GIST-Seoul National University Hospital develops AI technology for 3d cancer genome prediction that reduces costs and increases accuracy: Predicting Hi-C data, which is difficult to obtain at great cost, can identify abnormalities in gene expression regulation for individual cancer patients

- GIST Professor Hyunju Lee - Seoul National University Hospital Professor Sung-Hye Park's joint research team, developed AI model InfoHiC that predicts 3D cancer genome structure with high accuracy at low cost using whole genome information of cancer cells

- Compared to existing human reference genome-based models, prediction performance is greatly improved and verified by applying it to medulloblastoma patient data... Published as cover paper in international academic journal <Molecular Systems Biology>

▲ (From left) Professor Hyunju Lee of the GIST AI Graduate School, Professor Sung-Hye Park of Seoul National University College of Medicine, Department of Pathology, and Dr. Yeonghun Lee of GIST School of Electrical Engineering and Computer Science

While many studies are being conducted to identify mutations occurring in the genome of cancer cells in order to understand the pathogenesis of cancer, the importance of identifying not only point mutations occurring in genes but also the specific gene expression control mechanisms of cancer cells has been drawing attention recently.

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that the research team of Professor Hyunju Lee of the GIST AI Graduate School, together with the research team of Professor Sung-Hye Park of Seoul National University Hospital, developed an AI model, 'InfoHiC', that predicts the 3D cancer genome* using the whole genome information of cancer cells (the entire genome of a person).

In cancer cells, changes in the 3D genome play an important role in regulating gene expression patterns. Using Hi-C data*, neo-TAD* structures in 3D cancer genomes can be identified, but analysis is relatively difficult and expensive compared to whole genome data.

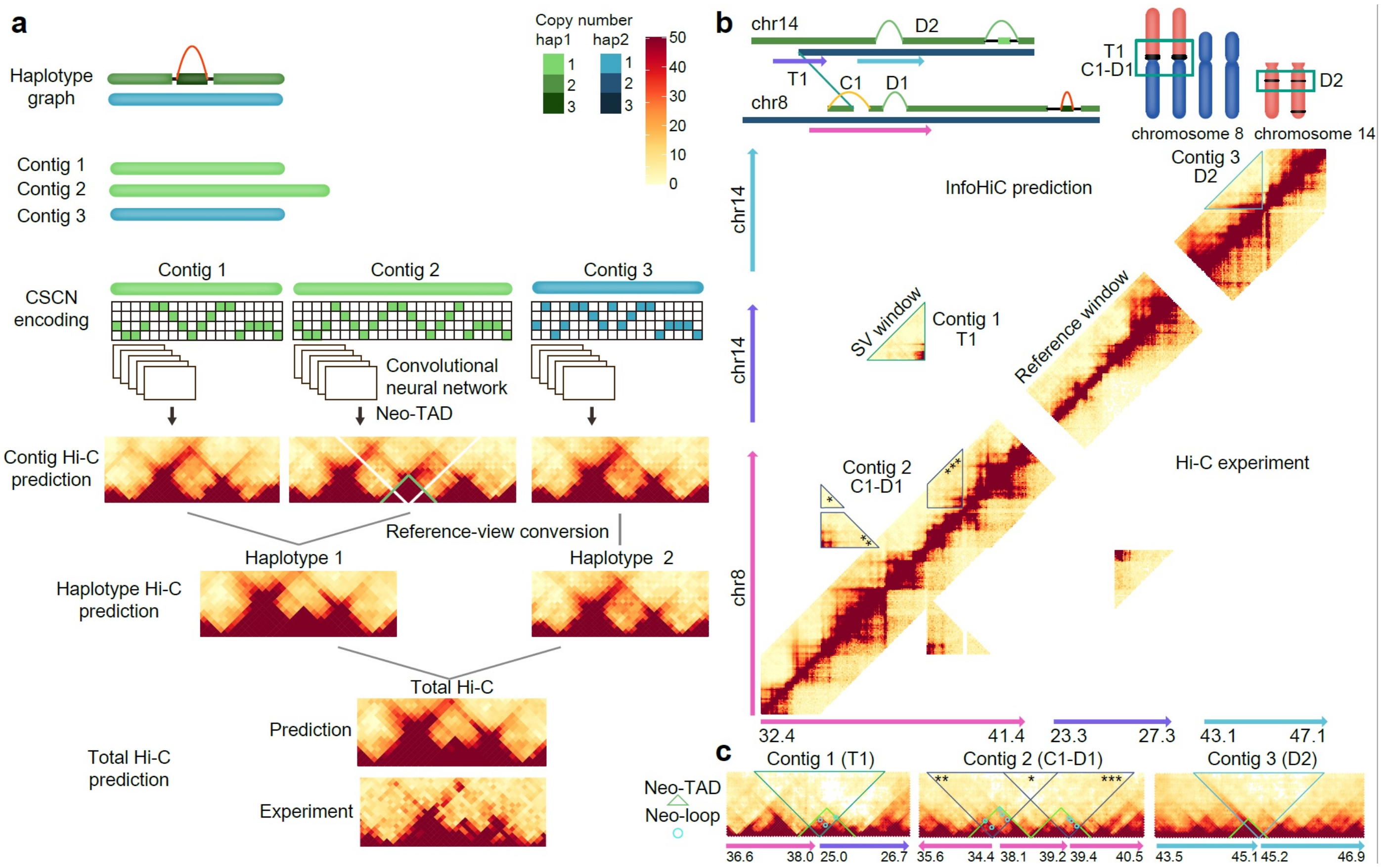
* genome: All genetic information of an organism. Except for the RNA of some viruses, all organisms have their genetic information composed of DNA, so it generally refers to genetic information composed of DNA.

* whole genome sequencing data: data that provides the base sequence of the entire DNA of an individual organism

* Hi-C data: data for analyzing the three-dimensional 3D structure and folding of DNA by measuring the relative spatial distance between two chromatins

* Neo-TAD (neo-Topologically Associating Domain): TAD is a concept that describes the topologically associated region where the genome is organized and operates in three dimensions in a cell. In Neo-TAD, the interaction between genes and regulators is changed due to modification of the existing TAD, which causes a new change in the gene expression pattern.

InfoHiC, developed by the research team, predicts Hi-C sequence data using whole-genome data of cancer cells rather than predefined human reference genome* sequences, unlike existing methodologies.

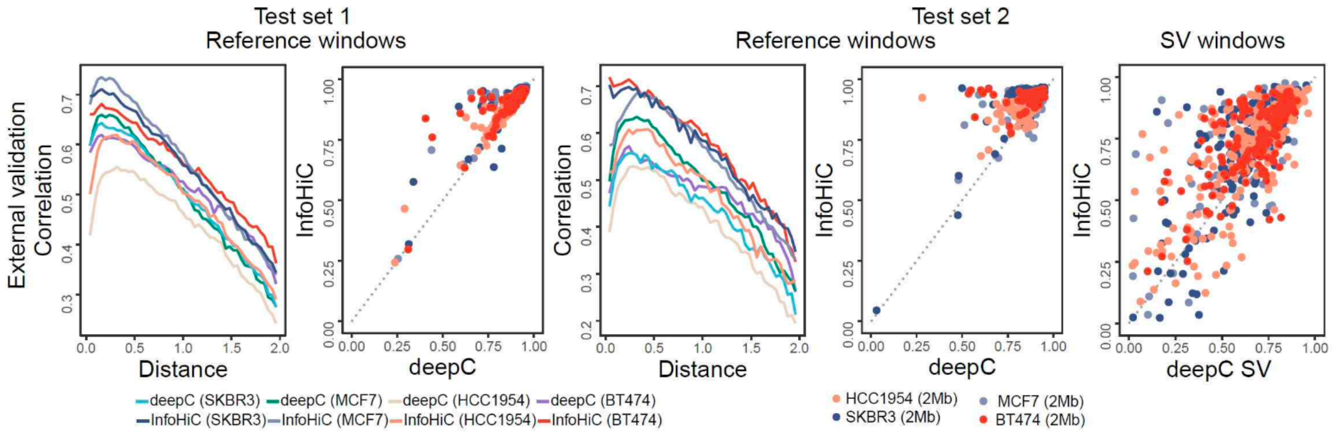


▲ Schematic diagram of InfoHiC, a cancer 3D genome prediction AI model. The model takes whole-genome data of cancer cells as input and predicts haplotype-specific HiC data.

Complex structural mutations* frequently occur in the chromosomes of cancer cells, and InfoHiC can predict neo-TADs caused by these complex structural mutations with higher accuracy.

* reference genome: a genetic map that serves as a guideline when reassembling DNA, a genome sequence representing a species of organism

* complex structural variation: mutations such as insertion, deletion, duplication, inversion, and translocation that change the structure of an individual's chromosome



▲ Accuracy of the developed prediction model (InfoHiC). The model showed improved performance compared to existing models in both areas without and with structural mutations.

Through this, the research team predicted the neo-TAD generation and enhancer hijacking* phenomenon caused by structural mutations in the non-coding DNA region, thereby enabling the accurate and low-cost identification of the impact of structural mutations in the non-coding DNA region on the occurrence and progression of cancer, as well as securing a technology that can directly observe it in cancer patients.

* enhancer hijacking: In normal cells, genes and enhancers that belong to different TADs belong to the same TAD due to neo-TADs, resulting in overexpression of genes due to the interaction between genes and enhancers.

When the research team applied InfoHiC to the whole genome data of patient A with medulloblastoma*, they predicted the enhancer hijacking phenomenon that causes abnormal gene expression, and through this, they were able to confirm the abnormality in gene expression regulation.

In addition, the research team used InfoHiC to identify gene expression abnormalities according to 3D genome mutations in patient B, who had difficulty selecting treatment target genes because mutations were not found in the coding DNA region of the tumor gene. It is expected that InfoHiC will contribute to recommending customized treatment for patients in the future.

* medulloblastma: A malignant brain tumor that mainly occurs in the cerebellum of children that occurs in the center of the cerebellum connected to the brainstem, but some occur in the cerebellar hemisphere outside the cerebellum.

The research team focused on the fact that complex structural mutations in cancer cells generate various haplotype contigs* and that neo-TADs are specifically formed according to these haplotypes, and predicted the 3D genome by reflecting this in the AI model.

In addition, the research team utilized InfoGenomeR, a genetic mutation discovery and genome restoration algorithm developed in a previous study, to construct haplotype contigs of the cancer genome.

In this way, the prediction results of Hi-C data that specifically correspond to each contig with different genetic variations were combined to finally predict the 3D genome. Hi-C data was encoded by inputting the base sequence and copy number variations of the contig, and then predicted through learning of the convolutional neural network (CNN) structure.

* haplotype: Generally refers to a set of genetic markers linked to a single chromosome

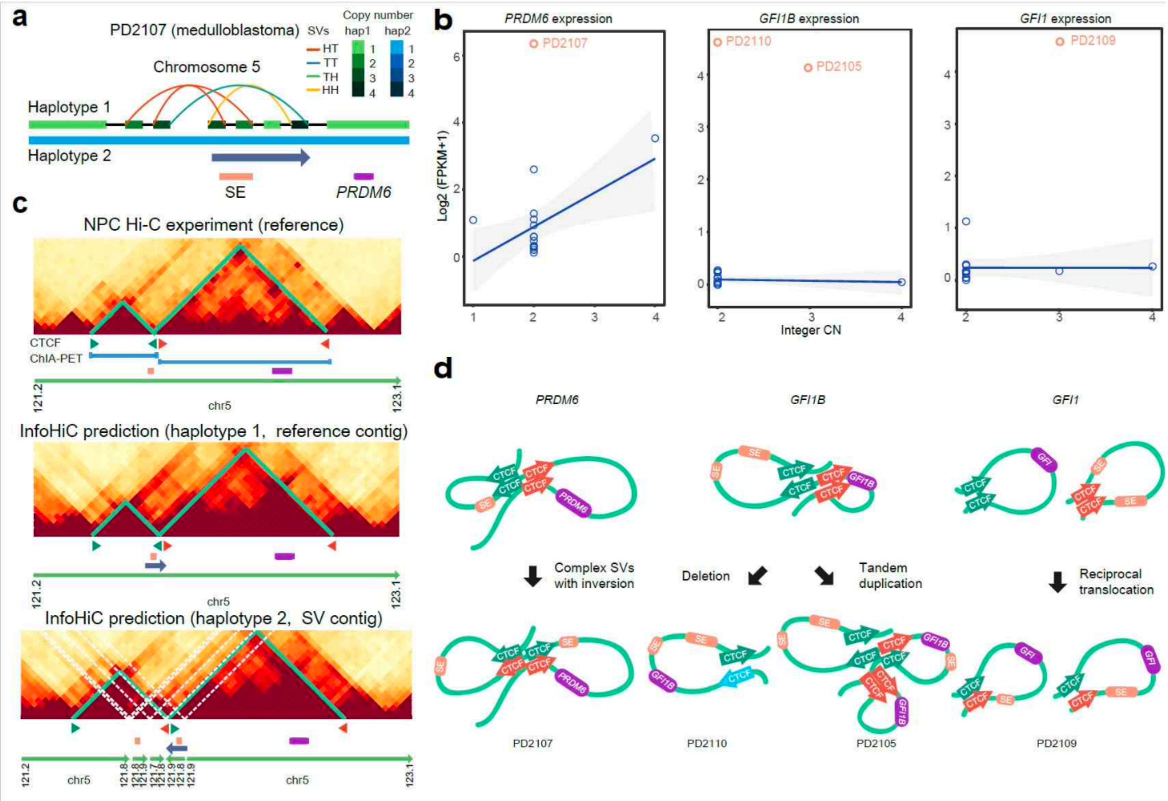
* contig: A new term that has emerged in the field of molecular biology as DNA sequencing methods have become widely used recently, meaning ‘a set of DNA fragments that overlap and are continuous’

Compared to models based on the existing human reference genome, InfoHiC developed by the research team significantly improved the 3D genome prediction performance of cancer cells with structural mutations.

As a result of verification using breast cancer cell lines, which are external data separate from the data used for model learning, the Pearson's R value* of the existing algorithm was 0.642, but InfoHiC improved it by 11% to 0.715.

More than 20% of the neo-TADs predicted in breast cancer cell lines were derived from complex structural mutations, which was concluded to be unpredictable by models based on the existing human reference genome.

* Pearson's R value: An indicator of the degree of correlation between the actual value and the predicted value.



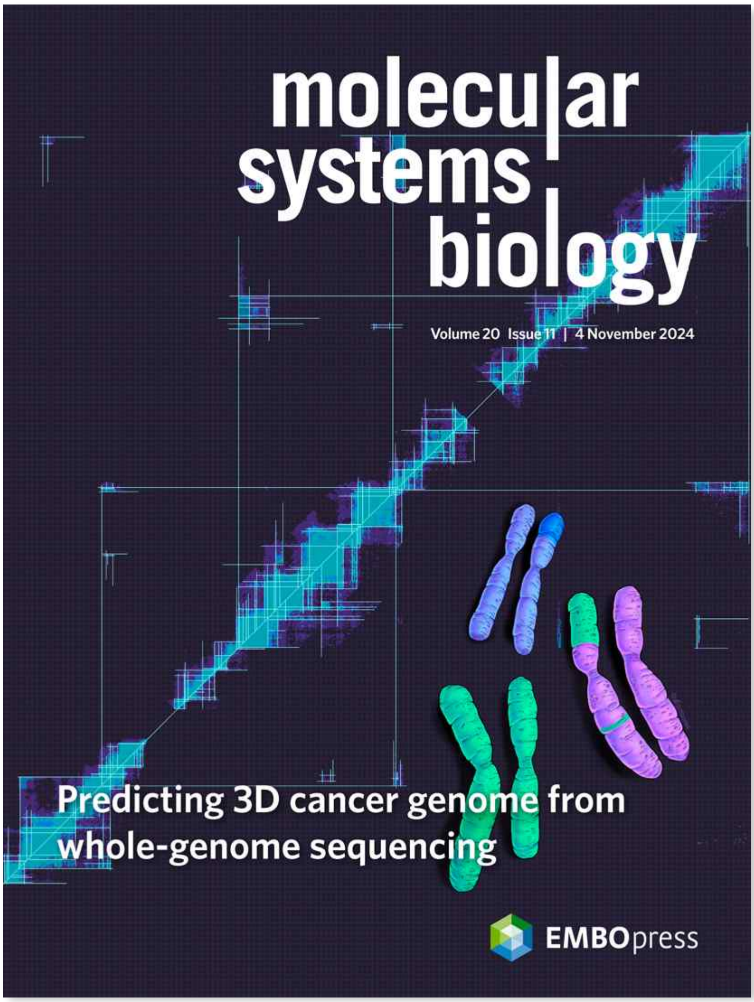
▲ Example of neoTAD generation prediction and enhancer hijacking when the developed prediction model (InfoHiC) is applied to patient data. It shows that PRDM6, GFI1B, and GFI1 genes were overexpressed by enhancer hijacking.

In addition, the research team applied InfoHiC to whole-genome data of 90 breast cancer patients to predict neo-TADs, and discovered neo-TAD-related genes that appeared repeatedly in multiple patients. It was also revealed that overexpression of these genes by enhancer hijacking was highly correlated with the survival rate of cancer patients.

Professor Hyunju Lee said, "Recently, with the decrease in the cost of sequencing data, whole-genome data of cancer patients are being produced in large quantities, but in contrast, Hi-C data that can confirm the 3D cancer genome is not easy to obtain due to the high cost. This study will contribute to personalized treatment of cancer patients with structural mutations in noncoding DNA regions through Hi-C data prediction."

This joint research by Professor Hyunju Lee of the GIST AI Graduate School and Professor Sung-Hye Park of Seoul National University College of Medicine's Department of Pathology was conducted by Dr. Yeonghun

Lee of the GIST School of Electrical Engineering and Computer Science and received support from the Institute of Information and Communications Technology Planning and Evaluation (IITP). The results of the research were published as the cover paper in the November 4, 2024, issue of Molecular Systems Biology, a top 10% international academic journal in the field of biochemistry and molecular biology.


molecular
systems
biology
Volume 20 Issue 11 | 4 November 2024
Predicting 3D cancer genome from
whole-genome sequencing
EMBOpress

▲ Published as the cover paper in 《Molecular Systems Biology》