters seriously. To overcome these difficulties, we are using next-generation sequencers for genome sequencing, the supercomputer SHIROKANE for sequence data analysis, and IBM Watson for Genomics and several databases for clinical interpretation of mutations. By uploading the mutation data of a patient’s specimen, driver genes and candidate drugs will be recommended together with their reasoning information. It is currently possible for the conducting doctor to return the diagnosis to the patient in four days that includes whole genome sequencing and data analysis time.
For cancer genome analysis, two big jigsaw puzzles, one for germline and the other for cancer, are solved by using supercomputer (Fig. 3). Detection of mutations is the first step to investigate cancer.

C. Big digitalized biomedical knowledge

In order to understand cancer genomes, we use various biomedical databases.

NIH PubMed Database (https://www.ncbi.nlm.nih.gov/pubmed/) compiles more than 27 million publication abstracts in 2017. If their full papers are printed and piled, its height exceeds 4km (higher than Mt. Fuji). As estimation, it will reach 100km in 2050. The database COSMIC, the Catalogue Of Somatic Mutations In Cancer (https://cancer.sanger.ac.uk/cosmic), is the world’s largest resource of somatic mutations in human cancer. In the v84 (released on the 13th February 2018), 5,448,850 coding mutations are linked to 25,807 papers by human-curation. Retrieving these databases is a heavy daily burden for cancer genomic scientists. ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) collects information about genomic variation and its relationship to human health. It is updated weekly. NIH Clinical.gov provides information of more than 250,000 clinical trials. The database of molecular interactions NDEx (http://www.ndexbio.org/) includes NIH National Cancer Institute’s Pathway Interaction Database (NCl-PID). It is a pathway interaction database that is a curated collection of information about known bio-molecular interactions and key cellular processes assembled into signaling pathways.

All are described with English, a natural language. Natural language processing technology should play a key role for the future of cancer genomics because no one can read them all.

D. Bottleneck

Millions of mutations can be found from the genome sequence data of one cancer patient. But their clinical/biological translation is a severe bottleneck [1]. We face with many mutations having biologically interesting features that may not have prognostic or therapeutic relevance. A major challenge is identification of clinically actionable mutations, where “actionable” means a potential target or risk factor that affects the treatment plan.

Genomon GO: Get Mutations!

Genomon (https://github.com/Genomon-Project) is a suite of bioinformatics tools for analyzing cancer genome and RNA sequencing data for Illumina NGS data (Fig. 4). It performs sensitive and accurate detection of most types of genomic variants (single nucleotide variants, short insertions/deletions, mid-size (20bp - 300bp) insertions/deletions and large scale structural variations), and transcriptomic changes (gene fusions, aberrant splicing patterns). Its good performance has been demonstrated through a large number of important cancer genome projects that led to very impactful discoveries [2-5]. Detection of mid-size insertions/deletions was considered as a “blind spot of Illumina NGS data” and Genomon solved this blind spot by sophisticated mathematical methods.

An efficient job scheduling framework realizes easy analysis of several hundreds of genome and transcriptome sequencing data simultaneously. Genomon is easy to install. After installation, users can start analysis just preparing simple sample configuration files. It is running on the supercomputer SHIROKANE at our Human Genome Center that consists of 550TFLOPS computation nodes at peak (Thin:5GB/Core; Fat: 2TB/node), 30PB Lustre File System and IBM Tape Archive System+1PB Nearline Disk (https://supcom.hgc.jp/english/). This combination of Genomon and SHIROKANE reduced the time for data analysis to less than an hour.

Watson for Genomics

A. Cancer clinical sequence system at Institute of Medical Science University of Tokyo

By the drastic advancement of sequencing technologies, we have extended the system to whole genome sequencing, and more in 2011. The ELSI team and the genetic counseling team are involved in this system. At the research hospital, currently, clinical sequencing covers colorectal cancer and blood disorders as research. We recognized that the bottleneck of the system is “clinical interpretation/translation” and the more serious problem is a lack of people who can do this. Human Genome Center and Health Intelligence Center are responsible for sequencing and data analysis with Genomon using the supercomputer system. Decision support system and effective utilization of biomedical big data are practical key issues. This is the main motivation for introducing IBM Watson for Genomics in our system. The concept is shown in Fig. 5. Various sequencers are employed in this system for precision medicine (Fig. 6). The total system is illustrated in Fig. 7 and Fig. 8.
B. Before Watson

When we started whole genome clinical sequence in 2011, Prof. Yoichi Furukawa, MD, PhD (Institute of Medical Science University of Tokyo, abbreviated to IMSUT) faced with a serious problem. A mystery was on a Familial Adenomatous Polyposis (FAP) patient who has no deleterious mutations on the coding regions (exons). Whole genome sequencing was done and we analyzed the sequence data by human eyes and database searches at that time. After heavy efforts of investigations, we found a deletion of 10kb (10,000 letters) in the upstream of \textit{APC} gene (Fig. 7) containing the promoter 1B. This deleted region causes reduced expression of APC-1B but not APC-1A by the deletion of promoter 1B that is responsible for familial adenomatous polyposis [6]. This interpretation was a result of huge labor and time. We thought we should have a new paradigm to go further. This was a main motivation to introduce Watson for Genomics in our clinical sequencing system.

C. After Watson

In July of 2015, our institute introduced IBM Watson Genomic Analytics (current version is called Watson for Genomics, abbreviated to WfG) as a research on supporting cancer clinical sequencing. WfG employs technologies of natural language processing and machine learning. It was trained at New York Genome Center for solid tumors with 20 million PubMed abstracts, 15 million patent data, COSMIC (Catalogue of Somatic Mutations in Cancer, UK), ClinVar (Genomic Variation and Health/Disease, NIH, USA), National Cancer Institute Pathways (NIH, USA), etc. From 2015, the hematology/oncology team directed by Prof. Arinobu Tojo, MD, PhD (IMSUT) joined in the project. The team started with the Illumina myeloid panel of 54 genes due to the cost and speed for diagnosis. WfG reads the list of mutations identified by Genomon as input, and suggest recommendations together with evidences in a way that clinicians can check and follow the reasons why such recommendations are generated. Initially, clinicians were skeptical about the recommendations by WfG and we found that WfG was not well-trained for blood tumors with the supervise of experts. After six months of training WfG by our hematologists/oncologists, its accuracy became the level of moderate acceptance. Simultaneously, we recognized only half of the patients could be diagnosed based on the 54 gene panel and we had to proceed to whole exome analysis. At that time, WfG covers only single nucleotide mutations. Currently, whole genome/exome, RNA sequencing as routine, and sometimes epigenetic analysis are done for diagnosis and search for drugs and therapies. Since WfG is still incomplete and under development, our in-house human...
Curation pipeline is inevitable, for example, RNA sequencing data, epigenetic data, time-course data and therapies other than drugs. The turnaround time from IC to diagnosis is now about four days for whole genome analysis. This was realized by the team work of concologists, cancer biologists, bioinformatics experts and in-house curation pipeline augmented with WfG and Genomon on supercomputers.

D. An example of analysis by WfG

The following is an analysis of colorectal cancer cell line RKO [7], not a patient data. Whole exome analysis revealed 4,237 single nucleotide variants (SNVs) in this cell line. The input for WfG is the list of these SNVs. WfG analysis finished in less than 30 minutes. Fig. 10 is a molecular profile tab that suggest strong driver genes mapped on chromosomes 1-22, X and Y.

Drug tab (Fig. 11) shows target genes and their candidate drugs. Fig. 12 is pathway map that provides the reasons for recommendations from the viewpoint of biological mechanisms. All gene names and drugs are linked to literature, documents for drugs and clinical trial information so that clinician can review the recommendations from the viewpoint of experts. Before introducing WfG, it was impossible to review all genes due to the limitation of time to diagnosis and labor workload to the experts.

Conclusion

Supercomputers reduced the time for data analysis for identifying genomic aberrations to the almost negligible level even for whole genome sequence data. However, interpretation/translation of aberrations is a sever bottleneck because of the number of mutations and the hugeness of background biomedical big data. This is getting severer year by year and already beyond human abilities. This is a reason why introduction of AI is inevitable. Diagnosis is responsible for medical doctors not AI. In our study, we learned that AI, in this study WfG, is not a technology to replace human experts but creates clinical experts equipped with AI exoskeleton.

Acknowledgments

We would like to thank all members of IMSUT cancer clinical sequence research team, especially, Prof. Yoichi Furukawa and his team for solid tumors, Prof. Arinobu Tojo and his team for hematology/oncology, Prof. Seiya Imoto and Prof. Rui Yamaguchi and their colleagues for informatics, and Prof. Yoichi Furukawa and Prof. Koichi Yuji for genetic counselling, and Prof. Kaori Muto for ELSI. This research is partly supported by MEXT grants: 15H05912, hp150265, hp160219, hp170227, and AMED grant JP18kk0205003.

References

[1] Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL. Organizing knowledge to enable personalization of medicine in


