# GIST-서울대병원, 비용 줄이고 정확도 높인 3차원 암 게놈 예측 AI 기술 개발

## 많은 비용 들어 확보하기 어려운 Hi-C 데이터 예측해 암 환자 개인별 유전자 발현 조절 이상 확인할 수 있어

- GIST 이현주 교수 - 서울대병원 박성혜 교수 공동연구팀, 암세포의 전장 유전체 정보 활용해 저비용으로 3차원 암 게놈 구조를 높은 정확도로 예측하는 AI 모델 InfoHiC 개발
- 기존 인간 참조 유전체 기반 모델 대비 예측 성능 크게 향상되고 수모세포종 환자 데이터에 적용해 검증... 국제학술지 <Molecular Systems Biology> 표지논문 게재

▲ (왼쪽부터) GIST AI대학원 이현주 교수, 서울대학교 의과대학 병리학교실 박성혜 교수, GIST 전기전자컴퓨터공학부 이영훈 박사

암의 발병 기전을 이해하기 위해 암세포의 유전체(게놈)에서 발생하는 돌연변이를 규명하려는 연구가 많이 시도되는 가운데, 최근에는 유전자에서 발생하는 점 돌연변이(point mutation)뿐 아니라 **암세포의 특이적 유전자 발현 조절 기전 규명의 중요성이 주목**받고 있다.

광주과학기술원(GIST, 총장 임기철)은 AI대학원 이현주 교수 연구팀이 서울대병원 박성혜 교수 연구팀과 함께 **암세포의 전장 유전체(한 사람의 전체 유전자) 정보를 활용하여 3차원 암(cancer) 게놈*을 예측하는 AI 모델, 'InfoHiC'를 개발**했다고 밝혔다.

암세포에서는 **3차원 게놈의 변화가 유전자 발현형의 조절에 중요한 역할**을 한다. Hi-C 데이터*를 사용하면 3차원 암 게놈의 neo-TAD* 구조를 확인할 수 있으나, 전장 유전체 데이터에 비해 **상대적으로 분석이 까다롭고 비용도 많이 든다**.

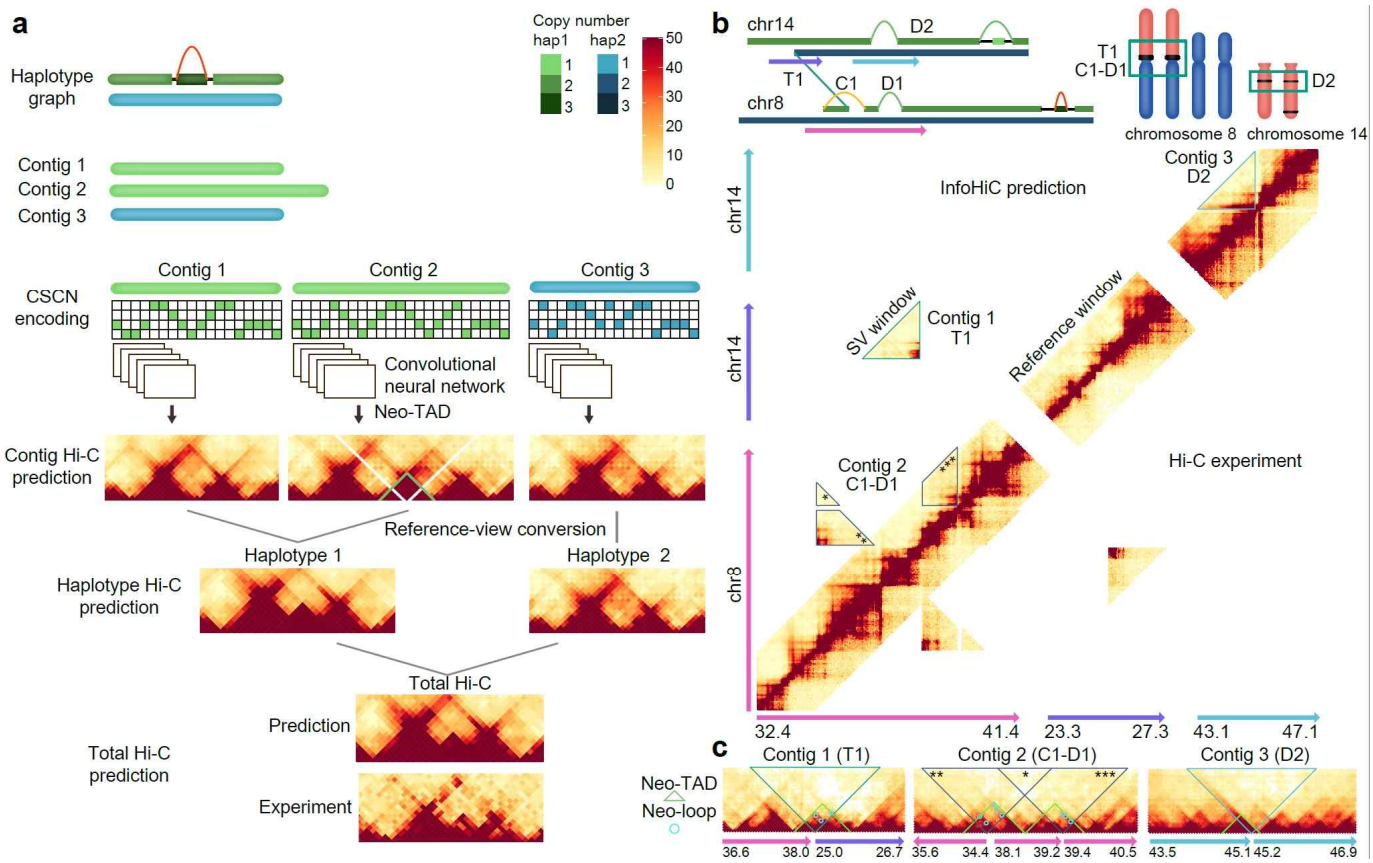
* **게놈(Genome):** 한 생물이 가지는 모든 유전 정보. 일부 바이러스의 RNA를 제외하고 모든 생물은 DNA로 유전 정보를 구성하고 있기 때문에 일반적으로 DNA로 구성된 유전 정보를 지칭함.

* **전장유전체 데이터(Whole Genome Sequencing data)**: 개별 개체의 전체 DNA의 염기 서열을 제공하는 데이터

* **Hi-C 데이터**: 두 염색질 사이의 공간상의 상대적 거리를 측정하여 DNA의 입체적 3차 구조와 접힘을 분석하기 위한 데이터

* **neo-TAD(neo-Topologically Associating Domain)**: TAD는 세포 속에서 유전체가 3차원적으로 구성되어 작동하는 위상학적으로 연관된 영역을 설명하는 개념임. Neo-TAD에서는 기존 TAD의 변형 때문에 유전자와 조절자 사이의 상호작용이 변경되며, 이로 인해 유전자 발현 패턴이 새롭게 변함.

연구팀이 개발한 InfoHiC는 기존의 방법론과 달리, 사전에 정의된 인간 참조 유전체* 서열이 아닌 **암세포의 전장 유전체 데이터를 사용하여 Hi-C 서열 데이터를 예측**한다.

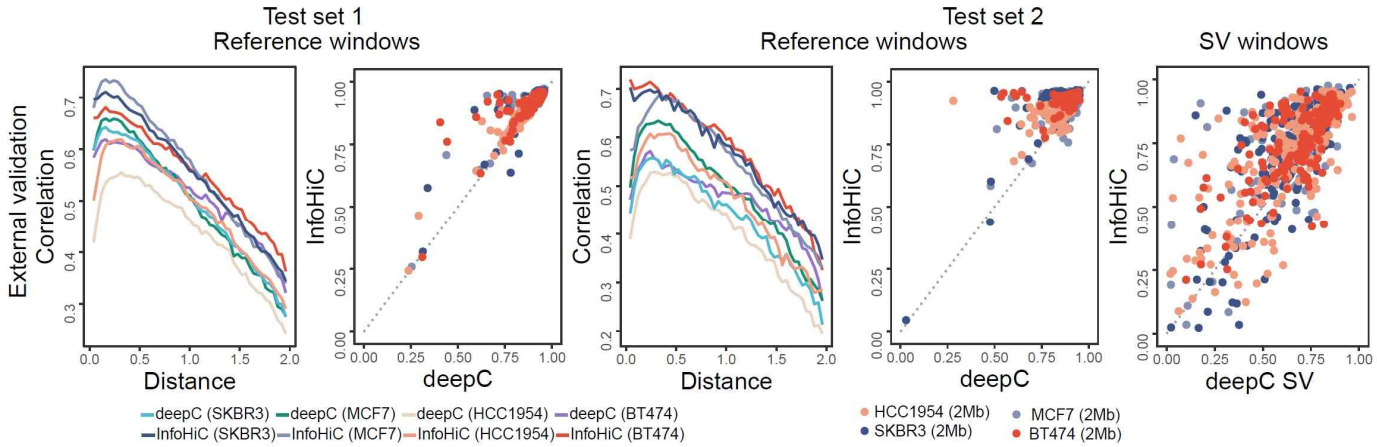



▲ **암 3D 게놈 예측 AI 모델인 InfoHiC의 개략도**. 해당 모델은 암세포의 전장 유전체 데이터를 입력으로 받아 하플로타입 특이적인 HiC 데이터를 예측한다.

암세포의 염색체에서는 복잡한 구조 변이*가 빈번하게 일어나는데, **InfoHiC는 이러한 복잡한 구조 변이에 의한 neo-TAD를 더 높은 정확도로 예측**할 수 있다.

* **참조 유전체(reference genome):** DNA를 재조립할 때 지침이 되는 유전자 지도, 생물의 종을 대표하는 유전체 서열

* **구조 변이(complex structural variation):** 개체 염색체의 구조를 변화시키는 삽입(insertion), 삭제(deletion), 중복(duplication), 역위(inversion), 전좌(translocation) 등의 변이




▲ **개발한 예측 모델 (InfoHiC)의 정확도**. 해당 모델은 구조 변이가 없는 영역과 있는 영역에서 모두 기존 모델보다 향상된 성능을 보였다.

연구팀은 이를 통해 비암호화 DNA(non-coding DNA) 영역에서 발생하는 **구조적 변이에 의한 neo-TAD 생성 및 인핸서 납치* 현상을 예측**함으로써, 비암호화 DNA 영역의 구조 변이가 암의 발생과 진행에 미치는 영향을 종전보다 **저비용으로 정확히 밝혀낼 수 있을 뿐만 아니라 암 환자에게서 직접 관찰할 수 있는 기술을 확보**하였다.

* **인핸서 납치(enhancer-hijacking):** 정상 세포에서는 서로 다른 TAD에 속해있던 유전자와 인핸서가 neo-TAD에 의해서 동일한 TAD에 속하게 됨으로써, 유전자와 인핸서의 상호작용에 의해서 유전자가 과발현 되는 현상임.

연구팀이 수모세포종* 환자 A의 전장 유전체 데이터에 InfoHiC를 적용한 결과, **비정상적인 유전자 발현을 유발하는 인핸서 납치 현상을 예측**하였고, 이를 통해 **유전자 발현 조절 이상을 확인**할 수 있었다.

또한 연구팀은 종양 유전자의 암호화 DNA(coding DNA) 영역에서 돌연변이가 발견되지 않아 치료 타깃 유전자 선정이 힘든 환자 B를 대상으로 InfoHiC를 활용하여 **3D 게놈 변이에 따른 유전자 발현 이상을 확인**하였는데, 이와 같은 방식으로 InfoHiC가 **추후 환자 맞춤형 치료 추천에 기여**할 것으로 기대된다.

* **수모세포종(medulloblastma):** 소아의 소뇌 부위에 주로 발생하는 악성 뇌종양으로 뇌간과 연결된 소뇌 중심부에서 발생하며 일부는 소뇌 바깥쪽 소뇌반구라는 부위에서 발생

연구팀은 암세포의 복잡한 구조 변이가 다양한 하플로타입(haplotype) 콘티그(contig)*를 생성하고, neo-TAD가 바로 이러한 **하플로타입에 따라 특이적으로 형성된다는 사실에 주목하여 이를 AI 모델에 반영함으로써 3차원 게놈을 예측**하였다.

또한 연구팀은 앞선 연구에서 개발한 바 있는 **유전 변이 발굴 및 유전체 복원 알고리즘인 인포지노머(InfoGenomeR)를 활용하여 암 유전체의 하플로타입 콘티그(contig)를 구성**하였다.

이로써 유전적 변이가 다른 각 **콘티그에 특이적으로 대응하는 Hi-C 데이터의 예측** 결과를 결합하여 최종적으로 3차원 게놈을 예측하였다. Hi-C 데이터는 콘티그의 염기서열 및 복제 수 변이를 입력하여 부호화(encoding)한 후, 합성곱 신경망(convolutional neural network, CNN) 구조의 학습을 통해 예측되었다.

* **하플로타입(haplotype):** 일반적으로 하나의 염색체(chromosome)에 연결되어 있는 유전적 마커(genetic marker)들의 세트(set)를 의미

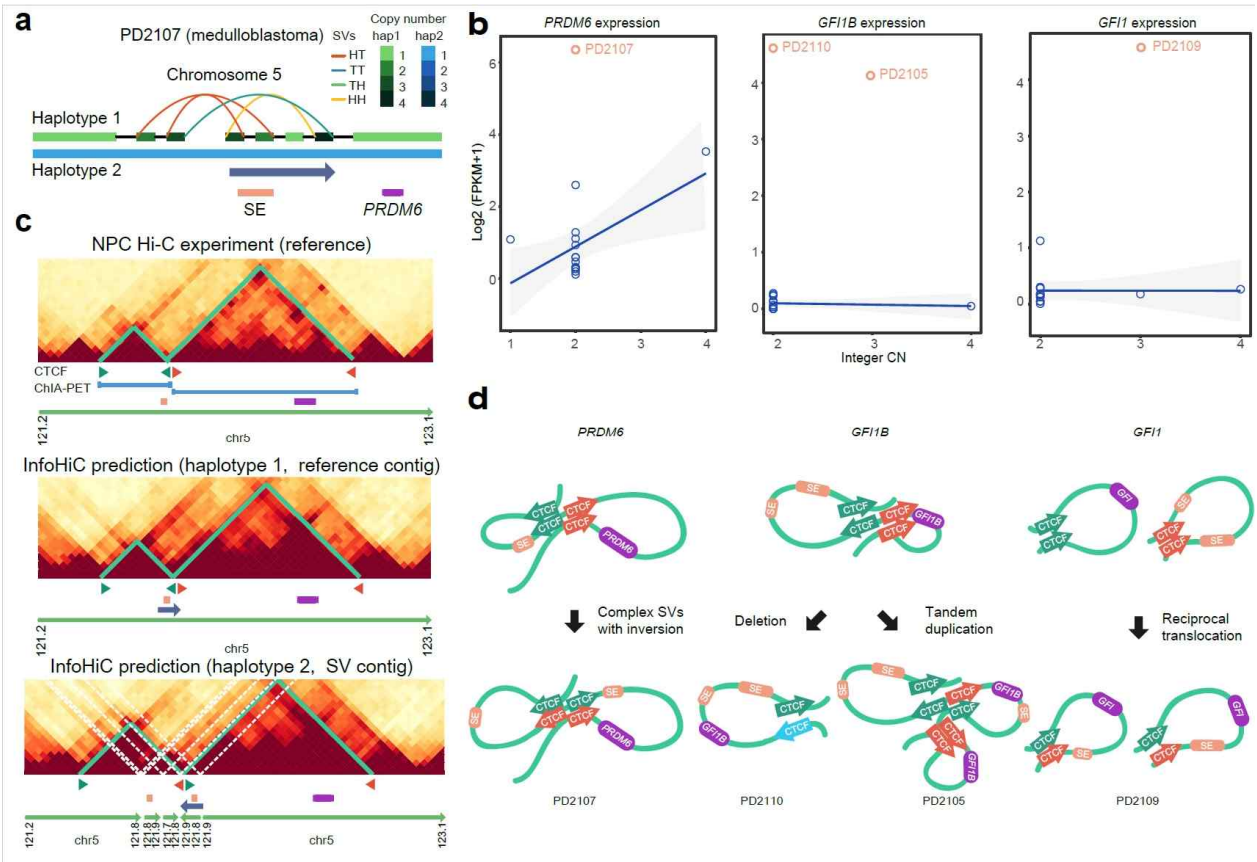
* **콘티그(contig):** 최근 DNA 시퀀싱 방법이 널리 사용되면서 분자생물학계에서 새롭게 등장한 용어로서 '서로 겹치면서 연속되어 있는 DNA 절편들의 집합'을 의미

**기존의 인간 참조 유전체에 기반한 모델과 비교하여** 연구팀이 개발한 InfoHiC는 **구조 변이가 있는 암세포의 3D 게놈 예측 성능이 크게 향상**되었다.

모델 학습에 사용된 데이터와는 별개의 외부 데이터인 유방암 세포주를 활용하여 검증한 결과, 기존 알고리즘의 Pearson's R 값*은 0.642이었으나, **InfoHiC는 0.715로 11% 향상**되었다.

유방암 세포주에서 예측한 neo-TAD 중 20% 이상이 복잡한 구조 변이에서 비롯된 것으로, 이것은 기존의 인간 참조 유전체에 기반한 모델에서는 예측할 수 없다는 결론을 내렸다.

*** Pearson's R 값:** 실제값과 예측값 사이의 상관관계의 정도를 나타내는 지표

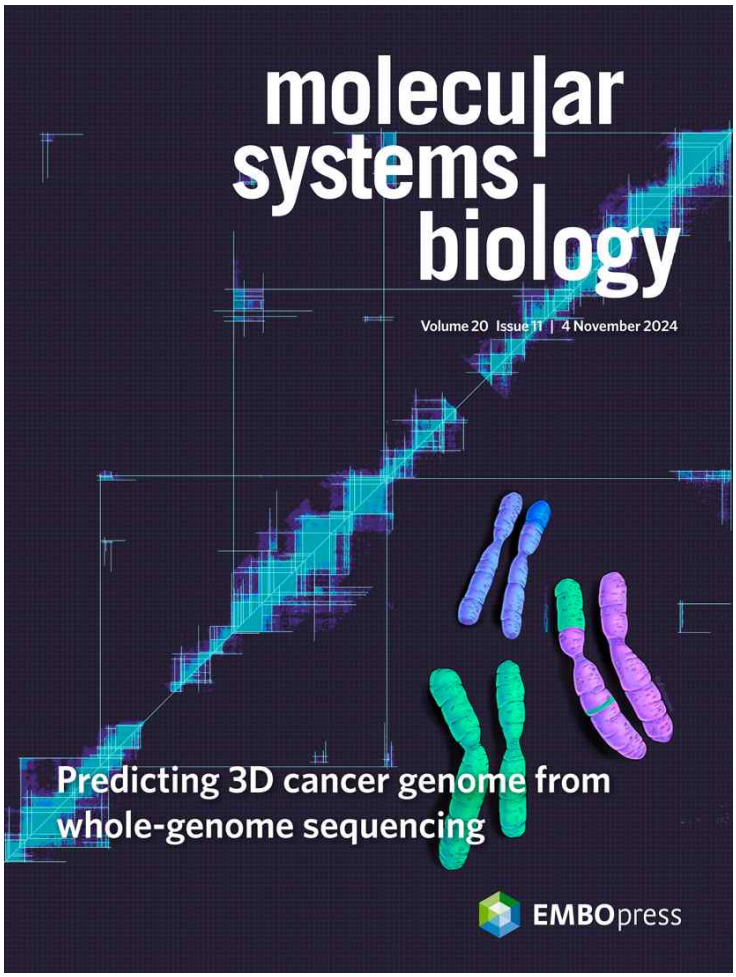



**▲ 개발한 예측 모델(InfoHiC)을 환자 데이터에 적용한 경우의 neoTAD 생성 예측 및 인핸서 납치의 예시**. 인핸서 납치에 의해서 PRDM6와 GFI1B, GFI1의 유전자가 과발현되었음을 보여준다.

또한 연구팀이 InfoHiC를 유방암 환자 90명의 전장 유전체 데이터에 적용하여 neo-TAD를 예측한 결과, **여러 환자들에게서 반복적으로 나타나는 neo-TAD 관련 유전자를 발견**했는데 **인핸서 납치에 의한 이들 유전자의 과발현이 암환자의 생존율과 연관이 높다**는 점도 밝혀졌다.

이현주 교수는 "최근에 시퀀싱 데이터 비용의 감소로 암 환자의 전장유전체 데이터는 많이 생산되고 있으나, 이에 반해 **3차원 암 게놈을 확인할 수 있는 Hi-C 데이터는 고비용 탓에 확보가 쉽지 않다**" 면서 "이번 연구는 Hi-C 데이터 예측을 통해서 **비암호화 DNA 영역에서의 구조 변이를 가진 암 환자의 개인 맞춤형 치료에 기여**할 수 있을 것"이라고 말했다.

GIST AI대학원 이현주 교수와 서울대학교 의과대학 병리학교실 박성혜 교수의 이번 공동연구는 GIST 전기전자컴퓨터공학부 이영훈 박사가 수행하였으며, 정보통신기획평가원(IITP)의 지원을 받았다. 연구 결과는 생화학 및 분자생물학 분야 상위 10% 국제학술지 《몰레큘러 시스템즈 바이올로지(Molecular Systems Biology)》에 2024년 11월 4일 표지논문으로 게재됐다.




▲《Molecular Systems Biology》에 표지논문으로 게재

# 논문의 주요 정보

**1. 논문명, 저자정보**

- 저널명: Molecular Systems Biology (IF : 8.5, 2023년 기준)
- 논문명 : Prediction of the 3D cancer genome from whole-genome sequencing using InfoHiC
- 저자 정보 : 이영훈 (제1저자, GIST 전기전자컴퓨터공학부 박사), 박성혜 교수 (공저자, 서울대학교 의과대학, 병리학교실), 이현주 교수 (교신저자, GIST AI대학원, 전기전자컴퓨터공학부)