

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크숍 (온라인)



Cancer multi-omics data analysis based on AI

홍동완 _ 가톨릭대학교



KSBI
KOREAN SOCIETY FOR
BIOINFORMATICS

| 한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBI-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크숍인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크숍은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크숍은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의를 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의를 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구 중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의를 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

Cancer multi-omics data analysis based on AI

암 멀티 오믹스 데이터는 전 세계적으로 지속적이고 엄청난 양으로 증가하고 있으나 분석 및 분석 결과의 해석을 최적화하기에는 구조가 상당히 이질적이고 복잡하며, 데이터의 크기 또한 방대하여 상당한 어려움을 겪고 있다.

본 강의에서는 암 멀티 오믹스 데이터를 효과적이고 빠르게 처리하기 위해 인공지능 기법을 활용하여 분석/해석할 수 있는 방법을 강의한다. 암 멀티 오믹스로부터 생성된 돌연변이 및 유전자 발현 등 데이터 특징에 따른 인공지능 활용 방법에 대해 이해를 한다. 특히, AlphaFold2를 활용하여 3차원 단백질 구조를 예측할 수 있는 응용 사례를 보이고, 3차원 단백질 예측 구조에 대해 특화된 자동화 평가 방법을 제안한다. 수강생의 인공지능 활용 범위를 고도화하며 인공지능 빅 데이터 시대에 필요한 생물정보학자의 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

- Multi-omics 데이터를 활용한 인공지능
- 암 유전자 돌연변이/유전자 발현/단백질 발현 데이터를 활용한 인공지능 응용
- Alphafold2를 이용한 단백질 3차원 구조 예측
- 구조 예측 결과의 평가 및 해석

* 참고강의교재: 담당교수 강의 노트

* 교육생준비물: 노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상, 온라인 접속)

* 강의 난이도: 초급

* 강의: 홍동완 교수 (가톨릭대학교 의과대학)

Curriculum Vitae

Speaker Name: Dongwan Hong, Ph.D.



► Personal Info

Name Dongwan Hong
Title Professor
Affiliation Catholic University of Korea, College of Medicine

► Contact Information

Address 222 Banpodae-ro, Soecho-gu, Seoul 06591,
Republic of Korea
Email dwhong@catholic.ac.kr
Phone Number +82-2-3147-8424

Research Interest

Cancer genomics, Big data, and Artificial Intelligence

Educational Experience

2002-2007 Ph.D. Dept. of Computer Engineering, Hallym University
1992-1996 B.S. Dept. of Computer Science Hallym University

Professional Experience

2020-present Professor, Catholic University of Korea, College of Medicine
2011-2022 Chief Researcher, National Cancer Center of Korea
2008-2011 Senior Researcher, Seoul National University, Medical Research Institute
2003-2007 Assistant Professor, Dept. of Multimedia, Songgok University

Selected Publications (5 maximum)

1. Park et. al., Clonal dynamics in early human embryogenesis inferred from somatic mutations, 597, 393-397, Nature, 2021.
2. Kim et. al., FIREVAT: finding reliable variants without artifacts in human cancer samples using etiologically relevant, Genome Medicine, 11(1):81, 2019.
3. Park et. al., Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma, CELL, 177, 1842-1857, 2019
4. Yang et. al., RhoGAP domain-containing fusions and PPAPDC1A fusions are recurrent and prognostic in diffuse gastric cancer, NATURE COMMUNICATIONS, 9(1):4439~4439, 2018.
5. Lee et. al., Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures, Nucleic Acids Research, 46(W1):W102~W108, 2018

KSBi-BIML 2023 (Cancer multi-omics data analysis based on AI)

Prof. Dongwan Hong, Ph.D.

(dwhong@catholic.ac.kr, <http://honglab.catholic.ac.kr>)

Catholic University of Korea, College of Medicine

본 강의 자료는 한국생명정보학회가 주관하는 KSBi-BIML 2023 워크숍 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다. 수업 목적으로 배포 및 전송 받은 경우에도 이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없습니다.

만약 이러한 사항을 위반할 경우 발생하는 모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고합니다.

Overview

- **Part I: Cancer Genome Analysis**
 - Mutational Signature Analysis: Mutalisk
- **Part II: Artificial Intelligence Study using Cancer Big Data**
 - Use AlphaFold

The History of Cancer Genomics

Vol 458 | 9 April 2009 | doi:10.1038/nature07943

nature

REVIEWS

The cancer genome

Michael R. Stratton^{1,2}, Peter J. Campbell^{1,2}

Integrated genomic analyses of ovarian carcinoma
Nature, 2011;474(7353):609-615. doi:10.1038/nature10551

2012

Comprehensive molecular characterization of human colon and rectal cancer
Nature, 2012;487(7407):330-337. doi:10.1038/nature11471

Comprehensive molecular portraits of human breast tumours
Nature, 2012;490(7418):61-70. doi:10.1038/nature11249

2013

Integrated genomic characterization of endometrial cancer
Nature, 2013;497(7447):67-73. doi:10.1038/nature12117

Genomic and epigenomic landscapes of adult de novo glioblastoma
N Engl J Med, 2013;368(22):2059-2074. doi:10.1056/NEJMoa1302442

Comprehensive molecular characterization of clear cell renal cell carcinoma
Nature, 2013;499(7456):43-49. doi:10.1038/nature12385

The Cancer Genome Atlas Pan-Cancer analysis project
Not Genet, 2013;45(10):1113-1120. doi:10.1038/ng.277

The somatic genomic landscape of glioblastoma
Cell, 2013;155(2):462-477. doi:10.1016/j.cell.2013.09.034

2014

Comprehensive molecular characterization of urothelial bladder carcinoma
Nature, 2014;507(7492):315-322. doi:10.1038/nature12965

Cell

Perspective on Oncogenic Processes at the Beginning of Cancer Genomics

Graphical Abstract

Substrates: DNA, DNA, RNA & Protein, Cellular

Associations: Germine & somatic, Driver & molecular subtypes, Immune cell & tumor

Correspondence: lding@wustl.edu (L.D.), wheeler@bcm.edu (D.A.W.), gadget@broadinstitute.org (G)

In Brief: A synthesized view on oncogenic processes based on PanCancer analyses highlights the complex of genome alterations on the site and multi-omic profiles of humans as well as their influence on the microenvironment.

Authors: Li Ding, Matthew H. Bailey, Eduard Porta-Pardo, ..., David J. Gad Getz, The Cancer Genome Research Network

~2,700 PanCan WGS

Article | Open Access | Published: 05 February 2020

Pan-cancer analysis of whole genomes

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

Nature 578, 82-93(2020) | Cite this article

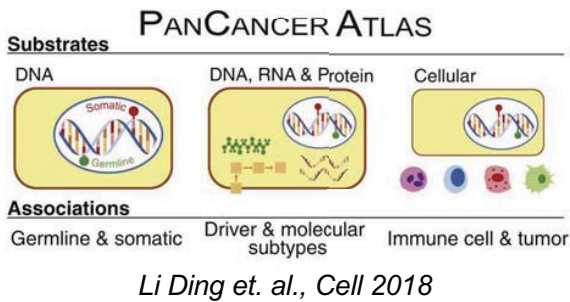
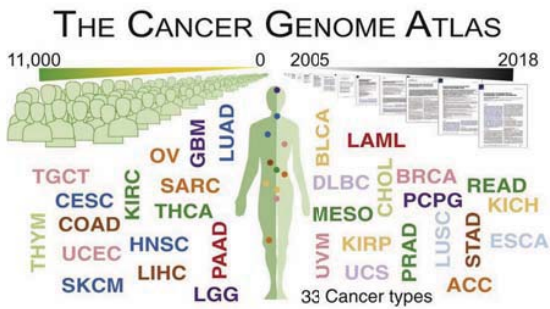
262k Accesses | 169 Citations | 1189 Altmetric | Metrics

Ding et al., Cell 2018

TCGA Research Network Publication

Nature, 2020

Cancer Big Project



F-22 Raptor stealth fighters

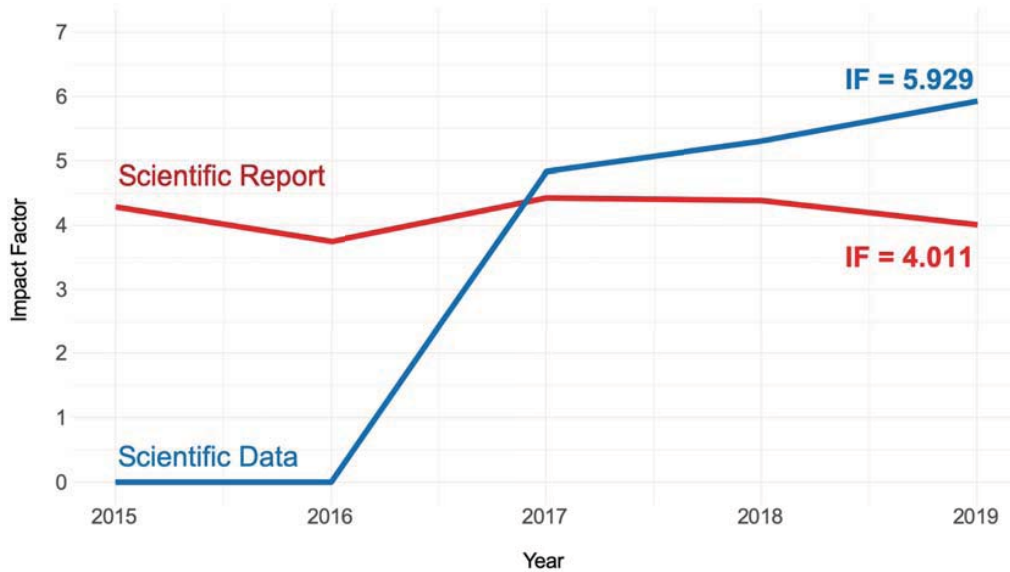


- TCGA project was finished with advancements in cancer genomics
 - Systemized large-scale genomics-based cancer research (33 types / 11,000 tumors)

HISTORY OF CANCER TREATMENT MODALITIES

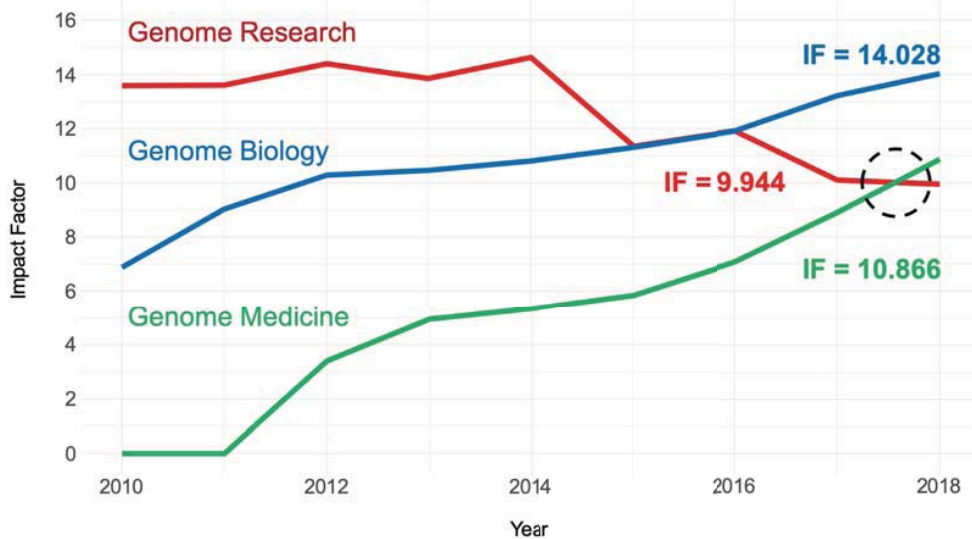
	SURGERY	RADIATION	CHEMO-THERAPY	TARGETED DRUGS	IMMUNO-THERAPY
APPROACH	Cut out accessible tumor cells to stop growth and prevent their spread	Use highly concentrated X-rays or radioactive isotopes to kill cancerous cells	Use cytotoxic drugs to kill or inhibit cancer cells	Interfere with a mechanism required for, or that supports tumor growth	Support the immune system's innate ability to recognize and eliminate tumor cells
SINCE	1800s	early 1900s	late 1940s	2000s	2010s
LIMITATIONS	Many inaccessible tumors ineligible; limited effectiveness if tumor has already begun to spread	Limited effectiveness if tumor has already begun to spread; potentially dangerous for tumors near vital organs	High toxicity and often does not destroy the whole tumor, leading to high rates of recurrence	Limited tumor types eligible; high efficiency but short durability driving high rates of recurrence	Applicable to all tumors at all stages of disease including metastatic tumors; responses are highly durable; potential for lower toxicity profiles; synergistic with other treatments

Trend 1. Data are More Impactful Than Individual Reports



Report < Data

Trend 2. Medicine Outweighs Research



Research < Medicine

Clinical Genomic Big Data

500 PB (2012) -> 25,000 PB (2020)

Exa -> Zeta -> Yotta

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

TCGA produced over
2.5
PETABYTES
of data

TCGA data describes
33
DIFFERENT
TUMOR TYPES

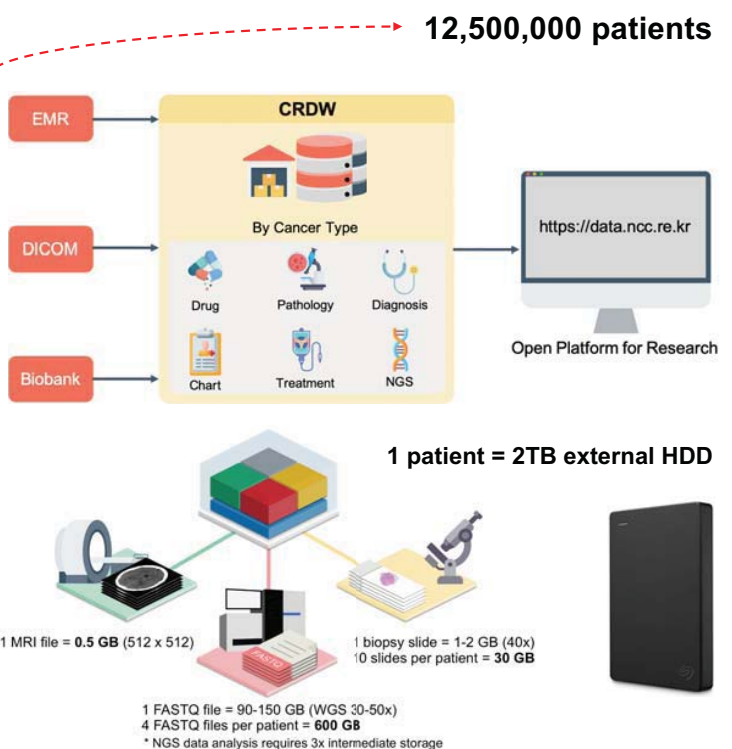
...including
10
RARE
CANCERS

To put this into perspective, 1 petabyte of data is equal to
212,000
DVDs

...based on paired tumor and normal tissue sets collected from
11,000
PATIENTS

...using
7
DIFFERENT
DATA TYPES

TCGA Research Network



ICGC Mutational Signatures Working Group

With which methods?



- For SBS, DBS and indel signatures, should we:
 - stick to the original ICGC PCAWG pipelines **SigProfiler (Sanger)** and **SigAnalyzer (Broad)**
 - test other methods (>25 **published tools** including EMu, SomaticSignatures, Palimpsest, **Mutalisk**, BayesNMF...). For review: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221235>
 - Choose one approach or combine several?
- For SV signatures:
 - not sure the PCAWG method can be easily reproduced (**ClusterSV** available to generate footprints but extracting signatures from that is not obvious)
 - alternatives include **Palimpsest** or **Signal**
- For signatures of repair deficiencies:
 - Mismatch repair deficiency → **MSIsensor**, **MSIseq**, **MOSAIC...**
 - Homologous recombination defects → **Signature 3**, **HRDetect**, **SigMA**, combined **HRD scores...**

Participants: Alvin Ng (**Sanger**), Dongwan Hong, Paz Polak (**Broad;NY**), Eric Letouzé (Paris genome center)



a web-based somatic MUTation ANalySIS toolKit for genomic, transcriptional and epigenomic signatures

Nucleic Acids Research (2018)

COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Genome Version ▾ Search COSMIC...

Mutational Signatures (v3.1 - June 2020)

Introduction

*Somatic mutations are present in all cells of the human body and occur throughout life. They are the consequence of multiple mutational processes, including the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA and defective DNA repair. Different mutational processes generate unique combinations of mutation types, termed "**Mutational Signatures**".*

In the past few years, large-scale analyses have revealed many mutational signatures across the spectrum of human cancer types, including the latest effort by the ICGC/TCGA [Pan-Cancer Analysis of Whole Genomes \(PCAWG\)](#) Network (Alexandrov, L.B. et al., 2020) using data from more than 23,000 cancer patients.



<https://cancer.sanger.ac.uk/cosmic/signatures>

Signature-based websites

As the number of mutational signatures and variant classes considered has increased, the need for a curated census of signatures has become apparent. Here, we deliver such a resource by providing a comprehensive overview of the key information known, suspected or widely discussed in the scientific literature for each of the identified mutational signatures on a dedicated website.

This summary includes the mutational profile, proposed aetiology and tissue distribution of each signature, as well as potential associations with other mutational signatures and how the signature has changed during iterations of analysis.

Currently, three different variant classes are considered, resulting in the following sets of mutational signatures.

Single Base Substitution (SBS)
Signatures

Doublet Base Substitution (DBS)
Signatures

Small Insertion and Deletion (ID)
Signatures

Versions

Mutational signatures version 3 was released as part of COSMIC release v89 (May 2019) and updated to version 3.1 in COSMIC release v91 (June 2020). The version 3.1 update expands and improves upon the version 2 signatures (March 2015) that were part of earlier COSMIC releases and can still be consulted.

Bioinformatic tools

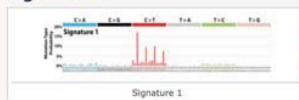
The current set of mutational signatures has been extracted using SigProfiler, a compilation of publicly available bioinformatic tools addressing all the steps needed for signature identification. SigProfiler functionalities include mutation matrix generation from raw data and signature extraction, among others.

Mutational Signatures
Version 2

SigProfiler
Bioinformatic Tools

Mutational Signature v2

Signature 1



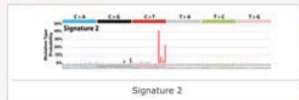
Cancer types: Signature 1 has been found in all cancer types and in most cancer samples.

Proposed aetiology: Signature 1 is the result of an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine.

Additional mutational features: Signature 1 is associated with small numbers of small insertions and deletions in most tissue types.

Comments: The number of Signature 1 mutations correlates with age of cancer diagnosis.

Signature 2



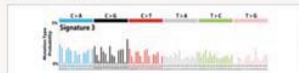
Cancer types: Signature 2 has been found in 22 cancer types, but most commonly in cervical and bladder cancers. In most of these 22 cancer types, Signature 2 is present in at least 10% of samples.

Proposed aetiology: Signature 2 has been attributed to activity of the AID/APOBEC family of cytidine deaminases. On the basis of similarities in the sequence context of cytosine mutations caused by APOBEC enzymes in experimental systems, a role for APOBEC1, APOBEC3A and/or APOBEC3B in human cancer appears more likely than for other members of the family.

Additional mutational features: Transcriptional strand bias of mutations has been observed in exons, but is not present or is weaker in introns.

Comments: Signature 2 is usually found in the same samples as Signature 13. It has been proposed that activation of AID/APOBEC cytidine deaminases is due to viral infection, retrotransposon jumping or to tissue inflammation. Currently, there is limited evidence to support these hypotheses. A germline deletion polymorphism involving APOBEC3A and APOBEC3B is associated with the presence of large numbers of Signature 2 and 13 mutations and with predisposition to breast cancer. Mutations of similar patterns to Signatures 2 and 13 are commonly found in the phenomenon of local hypermutation present in some cancers, known as kataegis, potentially implicating AID/APOBEC enzymes in this process as well.

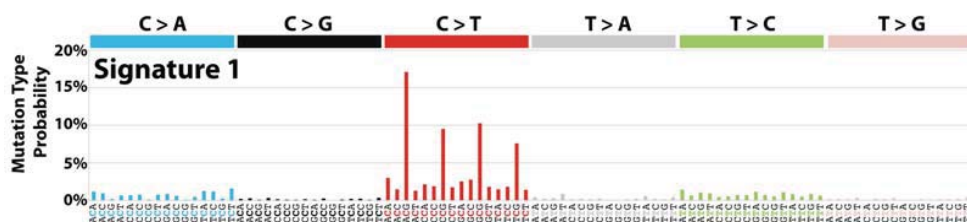
Signature 3



Cancer types: Signature 3 has been found in breast, ovarian, and pancreatic cancers.

Proposed aetiology: Signature 3 is associated with failure of DNA double-strand break-repair by homologous recombination.

Additional mutational features: Signature 3 associates strongly with elevated numbers of large (longer than 3bp) insertions and deletions with overlapping microhomology at breakpoint junctions.





Mutational Signature (v3 – May 2019)

- Single Base Substitution (SBS): 65 signatures (18 possible sequencing artifacts)
- Double Base Substitution (DBS): 11 signatures
- Small Insertion and Deletion Substitution (ID): 17 signatures



13

Mutational Signature v3.1

Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Genome Version ▾
 SEARCH
Login

Mutational Signatures (v3.1 - June 2020)

Mutational Signatures Home
Single Base Substitution (SBS) Signatures
Doublet Base Substitution (DBS) Signatures
Small Insertion and Deletion (ID) Signatures
Mutational Signatures Version 2
SigProfiler Bioinformatic Tools

Single Base Substitution (SBS) Signatures

Single base substitutions (SBS), also known as single nucleotide variants, are defined as a replacement of a certain nucleotide base. Considering the pyrimidines of the Watson-Crick base pairs, there are only six different possible substitutions: C>A, C>G, C>T, T>A, T>C, and T>G. These SBS classes can be further expanded considering the nucleotide context.

Current SBS signatures have been identified using 96 different contexts, considering not only the mutated base, but also the bases immediately 5' and 3'.


Click on any signature below to learn more about its details.

Signature extraction methods

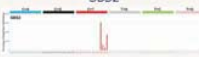
With a few exceptions, the signatures were extracted using SigProfiler (as described in [Alexandrov, L.B. et al., 2020](#)) from the 2,780 whole-genome variant calls produced by the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Network. The stability and reproducibility of the signatures were assessed on somatic mutations from an additional 1,865 whole genomes and 19,184 exomes. All input data and references for original sources are available from [synapse.org ID syn11801859](#).

The COSMIC v3 signatures are available in numerical form in [syn12009743](#), and attributions of the signatures to mutations in tumors are available in [syn11804040](#) and [syn11804058](#). The COSMIC v3.1 signatures can be downloaded [here](#).


SBS1




SBS2




SBS3




SBS4




SBS5



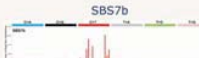
SBS6




SBS7a




SBS7b




SBS7c




SBS7d




SBS8




SBS9




SBS10a



SBS10b



SBS11

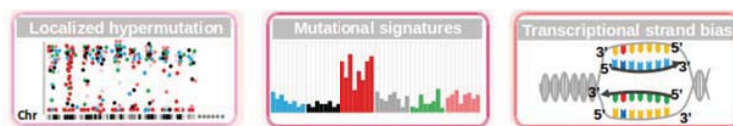


If there are 3,000 mutations
we'd expect a random dispersion of mutations (approx. 1 mutation/Mb)

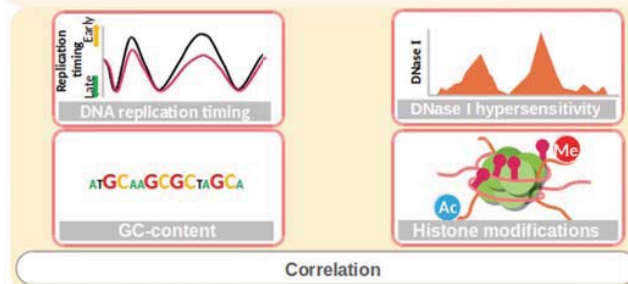


Human genome \approx 3,000-megabase (Mb)

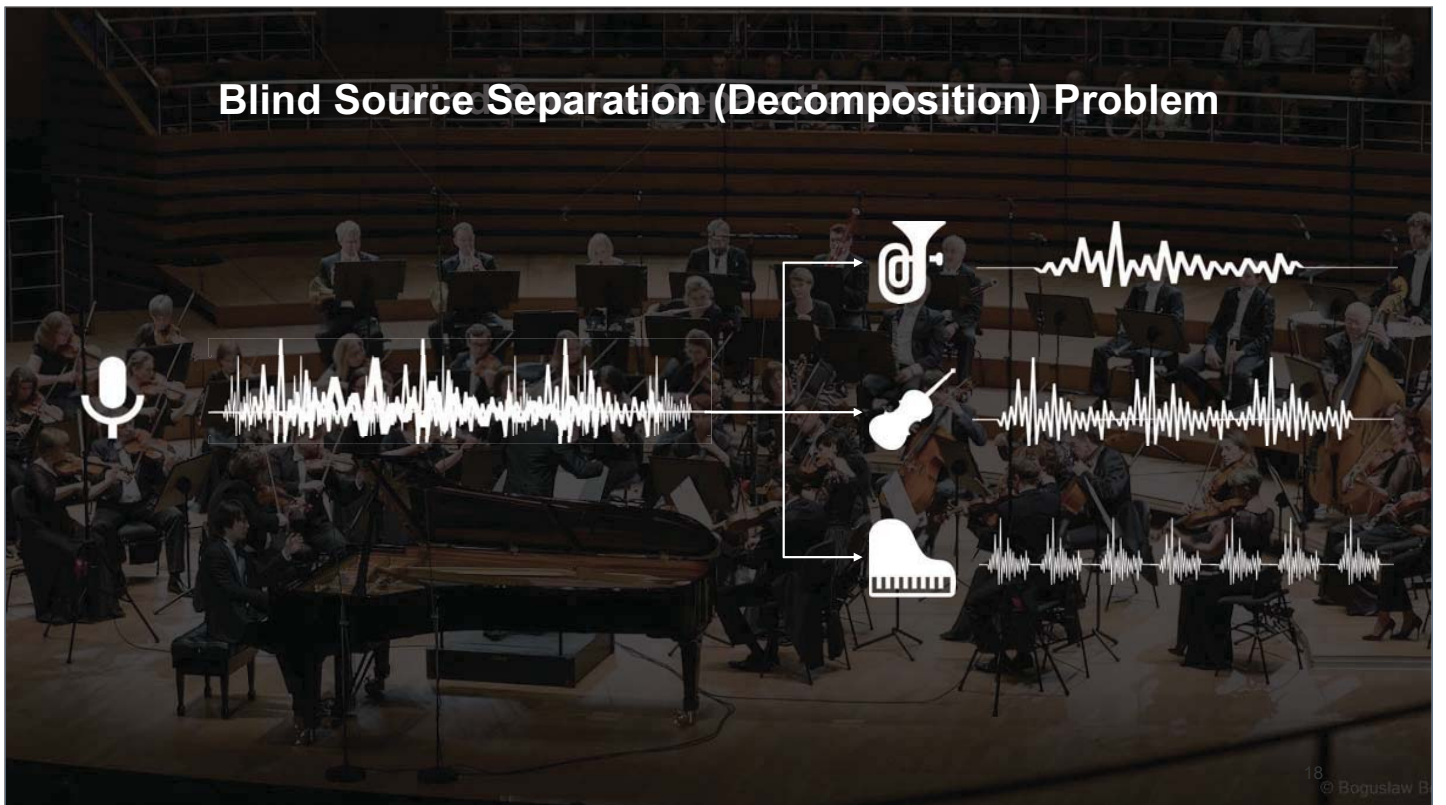
15



However, in reality, we observe uneven distribution of mutations
in genome due to favored contexts



16



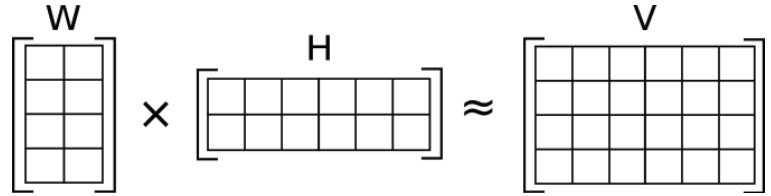


WIKIPEDIA
The Free Encyclopedia

Article Talk

Non-negative matrix factorization

From Wikipedia, the free encyclopedia



History [\[edit\]](#)

In *chemometrics* non-negative matrix factorization has a long history under the name "self modeling curve resolution".^[10] In this framework the vectors in the right matrix are continuous curves rather than discrete vectors. Also early work on non-negative matrix factorizations was performed by a Finnish group of researchers in the 1990s under the name *positive matrix factorization*.^{[11][12][13]} It became more widely known as *non-negative matrix factorization* after Lee and Seung investigated the properties of the algorithm and published some simple and useful algorithms for two types of factorizations.^{[14][15]}

Background [\[edit\]](#)

Let matrix **V** be the product of the matrices **W** and **H**,

$$\mathbf{V} = \mathbf{WH}.$$

Matrix multiplication can be implemented as computing the column vectors of **V** as linear combinations of the column vectors in **W** using coefficients supplied by columns of **H**. That is, each column of **V** can be computed as follows:

$$\mathbf{v}_i = \mathbf{W}\mathbf{h}_i,$$

where \mathbf{v}_i is the i -th column vector of the product matrix **V** and \mathbf{h}_i is the i -th column vector of the matrix **H**.

When multiplying matrices, the dimensions of the factor matrices may be significantly lower than those of the product matrix and it is this property that forms the basis of NMF. NMF generates factors with significantly reduced dimensions compared to the original matrix. For example, if **V** is an $m \times n$ matrix, **W** is an $m \times p$ matrix, and **H** is a $p \times n$ matrix then p can be significantly less than both m and n .

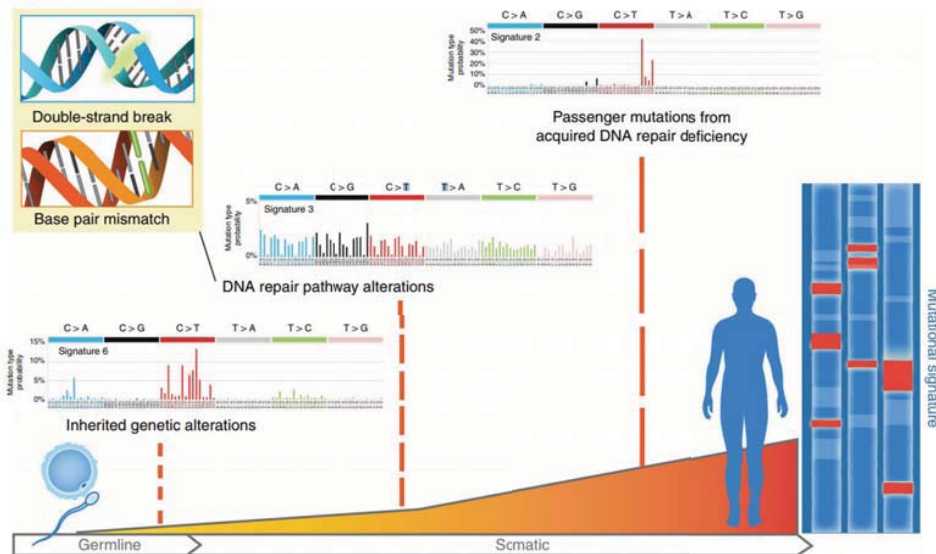
Here is an example based on a text-mining application:

- Let the input matrix (the matrix to be factored) be **V** with 10000 rows and 500 columns where words are in rows and documents are in columns. That is, we have 500 documents indexed by 10000 words. It follows that a column vector \mathbf{v} in **V** represents a document.
- Assume we ask the algorithm to find 10 features in order to generate a *features matrix* **W** with 10000 rows and 10 columns and a *coefficients matrix* **H** with 10 rows and 500 columns.
- The product of **W** and **H** is a matrix with 10000 rows and 500 columns, the same shape as the input matrix **V** and, if the factorization worked, it is a reasonable approximation to the input matrix **V**.
- From the treatment of matrix multiplication above it follows that each column in the product matrix **WH** is a linear combination of the 10 column vectors in the features matrix **W** with coefficients supplied by the coefficients matrix **H**.

This last point is the basis of NMF because we can consider each original document in our example as being built from a small set of hidden features. NMF generates these features.

It is useful to think of each feature (column vector) in the features matrix **W** as a document archetype comprising a set of words where each word's cell value defines the word's rank in the feature: The higher a word's cell value the higher the word's rank in the feature. A column in the coefficients matrix **H** represents an original document with a cell value defining the document's rank for a feature. We can now reconstruct a document (column vector) from our input matrix by a linear combination of our features (column vectors in **W**) where each feature is weighted by the feature's cell value from the document's column in **H**.

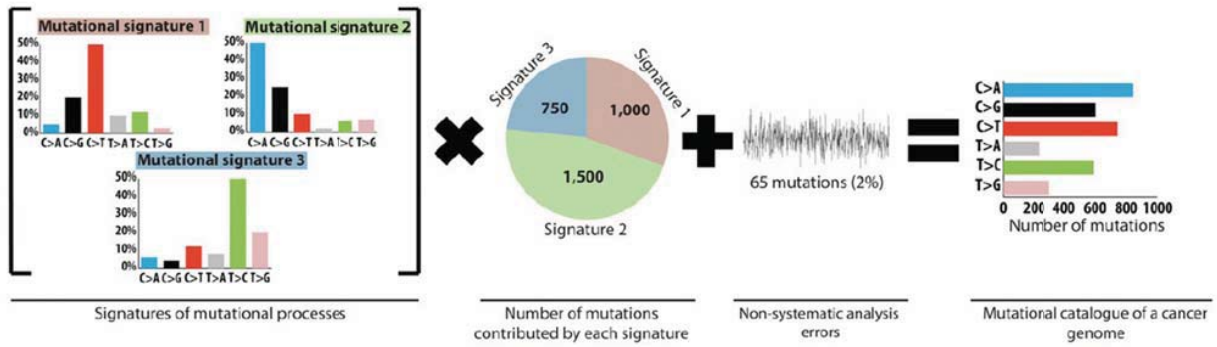
Mutations accumulated over lifetime can be decomposed



Jennifer Ma et al., *Nat Comm* 2017

20

Mutational signature decomposition



Alexandrov et al., *Cell* 2013

21

Mutational signatures

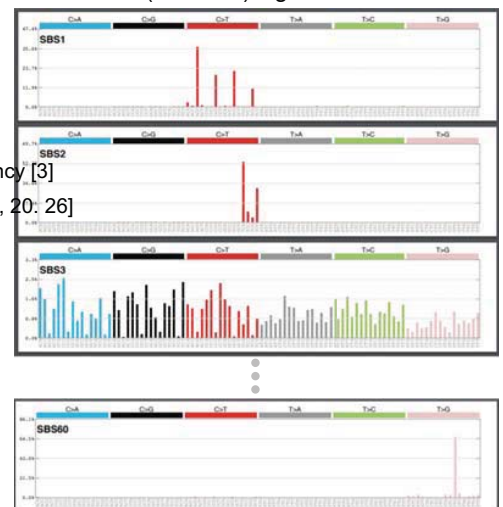
30 Catalogue of Somatic Mutations in Cancer (COSMIC) signatures



Alexandrov et al., *Nature* 2013

- Age [1, 5]
- APOBEC activity [2, 13]
- Homologous recombination deficiency [3]
- Defective DNA mismatch repair [15, 20, 26]
- Tobacco smoking/chewing [4, 29]
- Ultraviolet light exposure [7]
- *POLE* mutations [10]
- Exposure to aristolochic acid [22]
- Alkylating agent temozolomide [11]
- Aflatoxin [24]

65 PanCancer Analysis of Whole Genomes (PCAWG) signatures

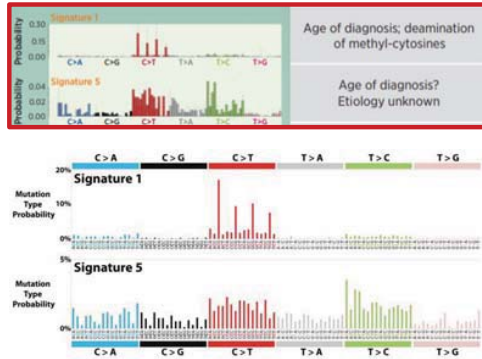


Campbell et al., *bioRxiv* 2017

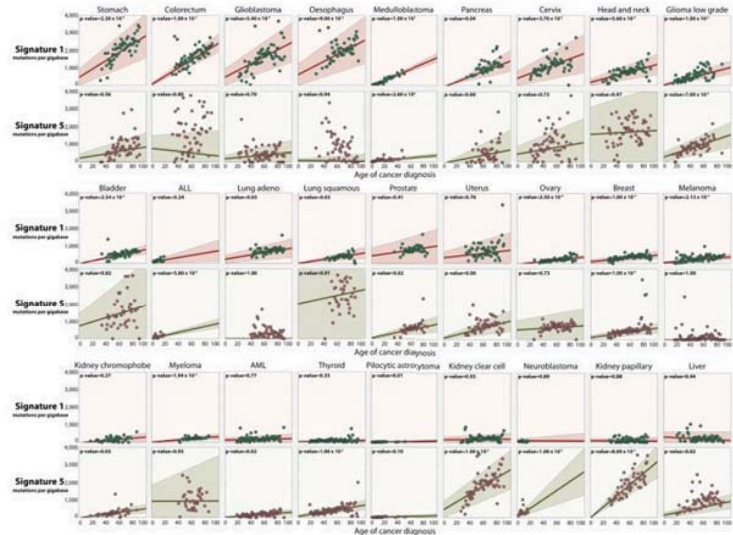
22

Clock-like mutational signatures

Nik-Zainal et al., *Clinical Cancer Research* 2017



Signature 1 and 5 is positively correlated to the age of patients



Alexandrov, Ludmil B., et al. *Nat Gen* 2015

23

Mutational signatures are clinically relevant

A	Associations	Prevalence in breast cancer	Presence in other cancer types
Signature 1	Age of diagnosis; deamination of methyl-cytosines	Common >75% of samples	Reported in nearly all other tumor types
Signature 5	Age of diagnosis? Etiology unknown	Common >75% of samples	Reported to be in many other tumor types, but not consistently
Signature 8	Increased in HR deficiency and late in cancer evolution but present at lower levels in many tumors	Common >60% of samples	Many tumor types
Signature 3	Homologous recombination repair deficiency	Common >20% of samples	Many tumor types
Signature 2	APOBEC cytidine deaminases	Common >75% of samples	More than half of all tumor types examined so far
Signature 13	APOBEC cytidine deaminases	Common >75% of samples	Many tumor types
Signature 6	Mismatch repair deficiency	Rare <10% of samples	Adrenocortical, colon, uterine, ovarian, pancreas
Signature 20	Mismatch repair deficiency	Rare <5% of samples	Gastric
Signature 26	Mismatch repair deficiency	Rare <5% of samples	Cervical, gastric, uterine

Nik-Zainal et al., *Clinical Cancer Research* 2017

“... **mutational signatures [3, 6, 20, 26]** are a direct pathophysiologic readout of the abrogation of a DNA repair gene/pathway and could be used as a **biomarker** to report **DNA repair/deficiency** in a tumor.”

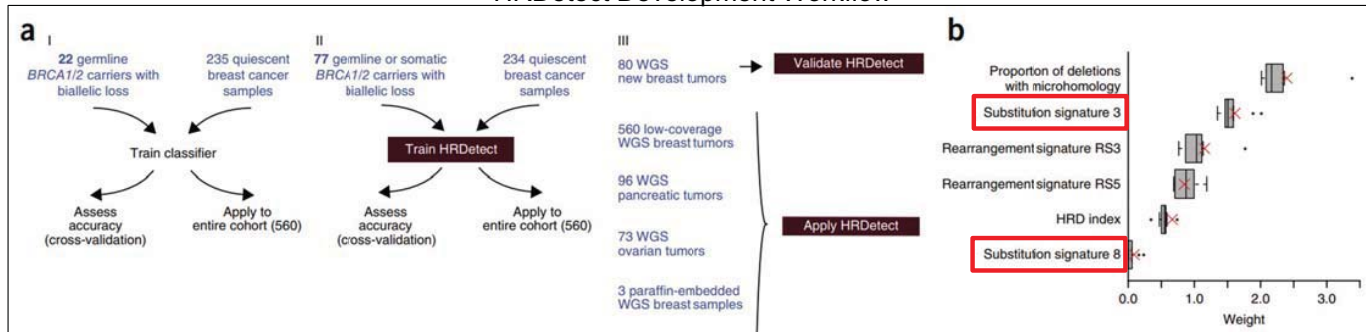
24

Mutational signature #3 (homologous recombination defect)

HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures

Helen Davies^{1,32}, Dominik Glodzik^{1,32}, Sandro Morganello¹, Lucy R Yates^{1,2}, Johan Staaf³, Xueqing Zou¹, Manasa Ramakrishna^{1,4}, Sancha Martin¹, Sandrine Boyault⁵, Anieta M Sieuwerts⁶, Peter T Simpson⁷, Tari A King⁸, Keiran Raine¹, Jorunn E Eyfjord⁹, Gu Kong¹⁰, Åke Borg², Ewan Birney¹¹, Hendrik G Stunnenberg¹², Marc J van de Vijver¹³, Anne-Lise Borresen-Dale^{14,15}, John W M Martens⁶, Paul N Span^{16,17}, Sunil R Lakhani^{7,18}, Anne Vincent-Salomon^{19,20}, Christos Sotiriou²¹, Andrew Tutt^{22,23}, Alastair M Thompson²⁴, Steven Van Laere^{25,26}, Andrea L Richardson^{27,28}, Alain Viari^{29,30}, Peter J Campbell¹, Michael R Stratton¹ & Serena Nik-Zainal^{1,31}

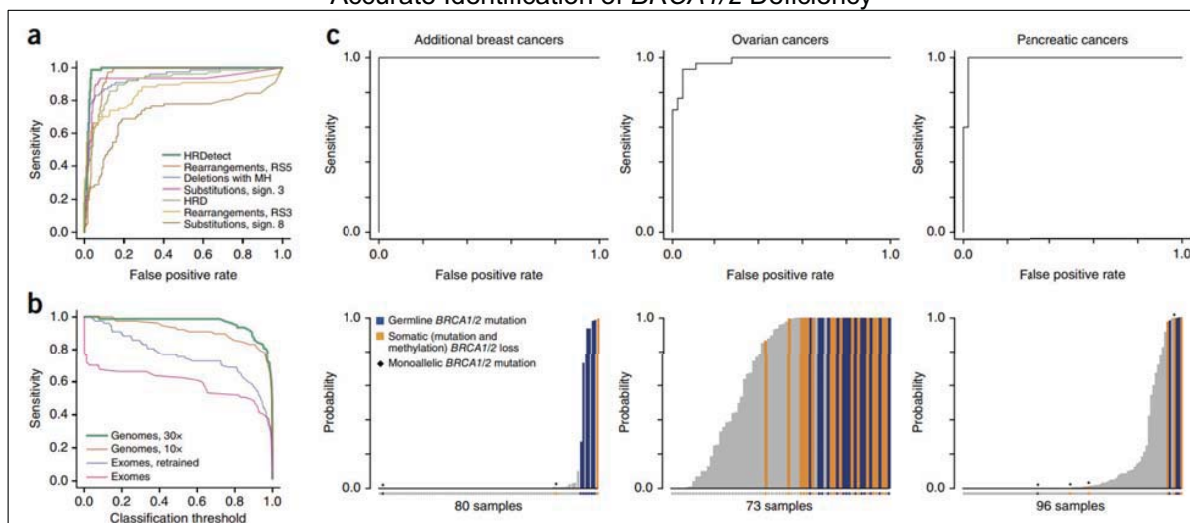
HRDetect Development Workflow



Helen Davies et al., *Nature Medicine* 2017

Mutational signature #3 (homologous recombination defect)

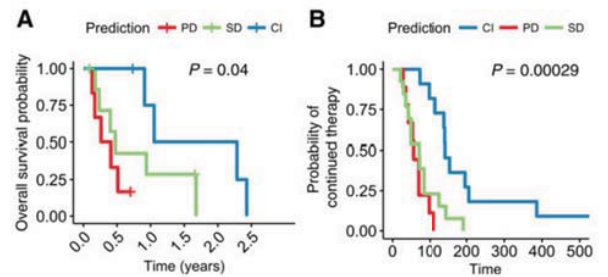
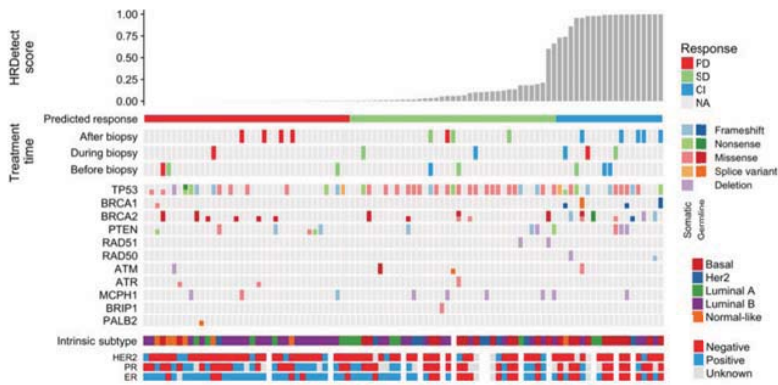
Accurate Identification of *BRCA1/2* Deficiency



Helen Davies et al., *Nature Medicine* 2017

Mutational signature #3 (homologous recombination defect)

Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer

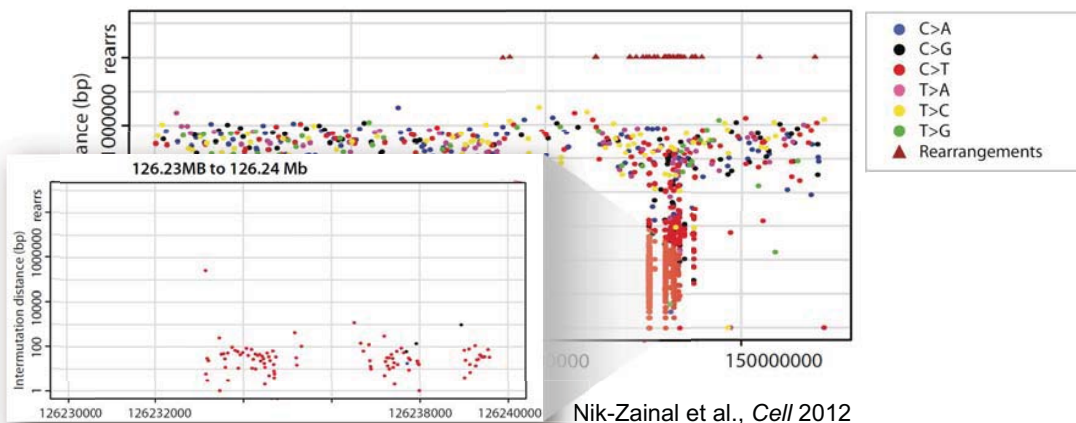


Eric Y. Zhao et al., *Clinical Cancer Research* 2017

27

Localized hypermutations (kataegis)

Chromosome 6 PD4107a

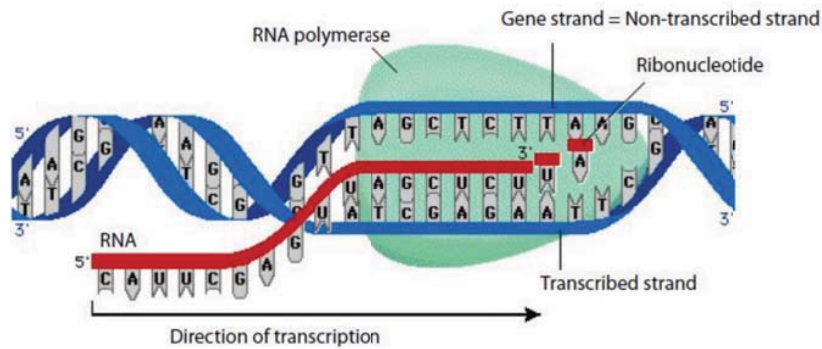


Nik-Zainal et al., *Cell* 2012

- There are clusters of mutations with short distances between mutations (kataegis)
- Kataegis is co-localized with rearrangements that have features of chromothripsis

28

Transcriptional strand bias



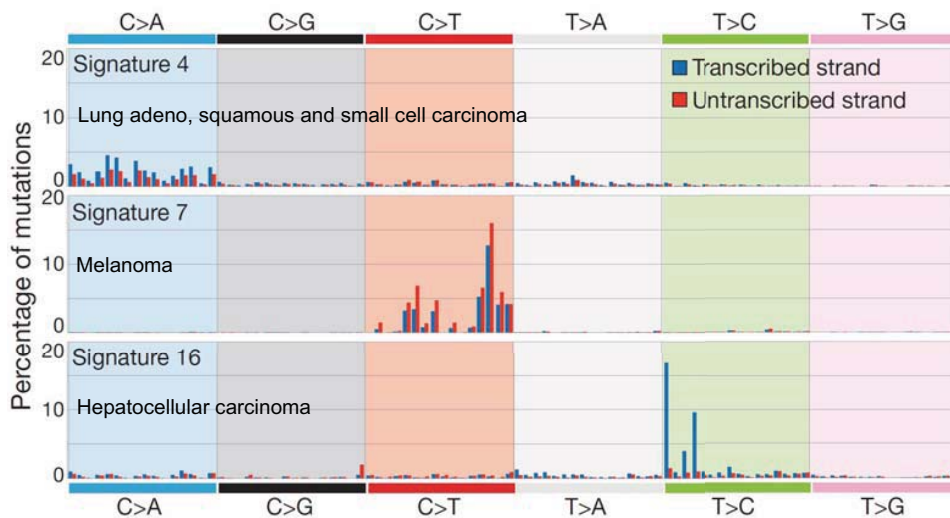
<ftp://ftp.sanger.ac.uk/pub/resources/theses/snz/chapter6.pdf>

Cause of transcriptional strand bias is transcription-coupled repair (TCR) of nucleotide excision repair (NER); DNA damage is repaired more efficiently on the transcribed strand

29

Somatic mutation rates are biased by transcriptional strand

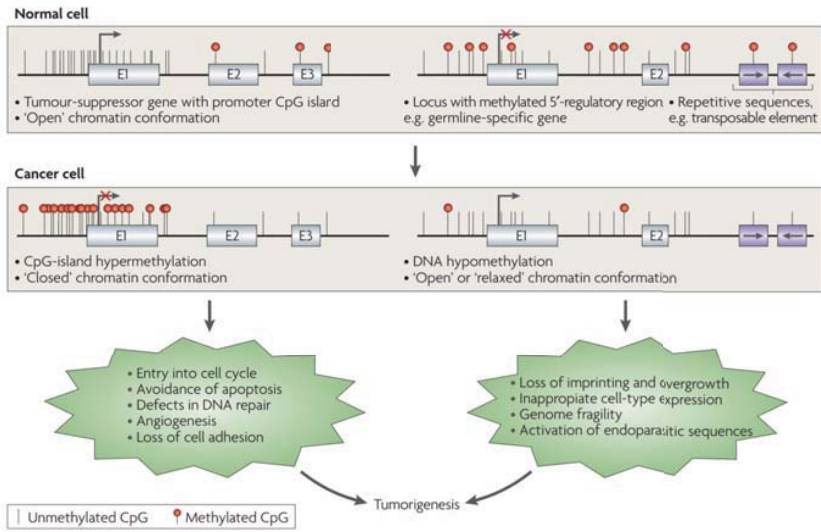
Fewer mutations accumulate on the transcribed strand



Alexandrov et al., *Nature* 2013

30

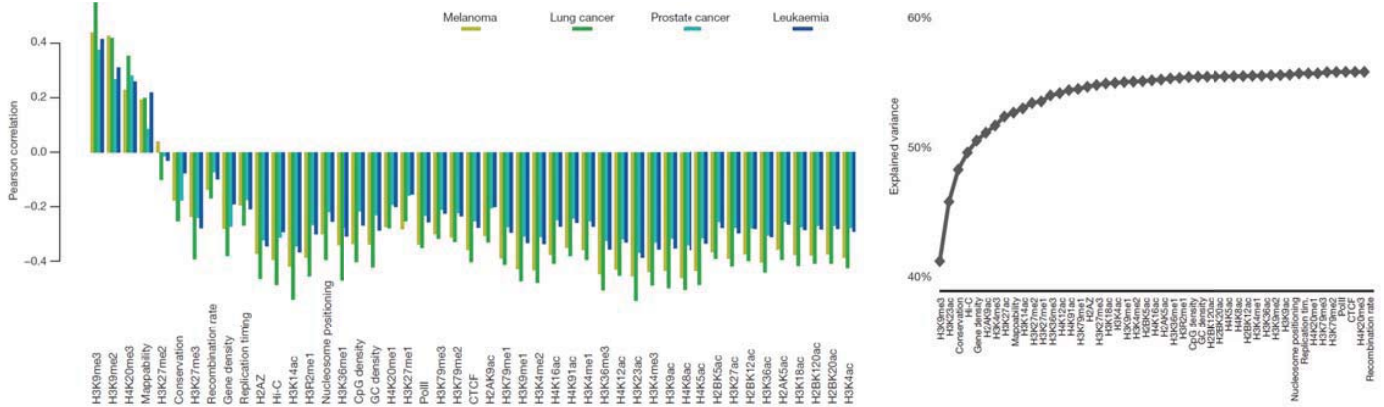
Cancer and epigenomics



Esteller, *Nature Reviews Genetics* 2007

31

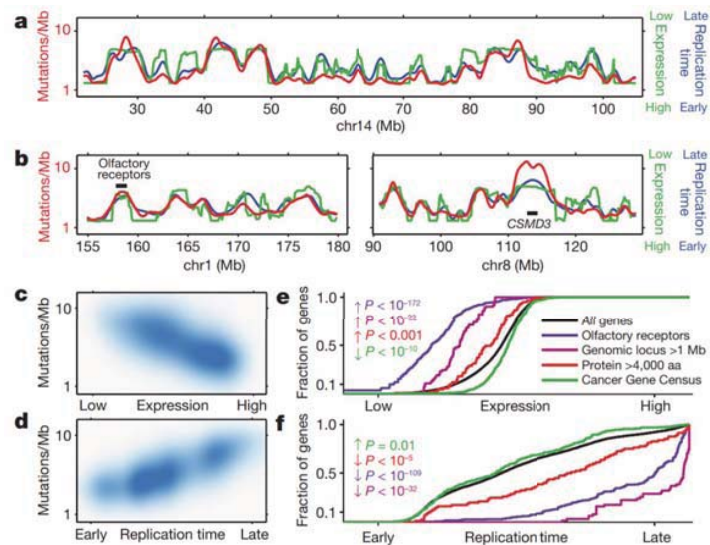
Mutation rates are correlated with genomic/epigenomic modifications



Schuster-Bockler and Lehner, *Nature* 2012

32

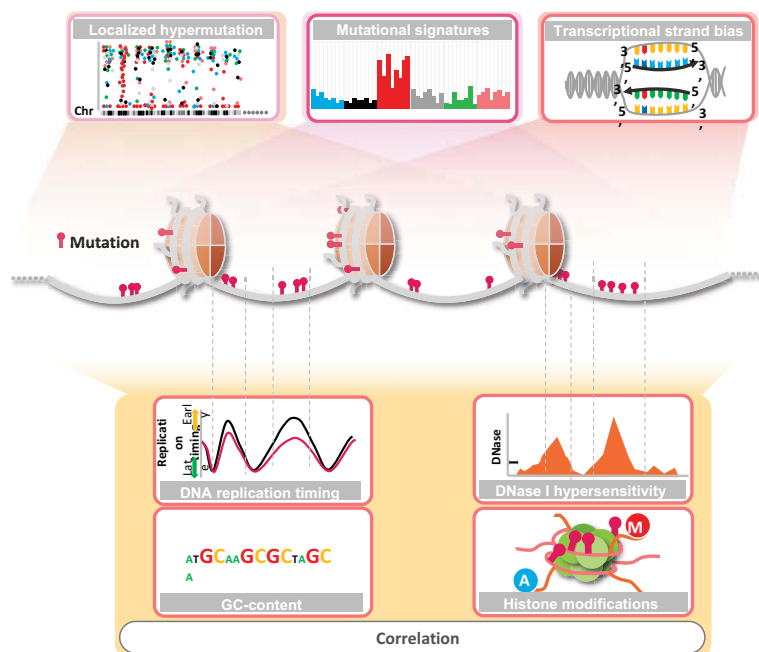
Mutation rates also vary with DNA replication timing



Lawrence et al., *Nature* 2013

33

Putting them altogether to better understand mutations operative in cancer



34

Existing mutational signature decomposition tools

Tool	Graphical user interface available?	Batch analysis possible?	Results downloadable?
WTSI (Alexandrov et al., Cell 2013)	No	Yes	Yes
deconstructSigs (Rosenthal et al., Genome Biology 2016)	No	Yes	Yes
MutaGene (Goncalves et al., Nucleic Acids Research 2017)	Yes	No	No

35

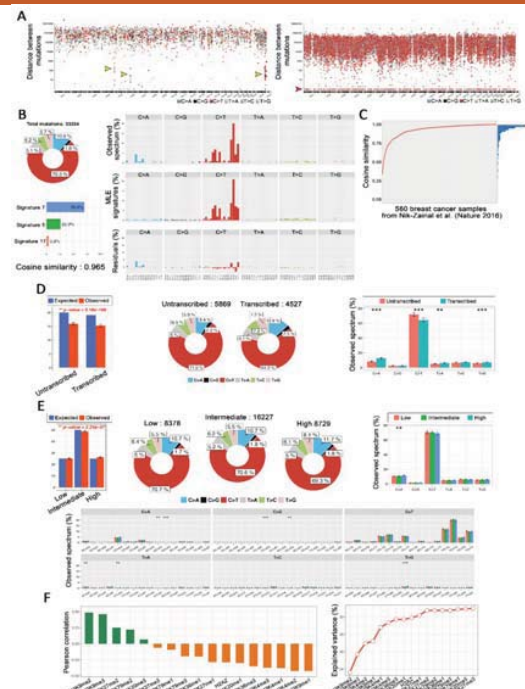
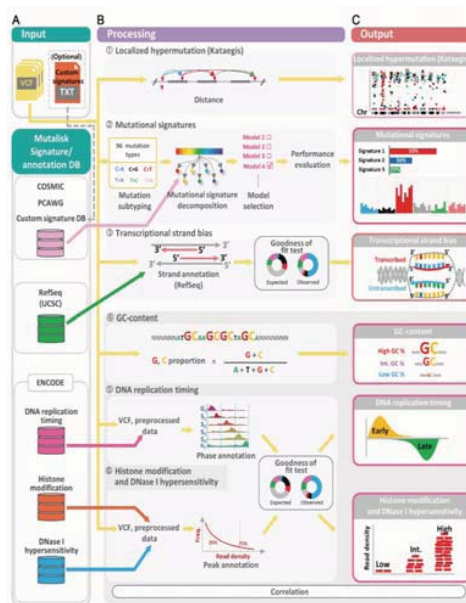
Mutalisk: a web-based somatic MUTation AnaLysis toolKit for genomic, transcriptional and epigenomic signatures (Nucleic Acids Research, 2018)

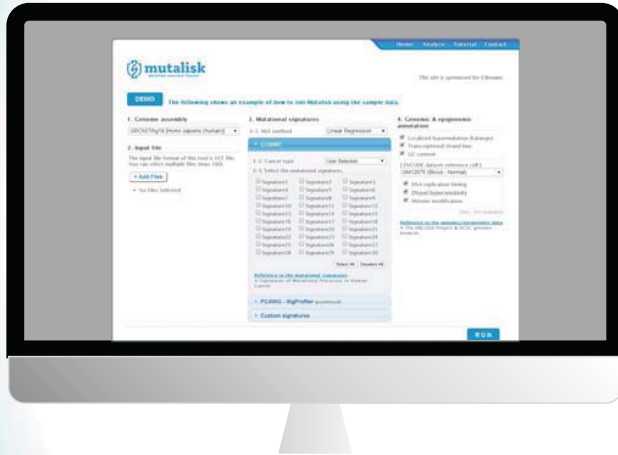
Mutalisk: a web-based somatic MUTation AnaLysis toolKit for genomic, transcriptional and epigenomic signatures
 Jonghwan Lee¹, Andy Anson Lee¹, Joon Won Lee¹, Jonghwan Park¹, Sanghvi Ram², Sanggeun Park³, Hyunsoo Chae⁴, Young Seok Ah⁵ and Donggeun Hong^{1*}
¹Department of Biomedical Science, National Cancer Center, Goyang-si, Gyeonggi-do, Korea; ²Department of Biostatistics, National Cancer Center, Goyang-si, Gyeonggi-do, Korea; ³Department of Biostatistics, National Cancer Center, Goyang-si, Gyeonggi-do, Korea; ⁴Department of Biostatistics, National Cancer Center, Goyang-si, Gyeonggi-do, Korea; ⁵Department of Biostatistics, National Cancer Center, Goyang-si, Gyeonggi-do, Korea; *Correspondence: Donggeun Hong, National Cancer Center, Goyang-si, Gyeonggi-do, Korea (E-mail: hongdg@krcr.go.kr)

ABSTRACT
 Somatic mutations are essential for cancer development and progression. However, the underlying mechanisms of somatic mutations are still unclear. To understand the underlying mechanisms of somatic mutations, we developed a web-based somatic mutational signature analysis tool, Mutalisk. Mutalisk is a web-based somatic mutational signature analysis tool that integrates genomic, transcriptional, and epigenomic signatures. Mutalisk is designed to be user-friendly and easy to use. Mutalisk is available at <http://mutalisk.krcr.go.kr>.

INTRODUCTION
 Somatic mutations are essential for cancer development and progression. However, the underlying mechanisms of somatic mutations are still unclear. To understand the underlying mechanisms of somatic mutations, we developed a web-based somatic mutational signature analysis tool, Mutalisk. Mutalisk is a web-based somatic mutational signature analysis tool that integrates genomic, transcriptional, and epigenomic signatures. Mutalisk is designed to be user-friendly and easy to use. Mutalisk is available at <http://mutalisk.krcr.go.kr>.

개방형 오픈 플랫폼





1. Identification of mutational signatures
1. Analysis of associations between various genome/epigenome regulatory elements and somatic mutation rates

Available at <http://mutalisk.org/>

37

Mutalisk (1) – analysis options

GRCh37/hg19, GRCh38/hg38, GRCm38/mm10, WBcel236/ce11, WS220/ce10

Multiple vcf files (max 100)

COSMIC, PCAWG (SigProfiler) signatures

User/custom signatures

Linear regression or multinomial test methods (signature decomposition)

31 cell lines

42 cancer types

1. Genome assembly: GRCh37/hg19 [Homo sapiens (human)]

2. Input file: + Add Files

3. Mutational signatures: 3-1. MLE method: Linear Regression; 3-2. Cancer type: User Selection; 3-3. Select the mutational signatures: COSMIC (Signature1-30), PCAWG - SigProfiler (provisional), Custom signatures

4. Genomic & epigenomic annotation: Localized hypermutation (kataegis), Transcriptional strand bias, GC content, [ENCODE dataset reference cell]: GM12878 (Blood - Normal), DNA replication timing, DNaseI hypersensitivity, Histone modification

RUN

National Cancer Center, 323 Ilan-ro, Ilсандong-gu, Goyang-si Gyeonggi-do, 10408, Republic of Korea

38

Mutalisk (1) – analysis options

Custom signatures

3-2. Select the mutational signatures.
Please upload your own signature file for the decomposition of mutational signatures. A tab-delimited sample signature file is available below:
[[Sample signature file](#)]

For additional information on the formatting of the signature file, please refer to the [Tutorials](#) page.

+ Add file • No File Selected

Provide a tab-delimited .txt signature file

Substitution Type	Trinucleotide	Somatic Mutation Type	Signature 1	...	Signature 7
C>A	ACA	A[C>A]A	0.011098326		0.0004
C>A	ACC	A[C>A]C	0.009149341		0.0005
C>A	ACG	A[C>A]G	0.00149007		0
C>A	ACT	A[C>A]T	0.006233885		0.0004
C>G	ACA	A[C>G]A	0.001801068		0
		⋮			⋮
T>G	TTT	T[T>G]T	0.004030128	...	0.0014

96 mutation subtypes

39

Mutalisk (1) – analysis options

mutalisk.org says

It takes about 3-5 minutes per file.
You can either wait or reconnect to the below address.
Copy to clipboard: Ctrl+C, Enter

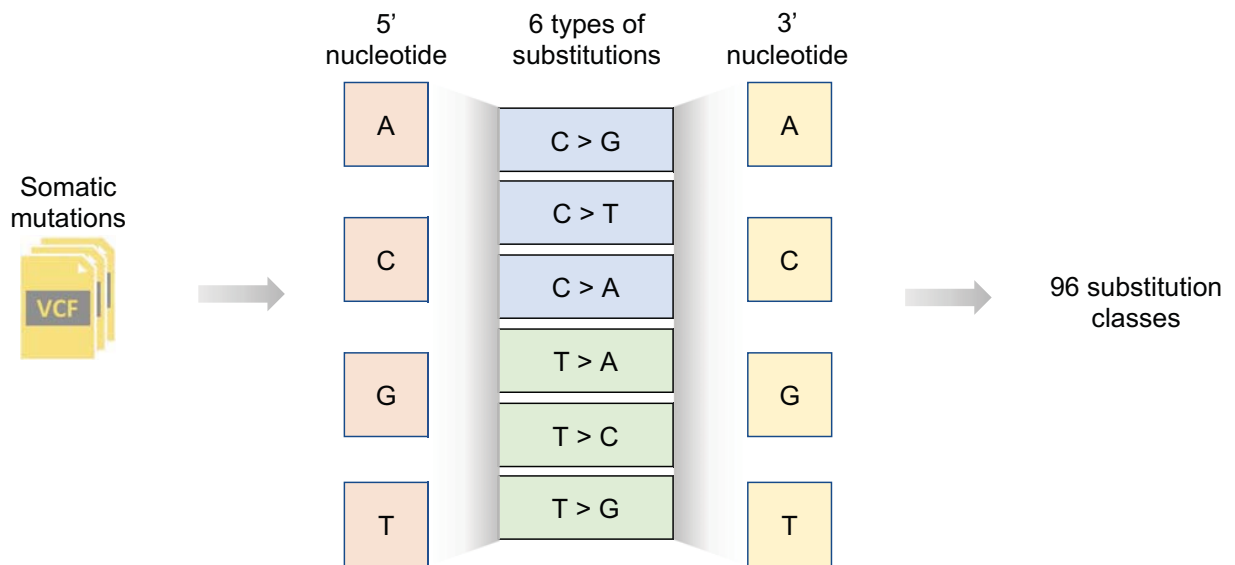
<http://mutalisk.org/result.php?rid=4P9iaD51mg>

OK
Cancel

User-uploaded data are permanently deleted after 48 hours;
you may access the analysis results for 48 hours

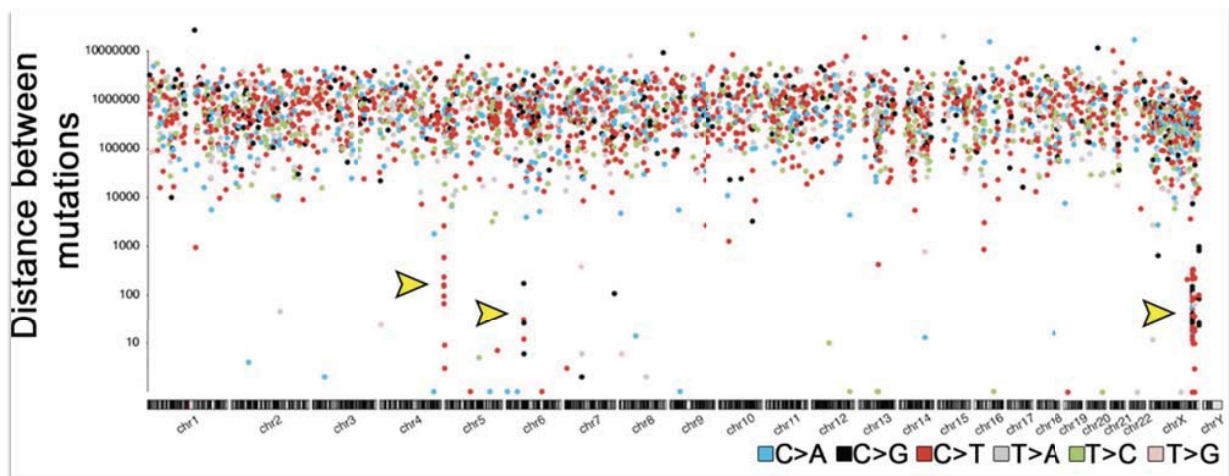
40

Mutalisk (2) – preprocessing



41

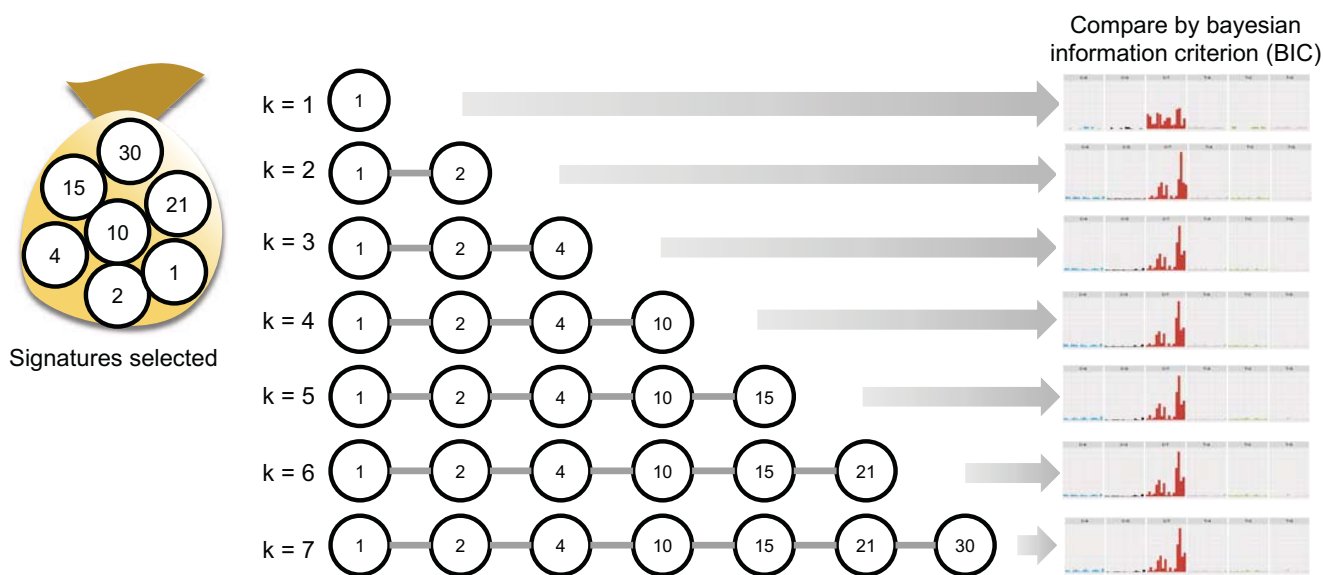
Mutalisk (3) – rainfall plot (*kataegis*)



Lung cancer sample (Lee et al., *J Clin Oncol* 2017)

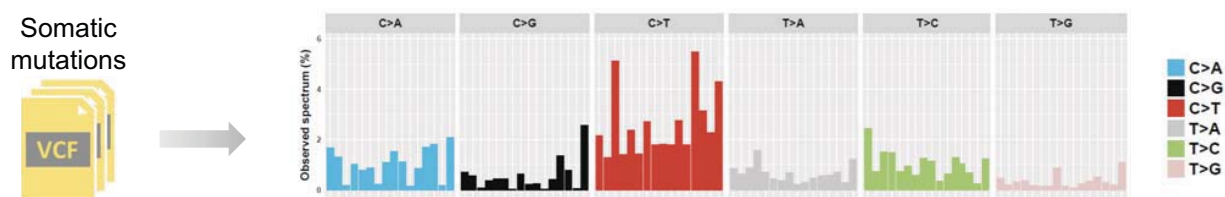
42

Mutalisk (4) – mutational signature identification



43

Mutalisk (4) – mutational signature identification



If we conjecture that there are 7 COSMIC signatures underlying the mutational profile, then

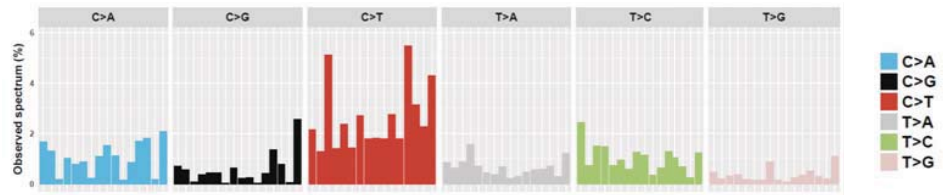
$$C(30,7) = \frac{30!}{7!(30-7)!}$$

2,035,800 different combinations of signatures

44

Mutalisk (4) – mutational signature identification

Somatic mutations



If we conjecture that there are 7 PCAWG signatures underlying the mutational profile, then

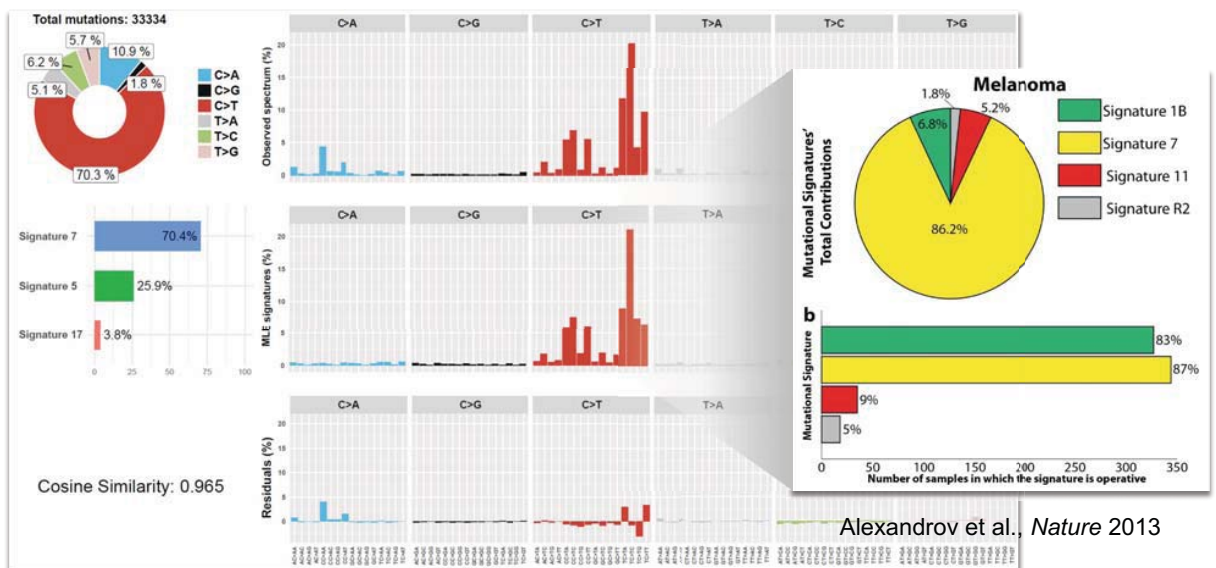
$$C(65,7) = \frac{65!}{(7! (65 - 7)!)}$$

696,190,560 different combinations of signatures

This is a computationally expensive task

45

Mutalisk (4) – mutational signature identification



Melanoma sample (Plesance et al., *Nature* 2010)

46

Mutalisk (4) – mutational signature identification

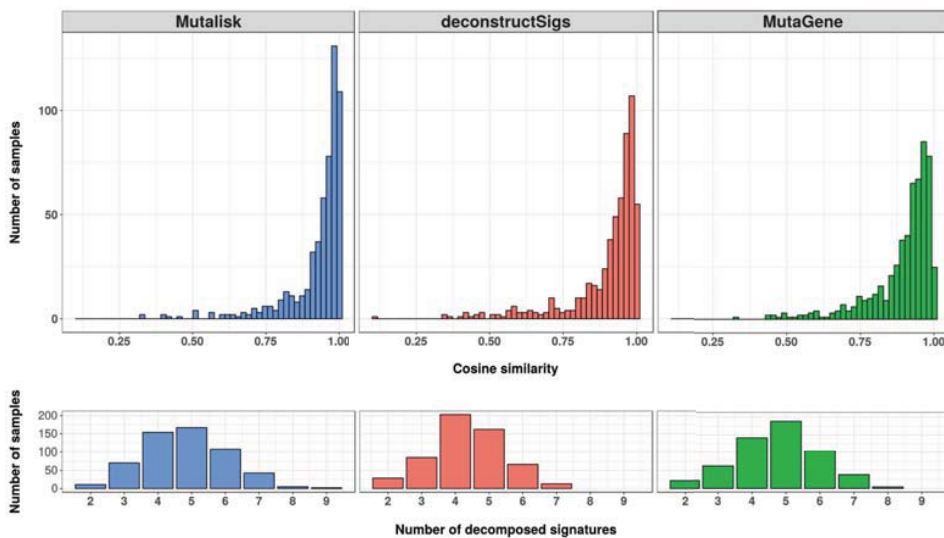
Mutalisk identifies up to 7 decomposition models

Mutational signatures					
No.	Signatures	Probabilities	Cosine similarity	BIC	Confidence Interval
7	7 8 2 11 18 17 12 <input checked="" type="checkbox"/> <i>Our Best</i>	0.59097 0.11749 0.10825 0.06858 0.05707 0.03374 0.02390	● 0.98200	222713.700	View
6	7 8 2 11 18 17	0.59163 0.14049 0.10844 0.06984 0.05328 0.03632	● 0.98200	222820.200	View
5	7 8 2 11 17	0.59492 0.19932 0.10851 0.06440 0.03285	● 0.98200	223568.700	View
3	7 8 2	0.64330 0.25263 0.10407	● 0.98000	224619.100	View
4	7 8 2 11	0.59826 0.22793 0.10916 0.06466	● 0.98100	224828.400	View
2	7 8	0.71863 0.28137	● 0.96800	225142.600	View
1	7	1.00000	● 0.78700	273566.100	View

47

Mutalisk (4) – mutational signature identification

Comparison of decomposition results using 560 breast cancer samples (Nik-Zainal et al., Nature 2016)



Tool	Median cosine similarity
Mutalisk	0.966
deconstructSigs	0.948
MutaGene	0.931

48

Mutalisk (5) – transcriptional strand bias analysis

Somatic mutations



UCSC RefSeq



Strand annotation



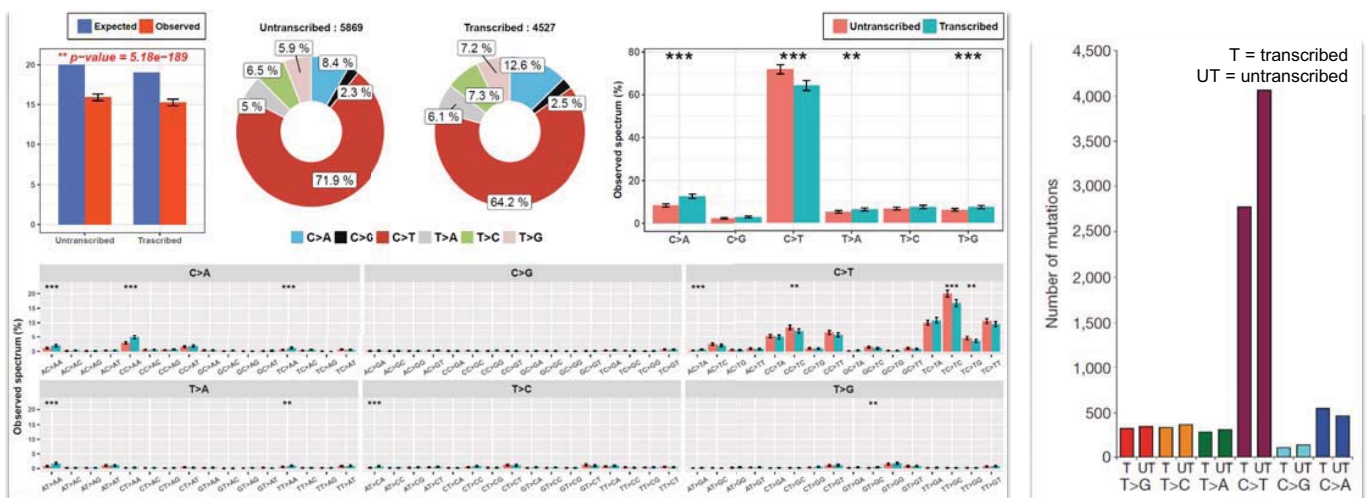
Goodness of fit test



Reference genome	Transcribed strand region proportion	Untranscribed strand region proportion
CRCh37/hg19	0.19044	0.20010
GRCh38/hg38	0.18885	0.28435

49

Mutalisk (5) – transcriptional strand bias analysis

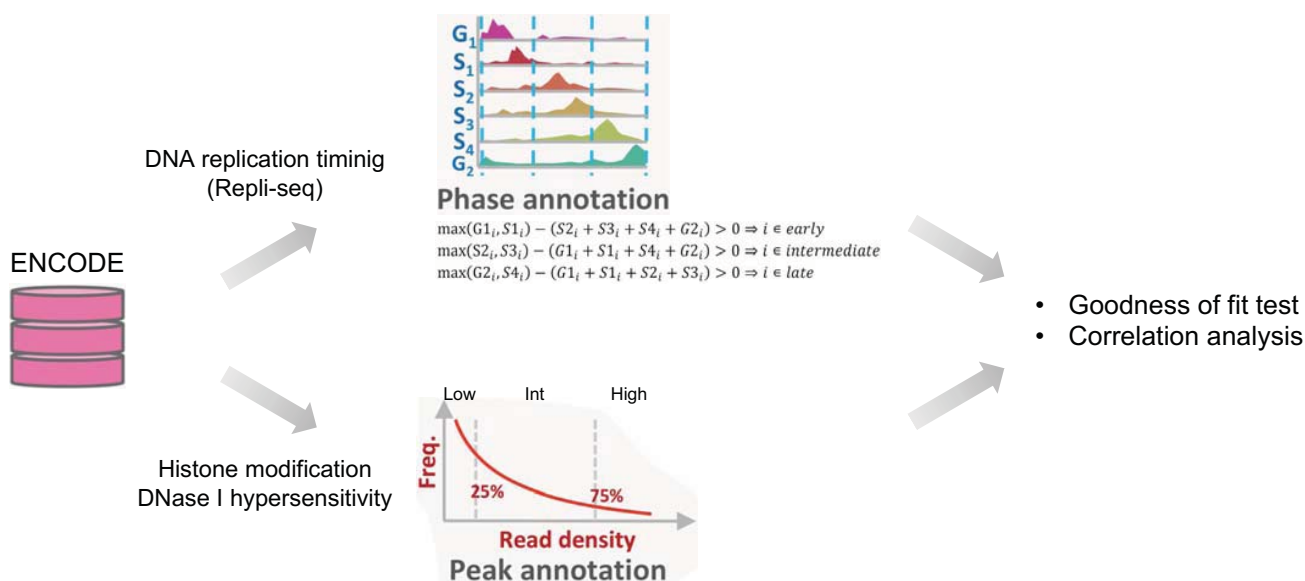


Melanoma sample (Plesance et al., *Nature* 2010)

Plesance et al., *Nature* 2010

50

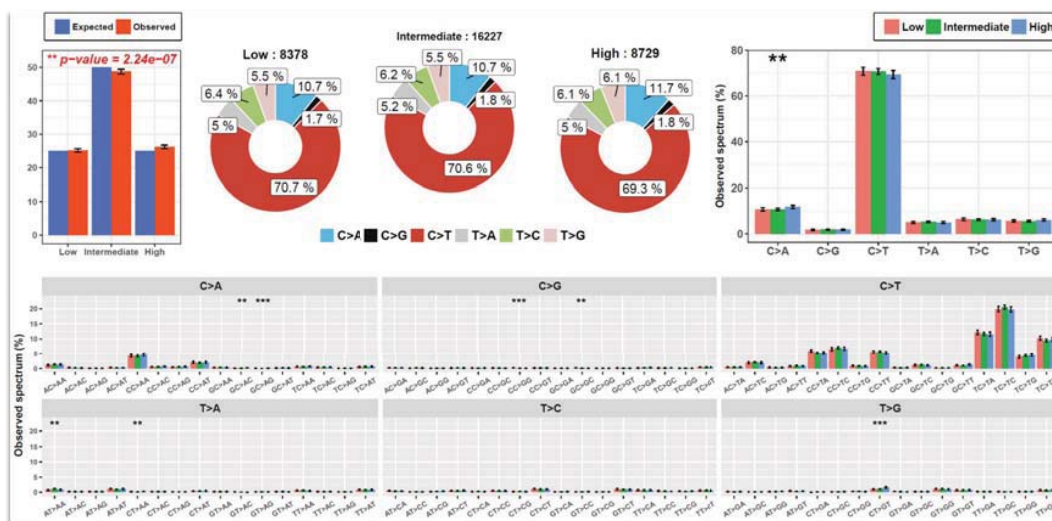
Mutalisk (6) – genomic/epigenomic modification analysis



51

Mutalisk (6) – genomic/epigenomic modification analysis

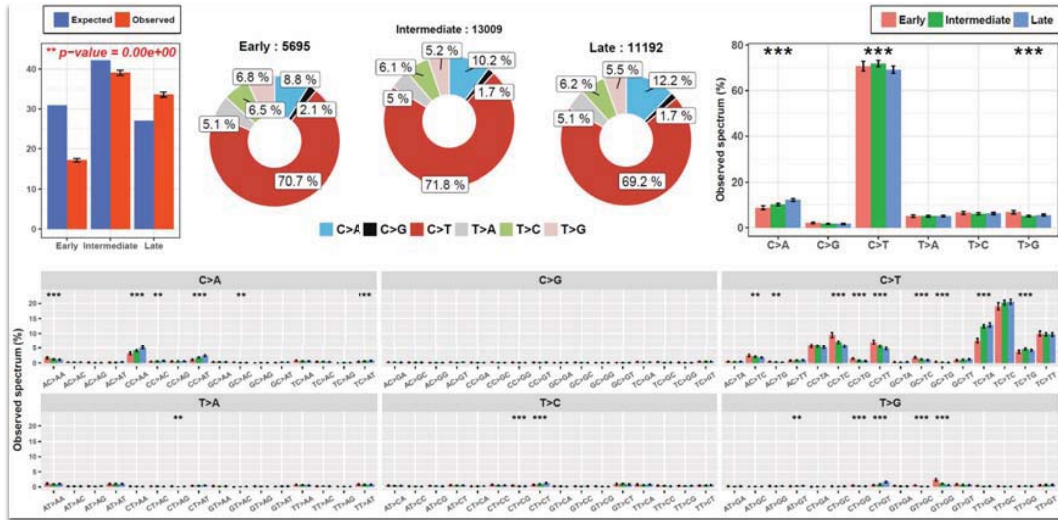
H3k9me3 , melanoma sample (Pleasant et al., Nature 2010)



52

Mutalisk (6) – genomic/epigenomic modification analysis

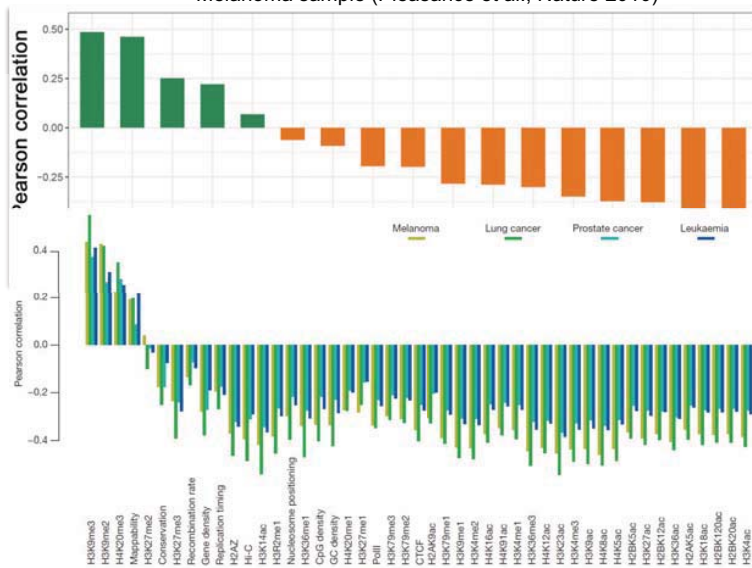
DNA replication timing, melanoma sample (Pleasant et al., Nature 2010)



53

Mutalisk (6) – genomic/epigenomic modification analysis

Melanoma sample (Pleasant et al., Nature 2010)



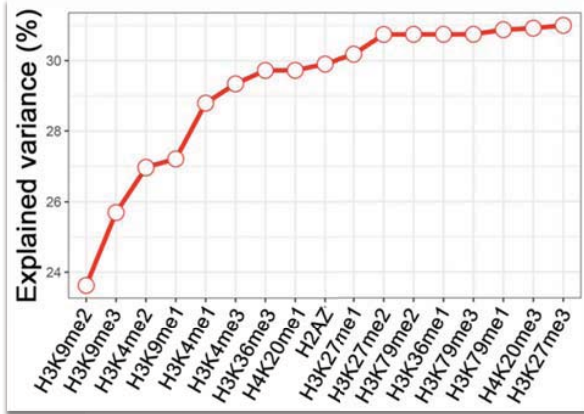
Mutalisk

Schuster-Bockler and Lehner, Nature 2012

54

Mutalisk (6) – genomic/epigenomic modification analysis

Melanoma sample (Pleasance et al., Nature 2010)



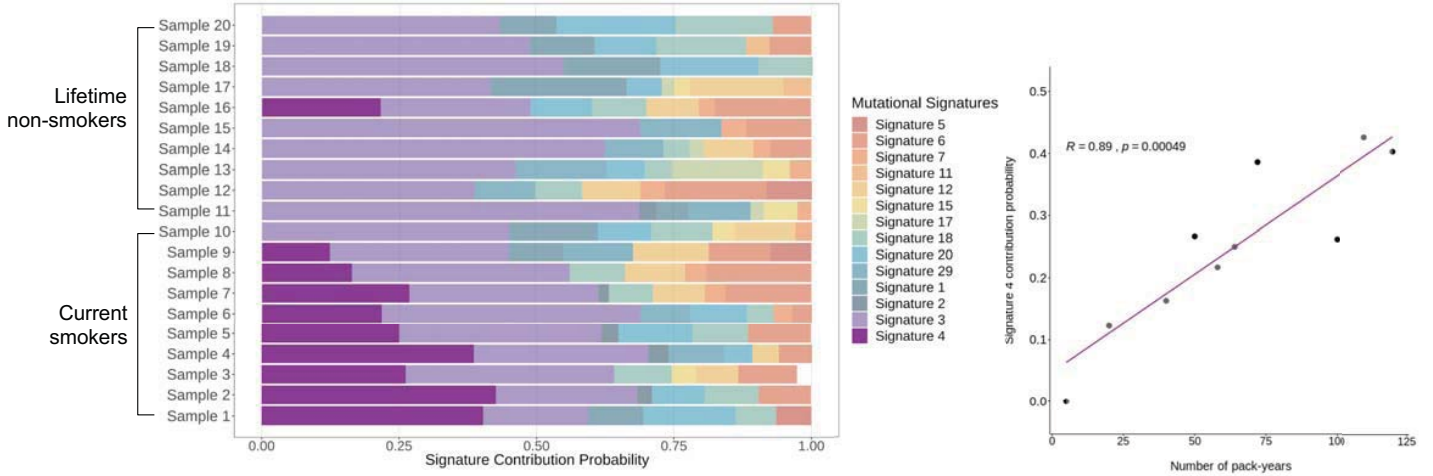
- Forward feature selection (generalized least-squares model and Akaike information criterion)
- Explained variance computed by linear regression

Mutalisk (7) – downloading results

The screenshot displays the Mutalisk web interface. On the left, a PDF viewer shows analysis results with bar charts and donut charts. Below the PDF is a 'TXT' download button. On the right, a list of analysis results is shown, including sample identifiers and analysis types. At the bottom right, there are two buttons: 'Get merged results' and 'Get all results at once'. The navigation menu at the top right includes 'Home', 'Analyze', 'Tutorial', and 'Contact'.

Tobacco-associated mutational signature #4

TCGA Lung Adenocarcinoma Samples



Run Mutalisk

Home Analyze Tutorial Contact



This site is optimized for Chrome.

DEMO

The following shows an example of how to run Mutalisk using the sample data.

Melanoma data example

Lung data example

§ Download the sample data: [Melanoma / Lung]

Source of the sample data: [Melanoma cancer] *E. D. Pleasance et. al. Nature 2010.* & [Lung cancer] *June-Koo Lee et. al. JCO. 2017.*

sample_lung.vcf

Variant Call Format (VCF): sample_lung.vcf

```
##fileformat=VCFv4.2
#CHROM POS ID REF ALT QUAL FILTER INFO
1 1910980 . A G . . .
1 5148788 . G C . . .
1 5867505 . C T . . .
1 7146801 . G A . . .
1 7665765 . C T . . .
1 7749485 . T G . . .
1 8710829 . G A . . .
1 9804956 . C T . . .
1 12381586 . G A . . .
1 17274833 . G T . . .
1 17365503 . T A . . .
1 18099894 . A T . . .
1 19665267 . G A . . .
1 20753216 . C A . . .
1 22905721 . T C . . .
1 28496930 . T C . . .
1 29464355 . A G . . .
1 30253004 . G T . . .
1 30346067 . G A . . .
1 30627515 . T C . . .
1 32251636 . T A . . .
1 34081408 . G A . . .
1 34435930 . A T . . .
1 35909337 . T C . . .
1 36062312 . T C . . .
1 36414737 . T A . . .
1 36729500 . A T . . .
1 36936453 . A T . . .
1 37261912 . G T . . .
1 37277934 . G A . . .
1 37708424 . C T . . .
1 39672350 . G C . . .
1 42038123 . C T . . .
1 45810549 . C T . . .
1 46950051 . C T . . .
1 46984444 . C T . . .
1 47098023 . G T . . .
1 47446596 . G T . . .
1 47619512 . C A . . .
1 48781219 . A G . . .
1 50032004 . G A . . .
1 50642579 . G T . . .
```

행 1, 열 1 - 3399 행

The screenshot shows the Mutalisk web application interface. At the top, there are navigation links: Home, Analyze, Tutorial, and Contact. The main heading is "mutalisk METASTASIS ANALYSIS TOOLKIT". Below this, a "DEMO" section states: "The following shows an example of how to run Mutalisk using the sample data."

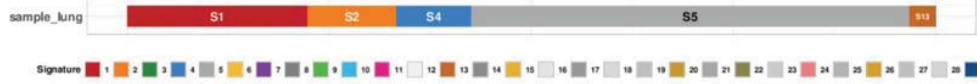
The interface is divided into four main sections:

- 1. Genome assembly:** A dropdown menu is set to "GRCh37/hg19 [Homo sapiens (human)]".
- 2. Input file:** A text box explains that the input file format is VCF and that multiple files (max 300) can be selected. A file named "sample_lung.vcf" is listed under an "Add Files" button.
- 3. Mutational signatures:**
 - 3-1. MLE method: "Linear Regression" (dropdown).
 - 3-2. Cancer type: "Lung Adeno" (dropdown, circled in red).
 - 3-3. Select the mutational signatures: A grid of checkboxes for COSMIC signatures 1 through 30. Signatures 1, 4, 5, 6, 13, and 17 are checked.
- 4. Genomic & epigenomic annotation:**
 - Checkboxes for "Localized hypermutation (kataegis)", "Transcriptional strand bias", and "GC content" are checked.
 - Reference cell: "GM12878 (Blood - Normal)" (dropdown).
 - Checkboxes for "DNA replication timing", "DNaseI hypersensitivity", and "Histone modification" are checked.

At the bottom right, there is a large blue "RUN" button. At the bottom center, the footer text reads: "National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang-si Gyeonggi-do, 10408, Republic of Korea".



Summary of the best results from our tool



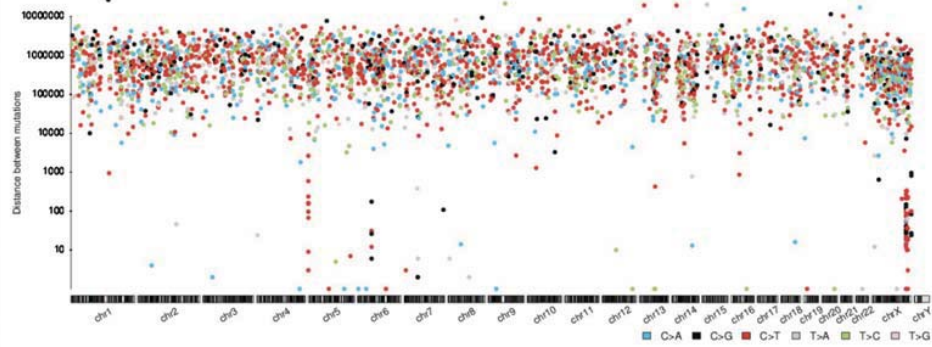
sample_lung

File : sample_lung

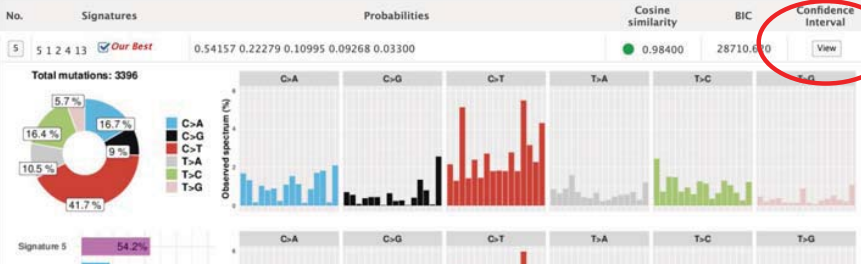
Get merged results

Get all results at once

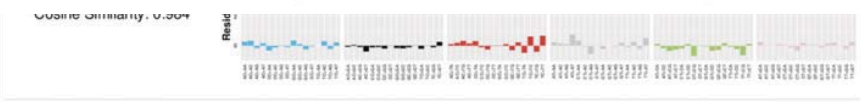
Localized hypermutation (kataegis)



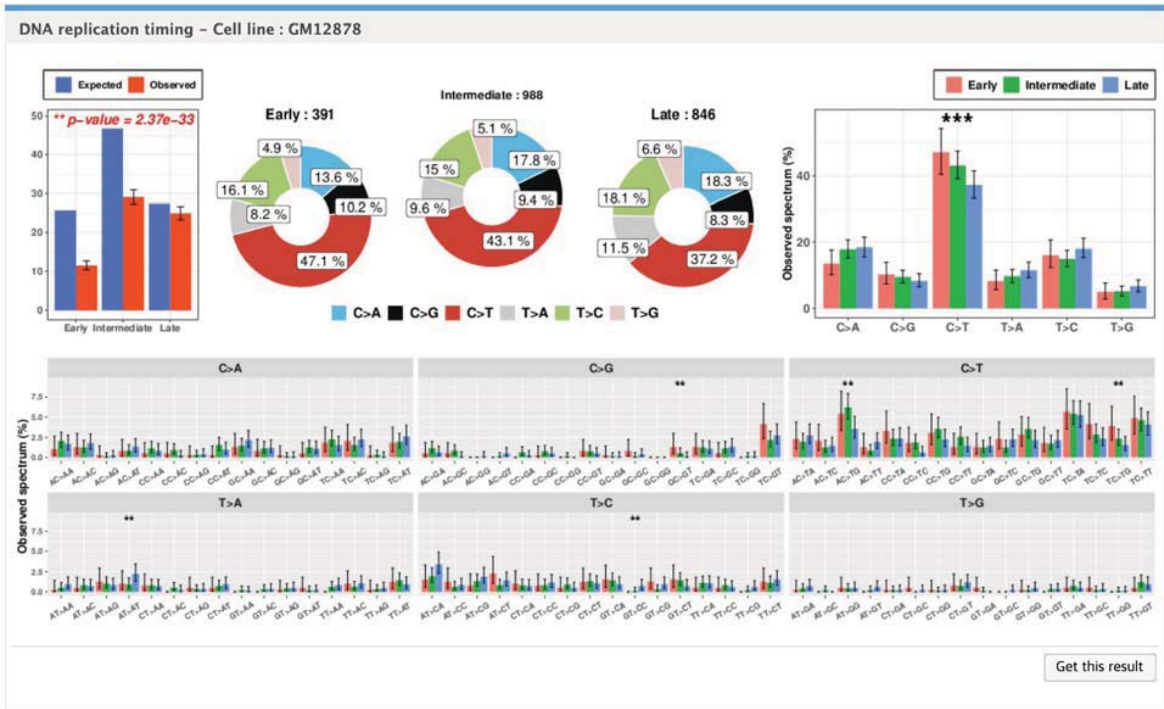
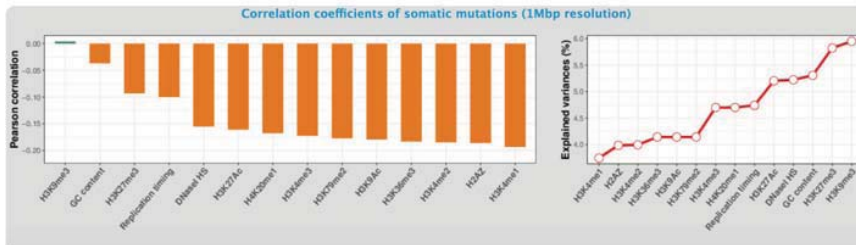
Mutational signatures



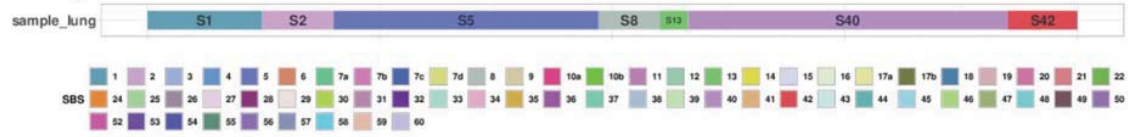
No.	Signatures	Probabilities	Cosine similarity	BIC	Confidence Interval
6	5 1 2 4 13 17	0.54363 0.22094 0.10903 0.09342 0.03298 0.00000	0.98400	28719.280	<input type="button" value="Close"/>



No.	Signatures	Probabilities	Cosine similarity	BIC	Confidence Interval
6	5 1 2 4 13 17	0.54363 0.22094 0.10903 0.09342 0.03298 0.00000	0.98400	28719.280	<input type="button" value="View"/>
7	5 1 2 4 13 6 17	0.53078 0.21571 0.11005 0.09363 0.03316 0.01000 0.00668	0.98400	28725.380	<input type="button" value="View"/>
4	5 1 2 13	0.63748 0.21879 0.10846 0.03527	0.98100	28732.020	<input type="button" value="View"/>
3	5 1 2	0.66229 0.21649 0.12121	0.97500	28778.940	<input type="button" value="View"/>
2	5 2	0.87389 0.12611	0.91700	29168.280	<input type="button" value="View"/>
1	5	1.00000	0.26800	29907.540	<input type="button" value="View"/>



Summary of the best results from our tool



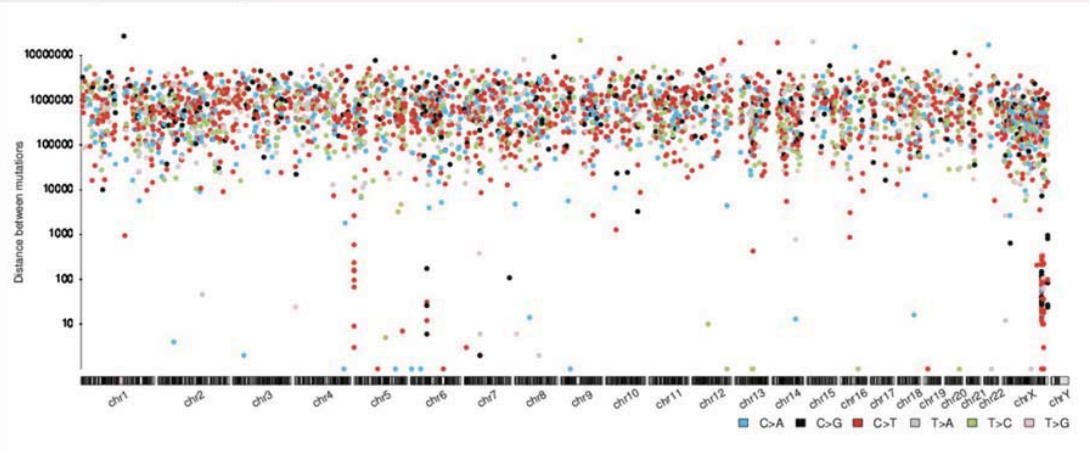
sample_lung

File : sample_lung

[Get merged results](#)

[Get all results at once](#)

Localized hypermutation (kataegis)



Mutational signatures



Interpretation of signature results

Proposed aetiology

Unknown.

Comments

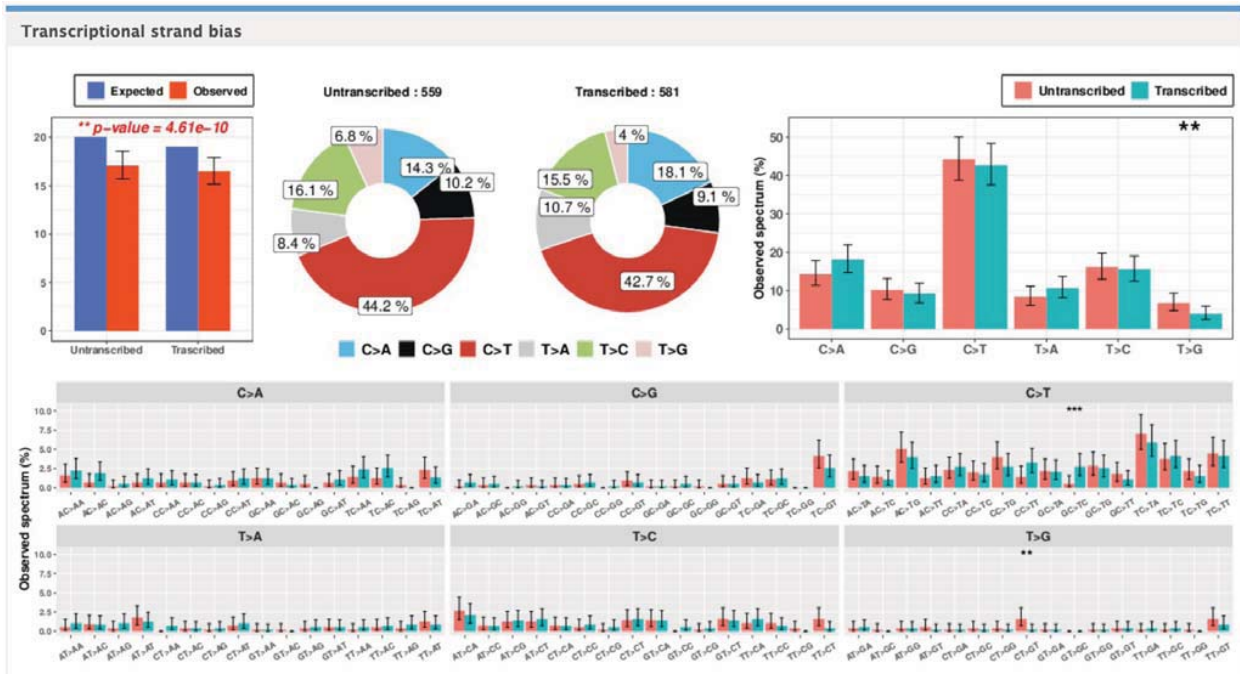
Numbers of mutations attributed to SBS40 are correlated with patients' ages for some types of human cancer.

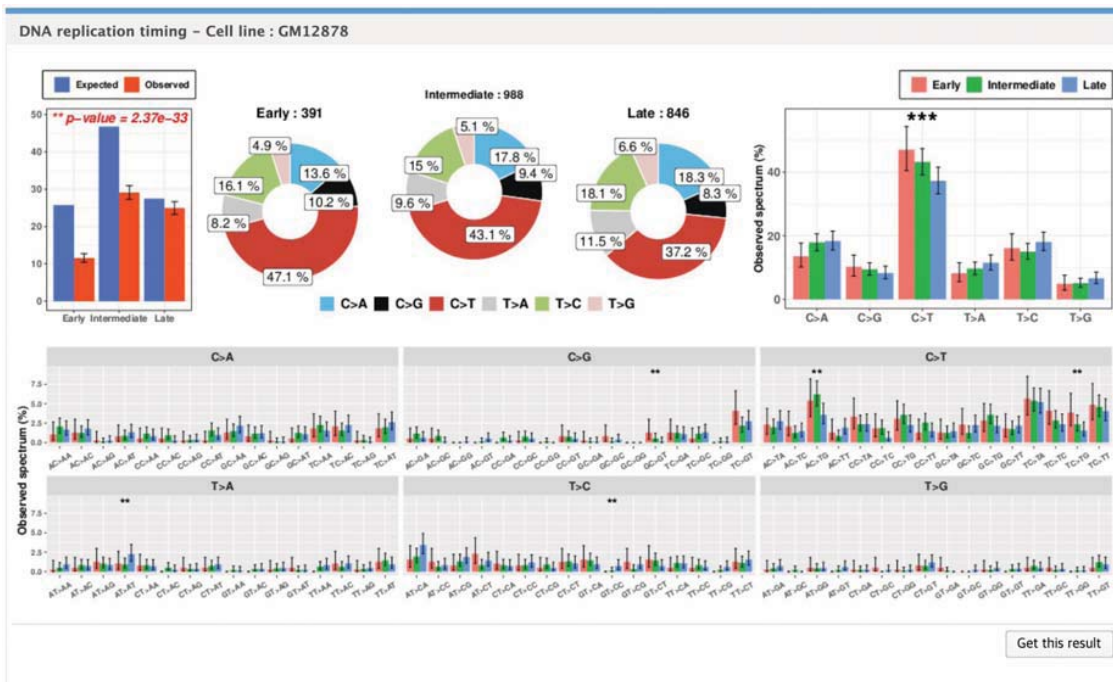
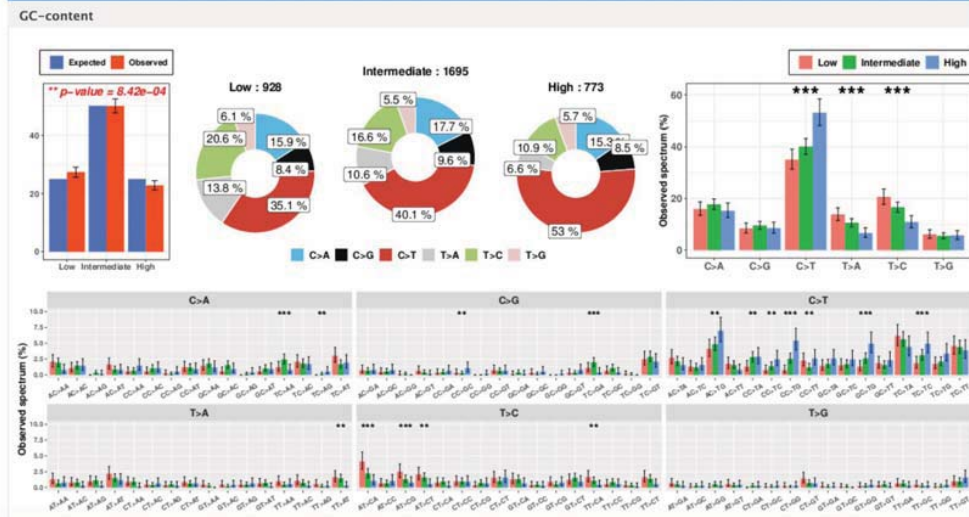
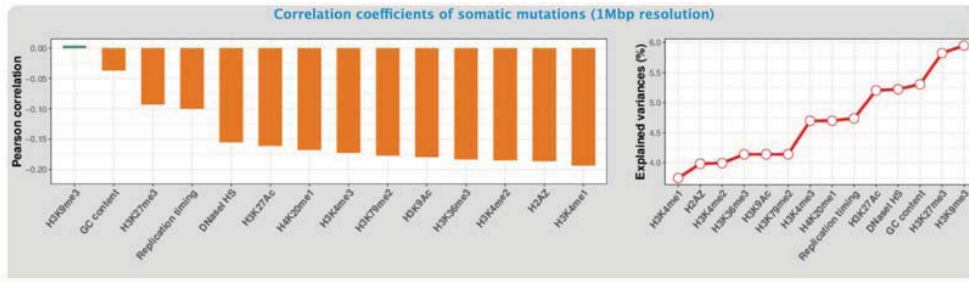
Proposed aetiology

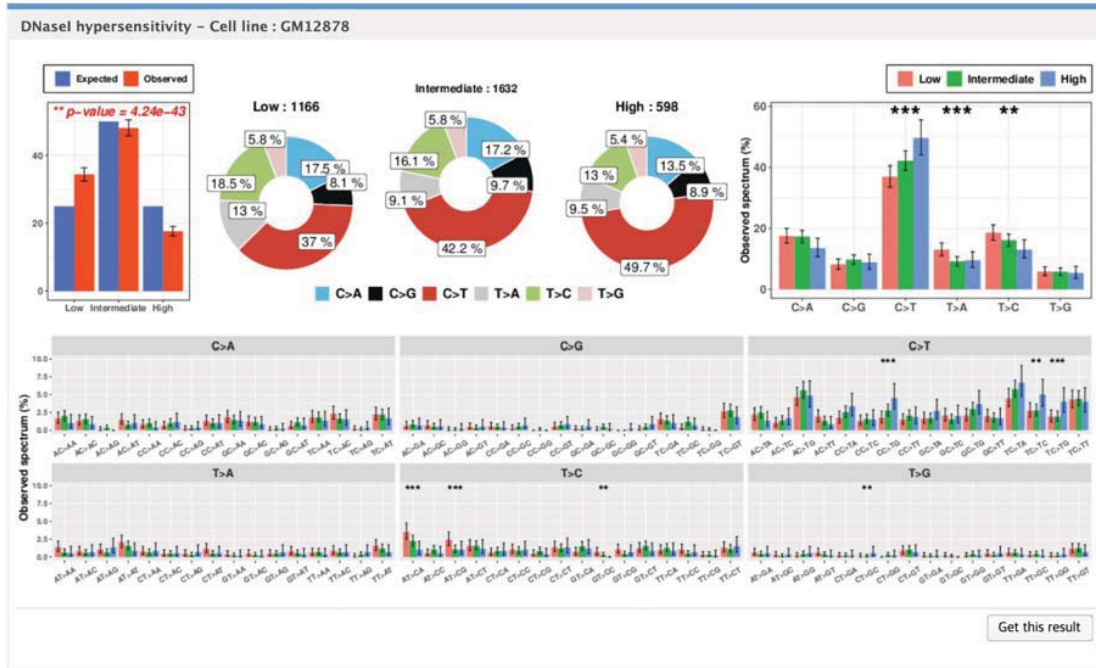
Unknown. SBS5 mutational burden is increased in bladder cancer samples with ERCC2 mutations and in many cancer types due to tobacco smoking.

Comments

SBS5 is clock-like in that the number of mutations in most cancers and normal cells correlates with the age of the individual. Rates of acquisition of SBS5 mutations over time differ between different cancer types and different normal cell types. These differences do not clearly correlate with estimated rates of stem cell division in different tissues nor with differences in SBS1 mutation rates. SBS5 may be contaminated by SBS16.







TO Part II

Artificial Intelligence Study using Cancer Big Data

AI

Deep-learning is the hottest topic in cancer genomic research area

Latest Cancer Research Trend Analysis - Top 15 Bigrams

~250,000 articles / 1 year (pubmed)

As of August 1, 2019



Rank	Keyword	2015	2016	2017	2018	2019	Cumulative growth ratio (2018/2015)
1	deep learning	127	284	790	2,107	2,285	16.69
2	pd1 blockade	0	6	8	9	9	10.00
3	circular rna	53	96	249	512	458	9.50
4	health-related quality	0	3	4	8	5	9.00
5	checkpoint inhibitor	99	239	488	784	679	7.85
6	circular rnas	80	140	308	612	567	7.57
7	mutational burden	30	61	107	225	230	7.29
8	blood-brain barrier	5	8	9	40	23	6.83
9	epstein-barr virus	2	1	5	18	8	6.33
10	next-generation sequencing	2	9	15	17	18	6.00
11	progression-free survival	4	5	10	28	20	5.80
12	checkpoint inhibitors	318	767	1,292	1,844	1,594	5.78
13	endogenous rna	82	119	239	467	463	5.64
14	receptor t	57	110	194	320	262	5.53
15	liquid biopsies	47	124	193	250	179	5.23

Latest Cancer Research Trend Analysis - Cancer Genomics

Keyword	2015	2016	2017	2018	2019	Cumulative growth ratio (2018/2015)
deep learning	127	284	790	2,107	2,285	16.47
nextgeneration sequencing	2	9	15	17	18	6.00
cancer database	213	274	515	685	598	3.21
machine learning	1,731	2,280	3,287	5,405	4,834	3.12
cancer genome	789	1,017	1,359	1,904	1,614	2.41
germline variants	71	85	117	159	143	2.22

AI ?

Strong AI = General AI

공상과학 영화



터미네이터



A.I



아이로봇



업그레이드

Weak AI = Narrow AI



Google 광고 개인 최적화



Apple Watch 부정맥 감지



TESLA
Tesla autopilot recognition

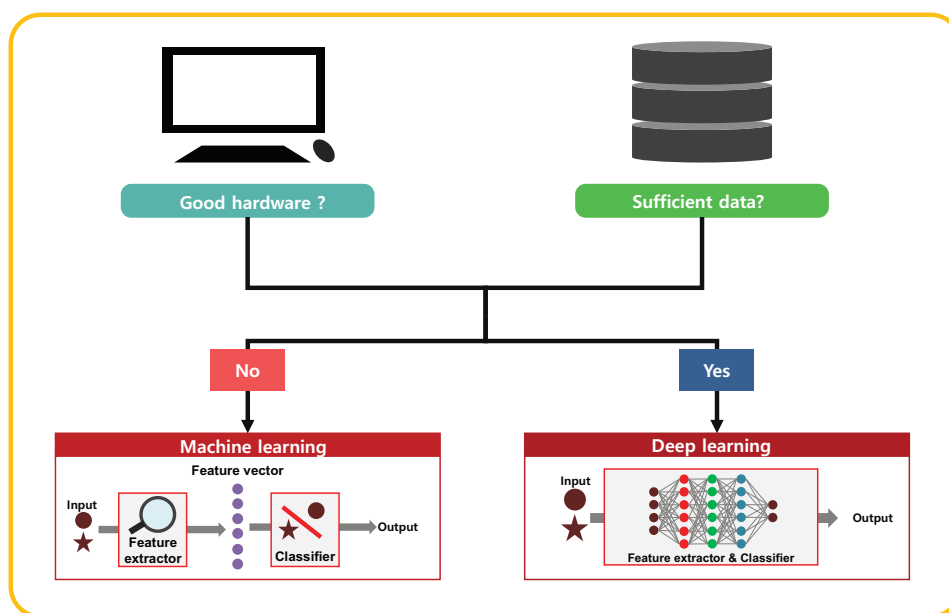
모든 부분에서 AR 가능한 로봇 또는 인간 수준 AI

특정 임무를 수행하는 AI

Cancer drugs

- Vemurafenib and trametinib
BRAF V600E mutations in melanoma
- Erlotinib and osimertinib
EGFR mutations in NSCLC; L858R, Del (19), T790M
- Pembrolizumab (immune checkpoint inhibitor): FDA approved
SOLID tumors from any tissue type
Mismatch repair deficiency (dMMR)
- Nivolumab, ipilimumab and atezolizumab
High tumor mutation burden

Which goal ?



컴퓨팅 파워의 발전

PS3
PlayStation 3



최소 출시: 2006년 11월 11일
최초 가격: \$499
CPU: 3.2 GHz 셀 브로드밴드 엔진 1 PPE, 8 SPEs
RAM: 시스템 RAM: 256MB (XDR DRAM)
비디오 RAM: 256MB (GDDR3)
GPU: 550 MHz 엔비디아/SCEI RSX 리얼리 신시사이저

PS3 1대 FLOPS = 230.4 GFLOPS

미공군 336대의 ps3 를 이용한 도시감시용 슈퍼컴퓨터

500 TFLOPS



*: Teraflops: 초당 10¹²회의 연산을 처리 할 수 있는 계산 능력

XBOX
SERIES X



최소 출시: 2020년 11월 10일
최초 가격: \$499
CPU: AMD ZEN 2 기반 커스텀 마이크로 아키텍처
제원: 8코어 16스레드, 3.6 GHz
RAM: 16 GB, GDDR6 ECC SGRAM
GPU: AMD RDNA 2 기반 커스텀 마이크로 아키텍처

XBOX 1대 FLOPS = 12.1472 TFLOPS



2001 슈퍼 컴퓨터 ACSI white **12.3 TFLOPS**
핵무기 관련 시뮬레이션용 사용 (농구장 2배 크기)



컴퓨팅 능력의 저변 확대

Data 분석용 workstation 성능 변화



CPU Intel Pentium III 1400S @ 1400MHz

Clockspeed 1.4 GHz

of Physical Cores 1 (Threads: 1)

CPU Mark **194**



CPU Intel Xeon Platinum 8268 @ 2.90GHz

Clockspeed 2.9 GHz

of Physical Cores 24 (Threads: 48)

CPU Mark **30,103**

15.48 배



인공 지능을 이용한 연구 환경 조성 ??

ex)

AlphaGo
AlphaGo Zero



3hours

19hours

70hours

$361! = 2.6 \times 10^{845}$

가장 큰 수의 단위: 10^{68} 무량대수
읽을 수 있는 수: 10^{71} 천무량대수

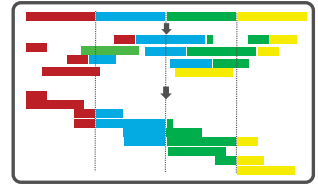
인간의 한계

오믹스 데이터

1) 일종의 디지털 데이터

A T G C
(30쌍)

2) 대량의 정보

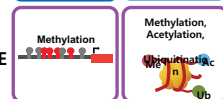


3) 복잡성

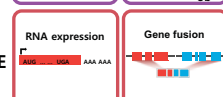
GENOME



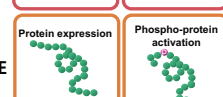
EPIGENOME



TRANSCRIPTOME

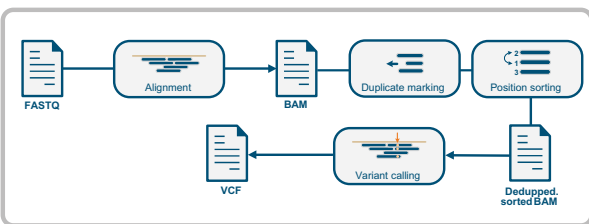


PROTEOME

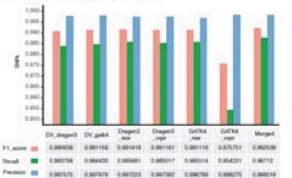


Pipeline 계산 능력

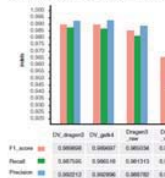
Germline variant calling pipelines



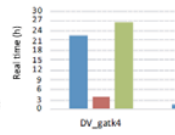
A. NA12878_PrecisionFDA - SNPs



B. NA12878_PrecisionFDA - indels



A. NA12878_PrecisionFDA



Sci Rep 2020 Nov 19;10(1):20222.

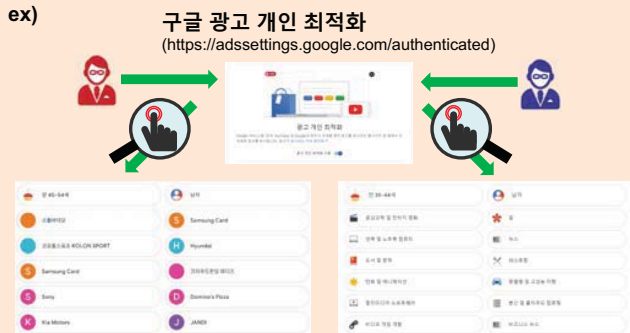
DRAGEN3 vs GATK4 : 오차 범위 내 동일 수준 정확성

동일 데이터 처리 (fastq to vcf) 기준;
9배(GATK4; 27시간, dragen; 3시간) 시간 소요

2.7. Machine learning 과 Deep learning

Machine learning

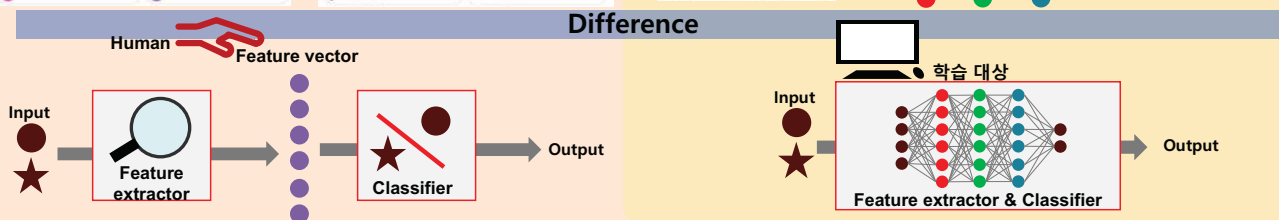
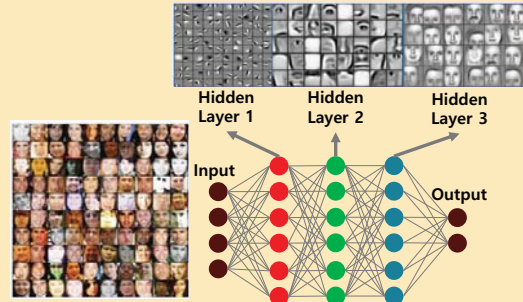
: 컴퓨터가 데이터를 통해 스스로 학습하여 예측이나 판단을 제공하는 기술



Deep learning

: 깊은 인공신경망을 알고리즘을 활용하는 머신러닝 기술

ex) 딥러닝 알고리즘을 활용한 얼굴 인식 프로세스



Three types of **research questions** driving application of **AI** in **genomics**
(Ching *et al.*, J R Soc Interface 2018)

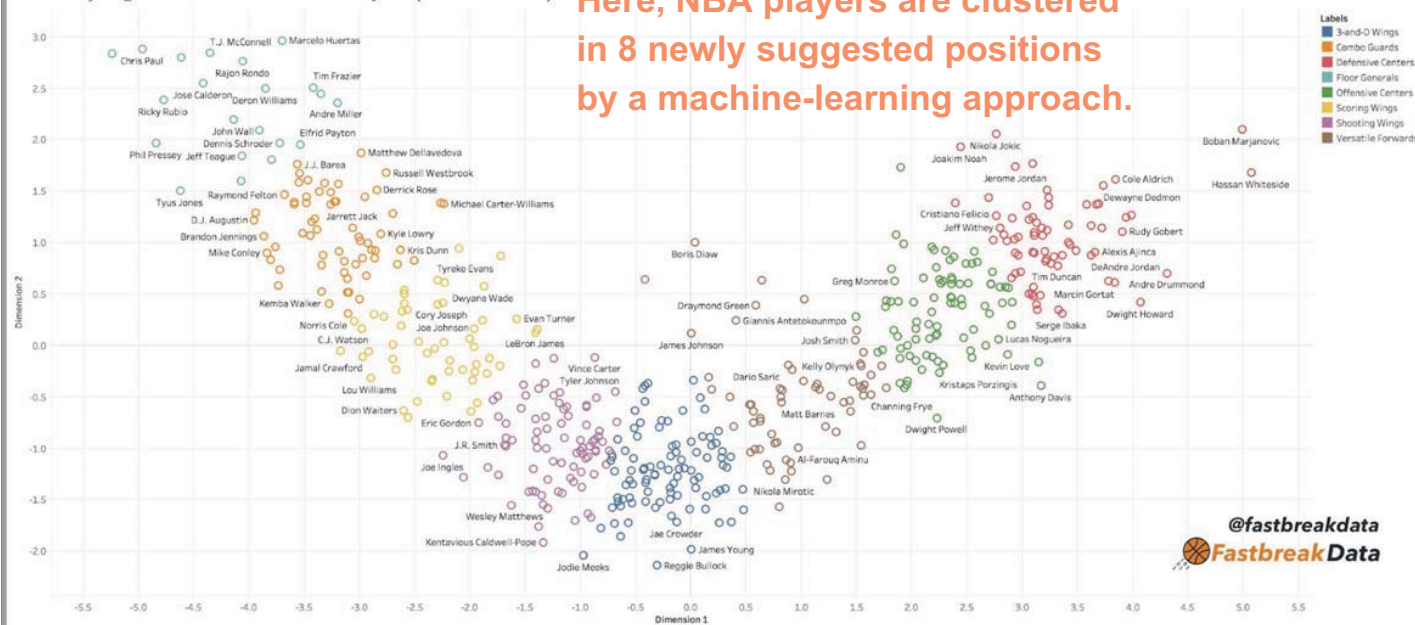
1. Disease and patient categorization
(**Gene expression, DNA methylation**)
2. Functional and biological study (DNA sequences)
3. Treatment of patients (Gene expression)

Three types of research questions driving application of AI in genomics (Ching *et al.*, J R Soc Interface 2018)

1. Disease and patient categorization
2. Functional and biological study
3. Treatment of patients

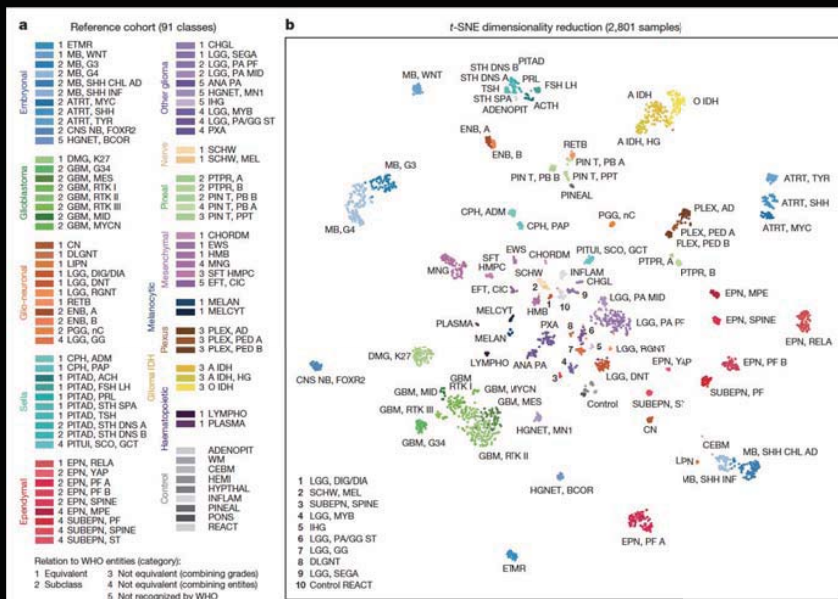
Classifying the Modern NBA Player (2014-2017)

Here, NBA players are clustered in 8 newly suggested positions by a machine-learning approach.



@fastbreakdata
Fastbreak Data

Machine Learning Research in Genomics



Capper *et al.*, Nature 2018
 Clustering of central nervous system (CNS) tumors based on DNA methylation data

Over 100 World Health Organization (WHO) CNS tumor subtypes

17 histological types

Three types of research questions driving application of AI in genomics (Ching *et al.*, J R Soc Interface 2018)

1. Disease and patient categorization
2. **Functional and biological study**
3. Treatment of patients

Different Tissues & Physiology



Caterpillar



Butterfly

Different Tissues & Physiology, Same Genome

Nucleotide sequences

```
...CGCCGCTGACCTATCATCAGTTC  
CAAGCGCTGATAGCGAGCATGCC  
CCCGCCTCCGTCCGCCGAACCCA  
CCATCAGTTTGGAGACTCAACC  
GCGCCGTTACACCTATCTCGGATA  
ATCAGGACGAACGATTTGGAGTGC  
CGACTCGAAGAACTTGGCTTCG  
ATACGGAAGGTCTTAAACCTCCAA  
TATGGATCGGCCGAGAAAACGAA  
GCTCTGTTGAGACT...
```

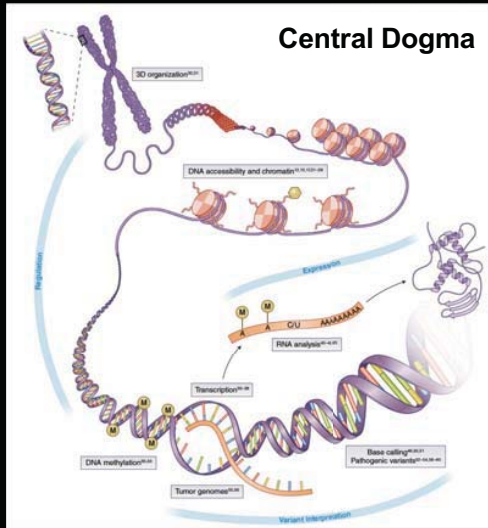
Caterpillar

=

```
...CGCCGCTGACCTATCATCAGTTC  
CAAGCGCTGATAGCGAGCATGCC  
CCCGCCTCCGTCCGCCGAACCCA  
CCATCAGTTTGGAGACTCAACC  
GCGCCGTTACACCTATCTCGGATA  
ATCAGGACGAACGATTTGGAGTGC  
CGACTCGAAGAACTTGGCTTCG  
ATACGGAAGGTCTTAAACCTCCAA  
TATGGATCGGCCGAGAAAACGAA  
GCTCTGTTGAGACT...
```

Butterfly

Many, So Many Components in Central Dogma



Zou *et al.*, Nat Genetics 2019

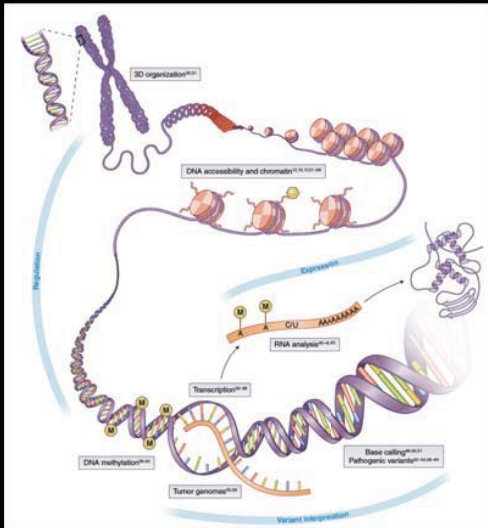


Current Problems in Applying Deep Learning to Genomics

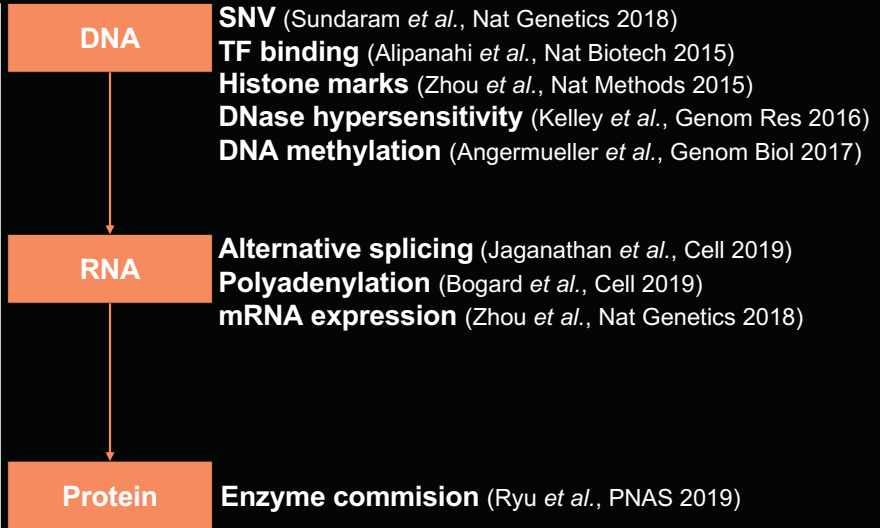
To conduct optimal learning, we need n^2 samples for n features for correlated features
(Hua *et al.*, Bioinformatics 2005)



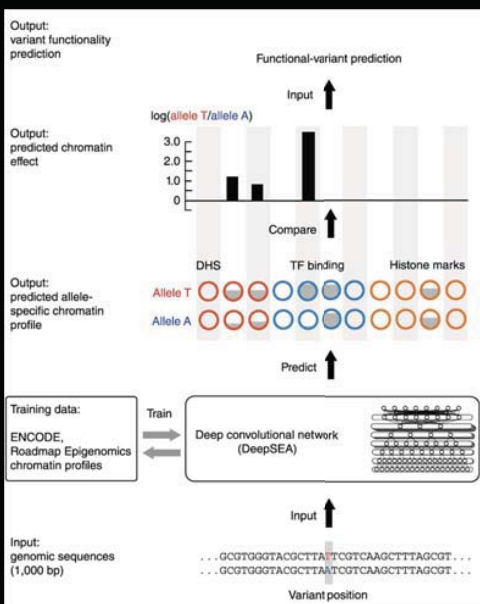
Deep Learning Research in Genomics



Zou *et al.*, Nat Genetics 2019

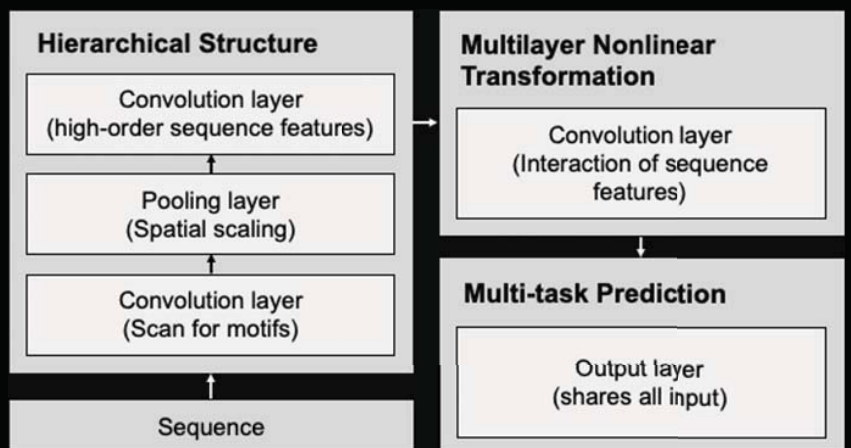


Deep Learning Research in Genomics - DeepSEA



DeepSEA (Zhou *et al.*, Nat Methods 2015)

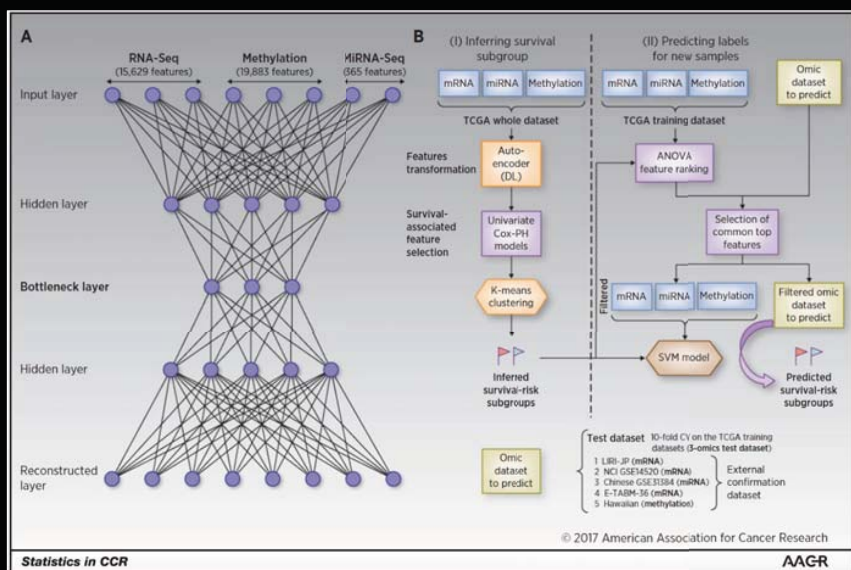
Using genomic sequences (1,000 bp) predict chromatin organization (transcription factor binding, histone marks, DNase sensitivity).



Three types of research questions driving application of AI in genomics (Ching *et al.*, J R Soc Interface 2018)

1. Disease and patient categorization
2. Functional and biological study
3. **Treatment of patients**

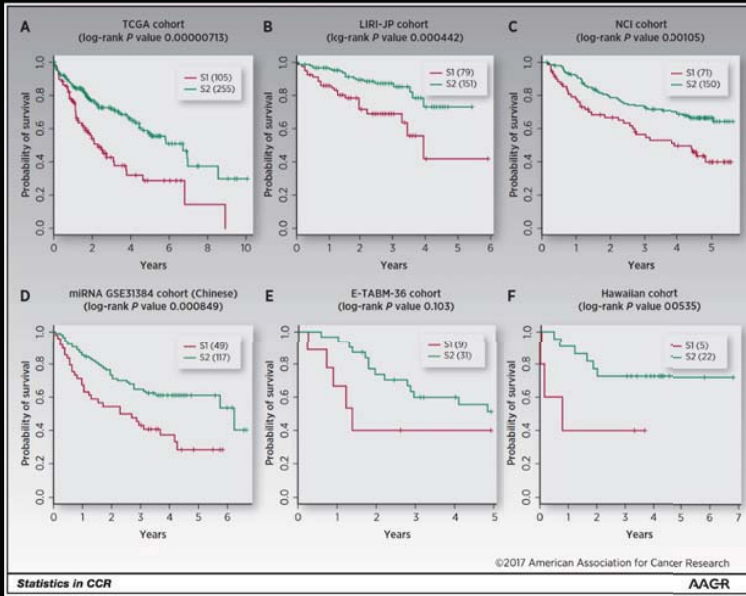
Deep Learning Research in Genomics - Predicting Prognosis



Chaudhary *et al.*, Clin Can Res 2018

- Predicting survival of liver cancer patients by training an **autoencoder**.
- Trained on 360 TCGA-HCC samples' **mRNA** and **miRNA sequencing** as well as **methylation data**

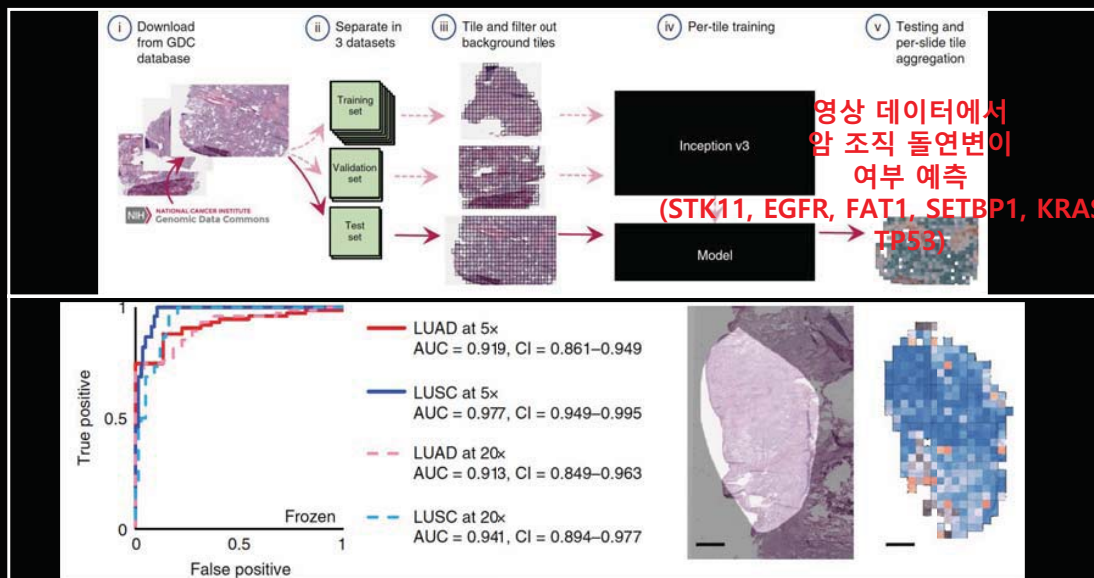
Deep Learning Research in Genomics - Predicting Prognosis



- Validated the deep learning model on 684 patients across 5 independent cohorts

Chaudhary *et al.*, Clin Can Res 2018

Deep Learning Research in Biomedicine



Coudray *et al.*, Nature Medicine 2018

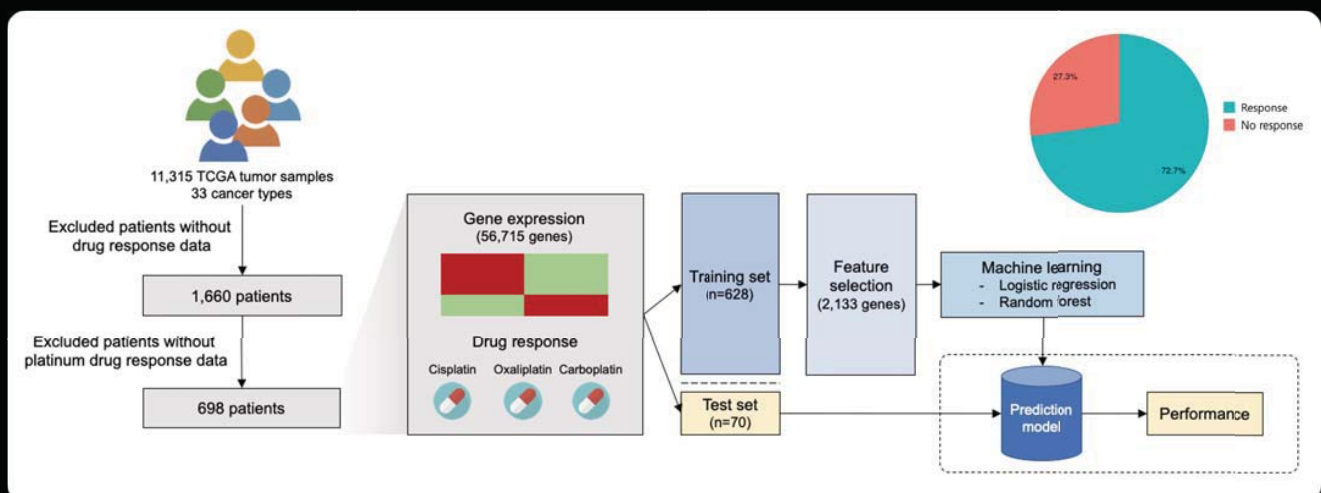
Use machine learning algorithm



ARBITR: An aRtificially intelligent Bayesian approach to predicting predisposition and evoluTionary deteRminants in human cancer

Challenges in Cancer Genomic Data for Deep Learning

Our pilot study predicting platinum therapy response using omics data





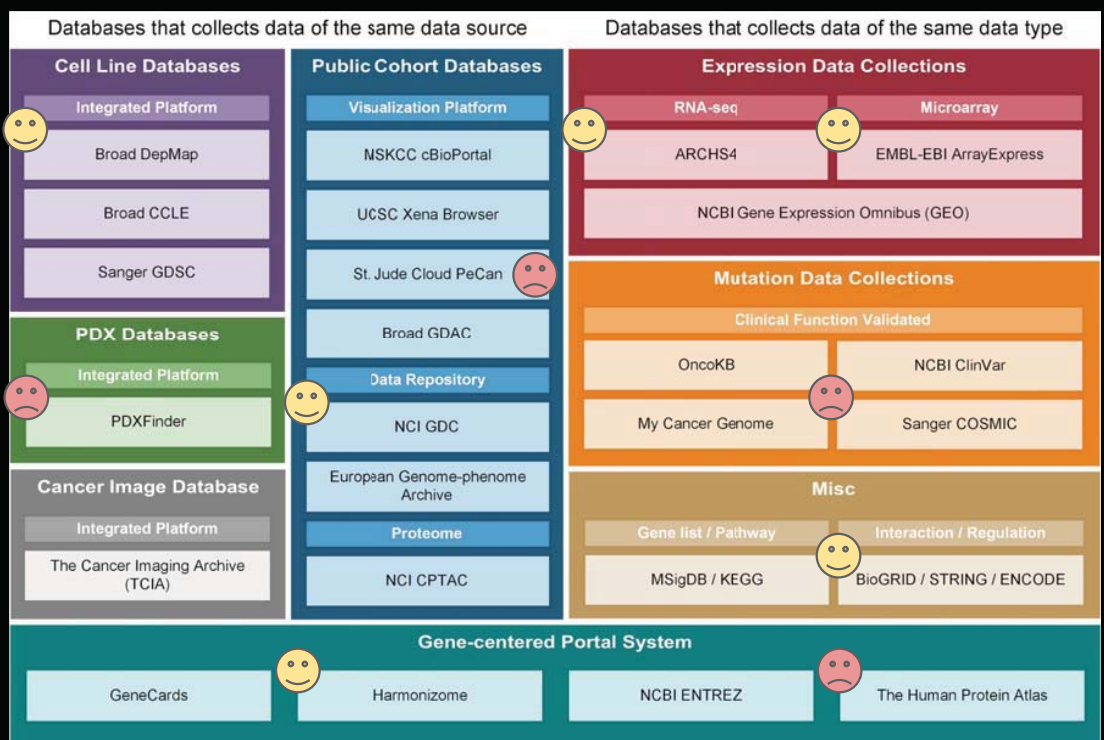
Challenges in Cancer Genomic Data for Deep Learning

Our pilot study predicting platinum therapy response using omics data

Method/Model	AUC	
Random forest	0.619*	Genomic/transcriptomic features (selection) matter a great deal. Due to currently limited sample size in publicly available datasets, we must devise new ways of tackling the problem of predicting therapeutic response in cancer patients
SVM	0.612	
Ada boost	0.609	So why are genomic/transcriptomic features problematic (besides the dimensionality problem)? Because of tumor heterogeneity Dagogo-Jack <i>et al.</i> , Nat Rev Clin Oncol 2018
Logistic regression	0.593	
MLP	0.593	

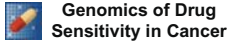
Current Problems in Applying Deep Learning to Genomics

-  Suitable for deep-learning
-  Not suitable for deep-learning



Cancer Omics Data

Cell line data



<https://www.cancerrxgene.org/>

Summary

	GDSC1	GDSC2
Cell line	987	809
Compounds	367	198
IC50	310,904	135242

Methods & dataset

Exome sequencing
 Array exome sequencing
 Mutation
 Copy Number
 Methylation
 Expression
 Drug Screening – IC50s



<https://portals.broadinstitute.org/ccle>

Summary

	CCLE
Cell line	1,457
Compounds	24 (504 cell line test)

Dataset

WES
 Mutation
 Fusion
 structural variant
 miRNA
 Global Chromatin
 RPPA
 Antibody
 RPKM
 RSEM
 Methylation
 copy number
 metabolomics

Cancer Omics Data

Human genomic data



<https://portal.gdc.cancer.gov/>

Summary

	GDC
Projects	67
Primary sites	68
Cases	84,392
Genes	23,399
Files	596,758
Mutations	3,287,299

Method

Clinical data
 Biospecimen data
 Pathology Reports
 SNP microarray
 Copy number microarray
 Low-Pass DNA Sequencing
 Whole exome
 Whole genome
 SNP microarray
 Sequence trace
 Diagnostic image
 Tissue image
 Radiological image
 Bisulfite sequencing
 Bead array
 miRNA Sequencing
 Total RNA Sequencing
 Microarray
 Reverse-Phase Protein Array



<https://dcc.icgc.org/>

Summary

	ICGC
Data release 28 cancer project	86
Cancer primer sites	22
Donor with molecular data in DCC	22,230
Total donors	24,289
Simple somatic mutations	81,782,588

Method

Clinical data
 Biospecimen Data
 WXS
 WGS
 RNA-Seq
 miRNA-seq
 Bisulfite-seq



<https://cptac-data-portal.georgetown.edu/>

Summary

	ICGC
Studies	55
Tumor sites	12
Cases	2,549
Samples	3,639
Files	107,493
Data	25,576GB

Method

Proteome
 Phosphoproteome
 Acetylome
 Glycoproteome
 Ubiquitylome

Issues to be considered

데이터 표기의 문제

▶ 각각의 데이터 표기 기준 안에 따른 데이터 구성의 문제 발생

◇ 약품 표기법



◇ 의학용어 표기법



ex)1. IL2 = IL-2 = interleukin - 2 = interleukin-2
2. typing error

Table S1: Correcting drug names from TCGA to standard names

Recorded Name from TCGA	DrugBank ID	Standard drug name
IL2	DB00041	Aldesleukin
IL-2	DB00041	Aldesleukin
Interleukin - 2	DB00041	Aldesleukin
Interleukin-2	DB00041	Aldesleukin
Alvesin	NA	NA
Anastrozole	DB01217	Anastrozole
ANASTROZOLE	DB01217	Anastrozole
Arimidex	DB01217	Anastrozole
ARIMIDEX	DB01217	Anastrozole
PF-04605412	NA	anti-ASB1 integrin monoclonal antibody PF-04605412
MORAb-004	NA	anti-endostalin/TEM1 monoclonal antibody MORAb-004?
autologous vaccine	NA	autologous vaccine
Axitinib	DB06626	Axitinib
Axitinib	DB06626	Axitinib
Cediranib	DB04849	AZD2171
Bacillus Calmette-Guerin (BCG)	NA	BCG
BCG	NA	BCG
avastin	DB00112	Bevacizumab
Avastin	DB00112	Bevacizumab
bevacizumab	DB00112	Bevacizumab
Bevacizumab	DB00112	Bevacizumab
BEVACIZUMAB	DB00112	Bevacizumab
Bicalutamide	DB01128	Bicalutamide
casodex	DB01128	Bicalutamide
Casodex	DB01128	Bicalutamide
bleomycin	DB00290	Bleomycin
Bleomycin	DB00290	Bleomycin
BRAF inhibitor	NA	BRAF inhibitor
cabazitaxel	DB06772	Cabazitaxel
jeviana	DB06772	Cabazitaxel
Cabozantinib	DB08875	Cabozantinib
Cancer Vax	NA	Cancer Vax
capecitabine	DB01101	Capecitabine
Capecitabine	DB01101	Capecitabine

Bioinformatics, 32(19), 2891-2895.



Preliminary study - methodology

AutoEncoder + 예측 모델



중요 특징 차원 축소

강화학습 + 예측 모델
(Supervised Reinforcement Learning)



중요 특징 선택

지도 학습
(Supervised Learning)

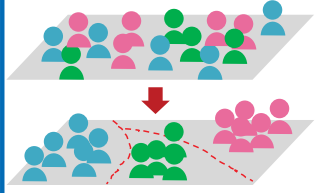
문제와 정답을 모두 알려주고 공부시키는 방법



예측, 분류

비지도 학습
(Unsupervised Learning)

답을 가르쳐주지 않고 공부 시키는 방법



연관 규칙, 군집

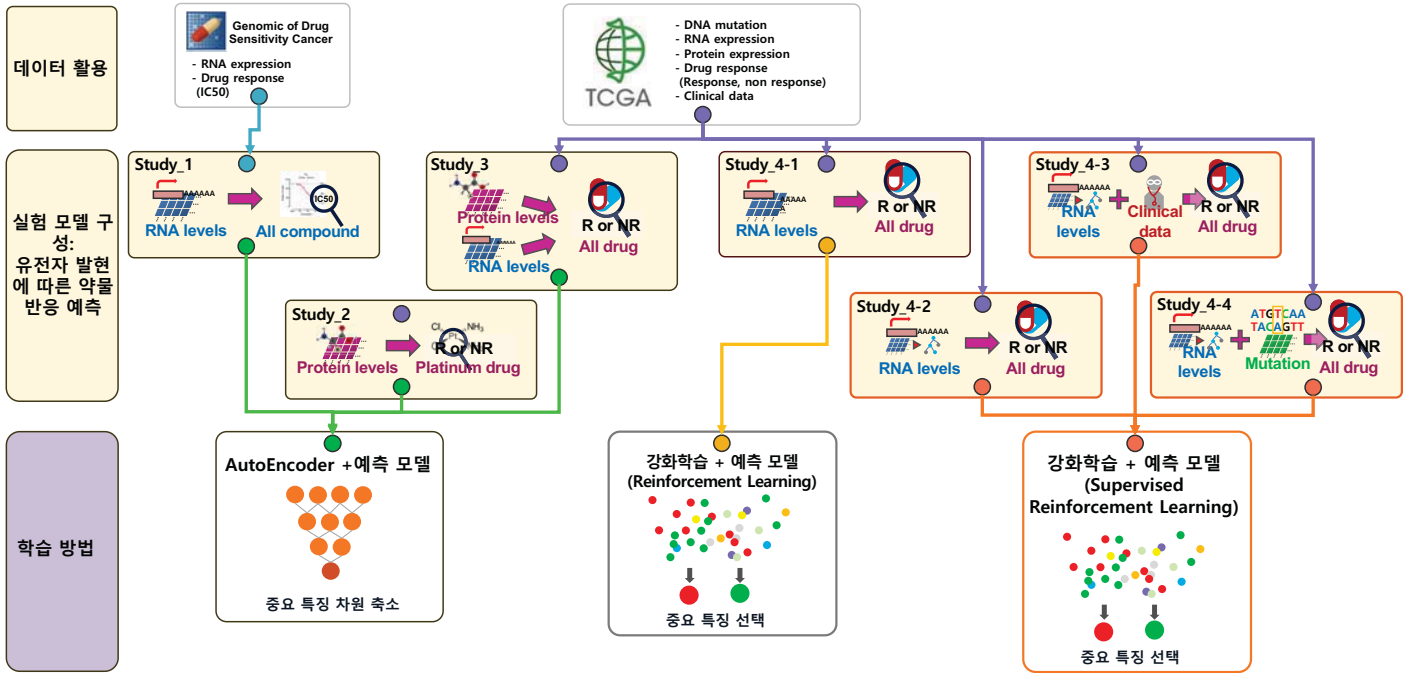
강화 학습
(Reinforcement Learning)

보상을 통해 상을 최대화, 벌은 최소화하는 방향으로 행위를 강화하는 학습

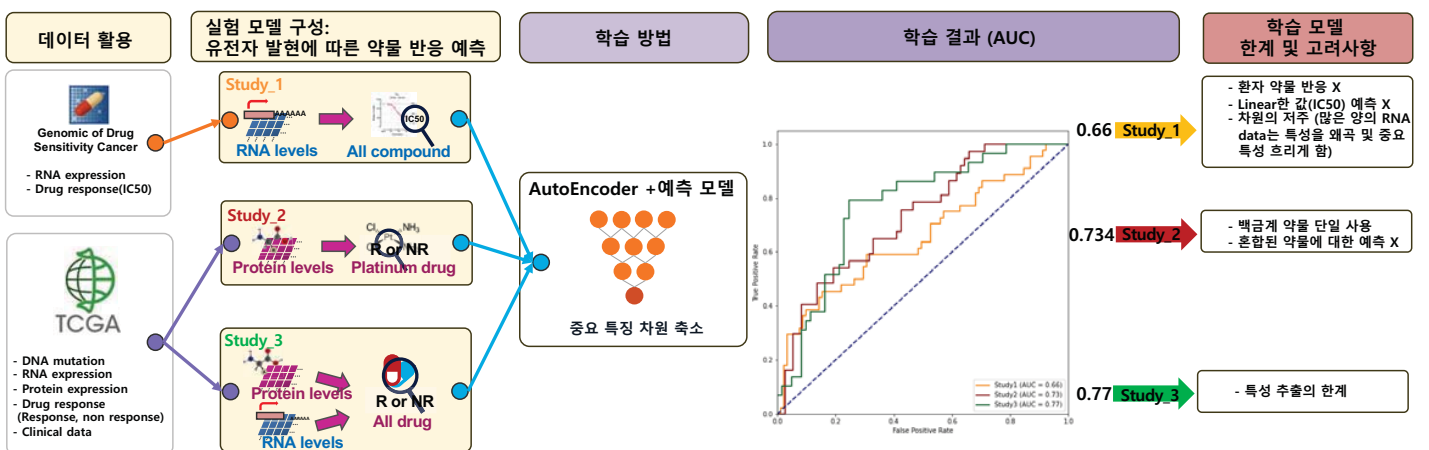


보상

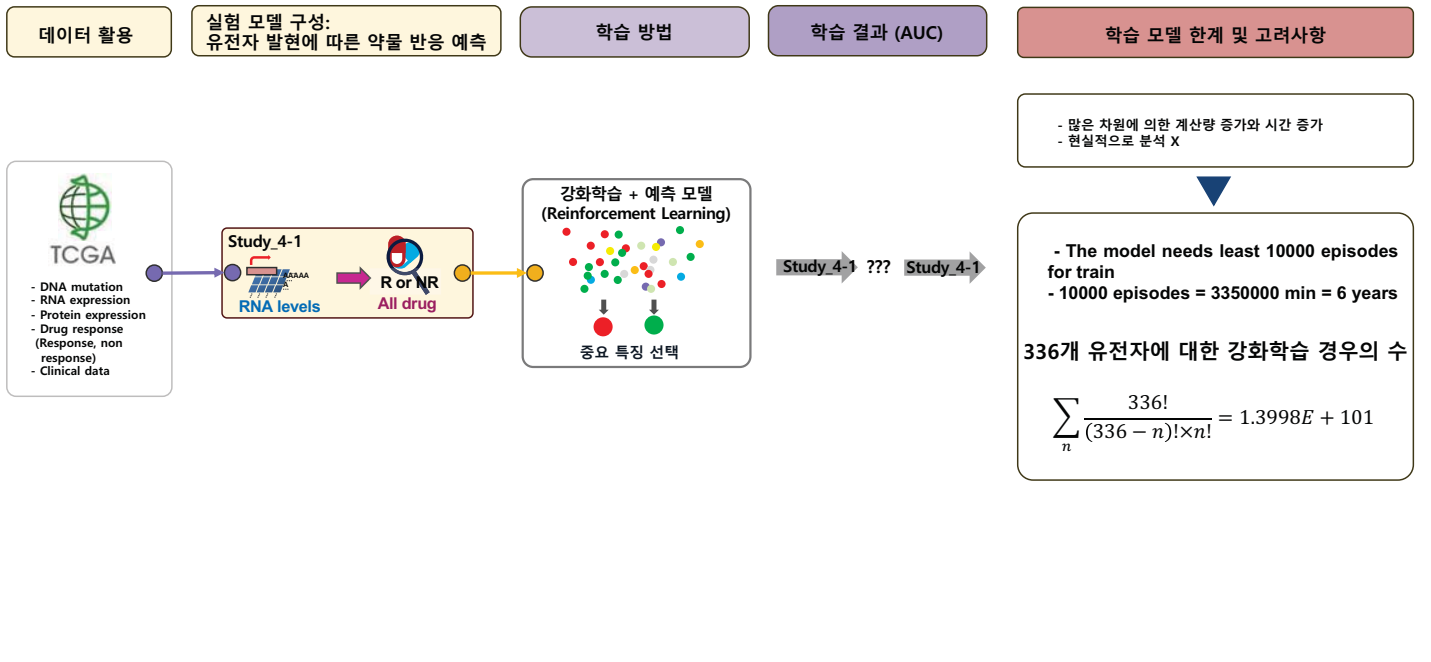
Preliminary study



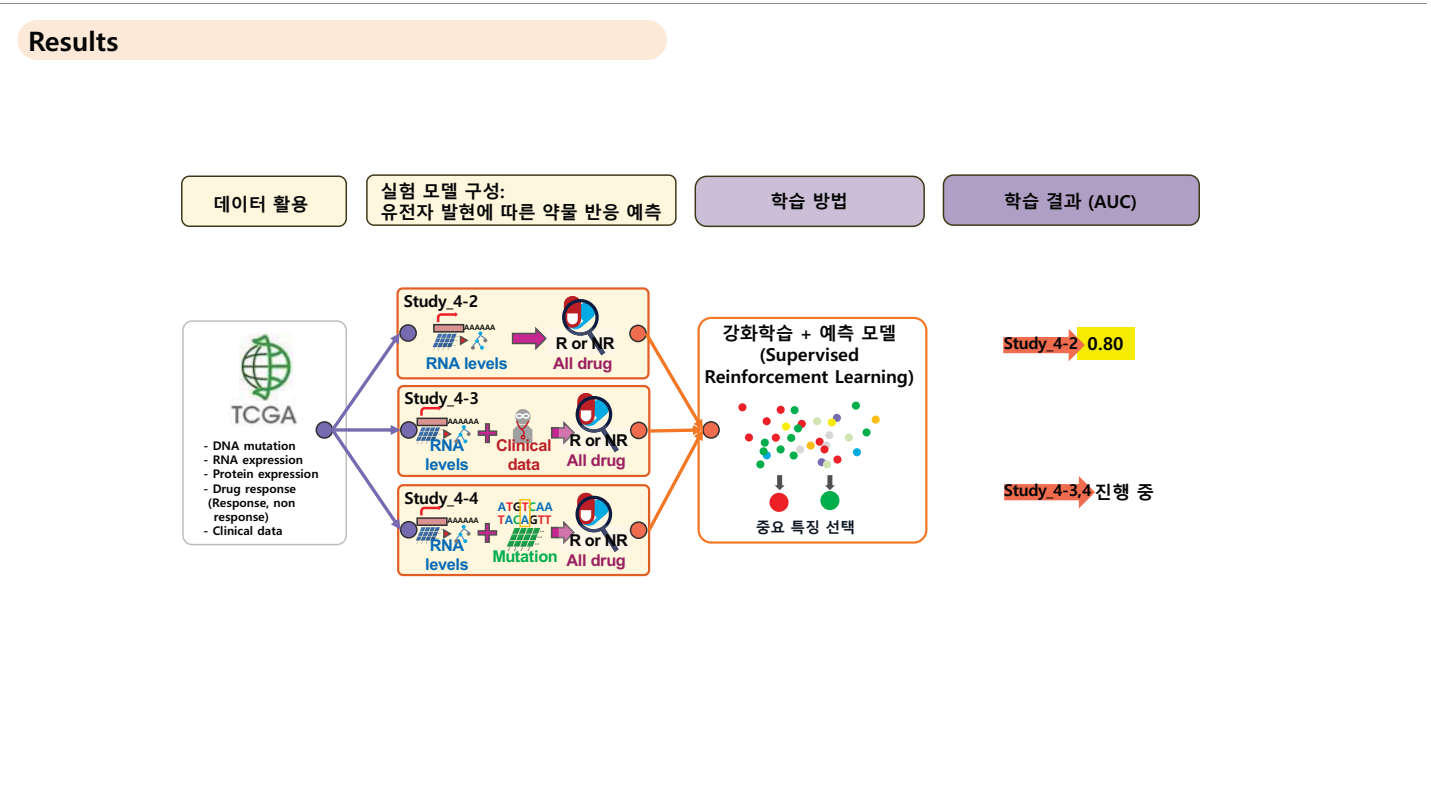
Results



Results



Results



Results

학습 결과

학습 모델
한계 및 고려사항

Study_4-2 결과

Cancer type	Sample size (R:NR)	Total gene (20531 genes)		KEGG cancer pathway (525 genes)		COSMIC cancer gene census (683 genes)	
		AUROC	Loss	AUROC	Loss	AUROC	Loss
		PANCAN	1794 (42:58)	0.791	0.795	0.782	0.875
BRCA	278 (15:85)	0.902	0.579	0.913	0.722	0.910	0.624
LGG	176 (80:20)	0.939	0.480	0.908	0.591	0.840	0.930
STAD	150 (35:65)	0.586	2.651	0.620	1.598	0.584	0.941
HNSC	59 (25:75)	0.989	0.10	0.958	0.225	0.989	0.139
KIRC	18 (89:11)	0.750	21965.61	0.750	8.888	0.750	12.087
KIRP	11 (73:27)	0.333	141.36	1.000	0.141	1.000	0.041
KICH	8 (87.5:12.5)	0.500	32591.15	0.500	13.325	0.500	17.119
LUAD	112 (47:53)	0.752	1.143	0.811	0.921	0.750	1.016
PRAD	43 (33:67)	0.593	1.230	0.704	1.371	0.580	1.827

- 특성 추출을 위한
여러 조건 연구

Study 4-2

Results

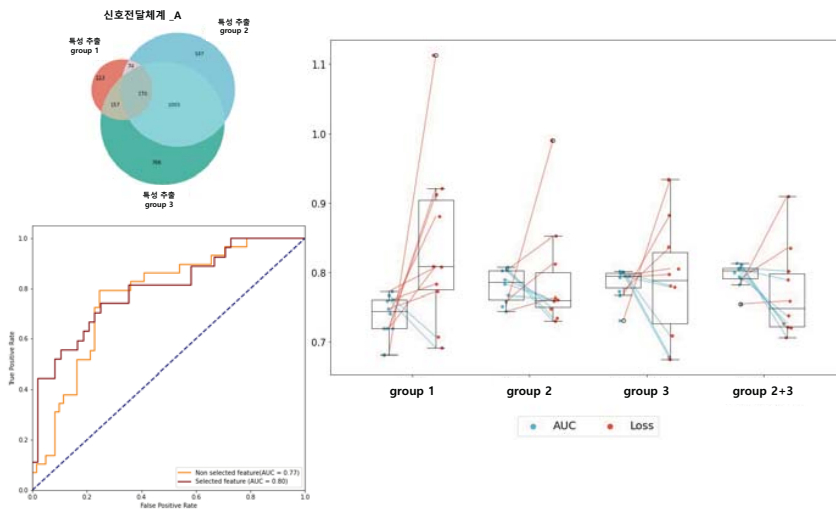
학습 결과

개선점

학습 모델
한계 및 고려사항

Study_4-2 결과

- 암세포의 신호전달체계 중 항암제 관련 신호전달체계 선택 후, 특성 추출
- 추출된 특성으로 예측모델 수행



- 항암제 관련 신호전달체계에서 특성 추출 후 예측모델 수행 시 학습 능력 향상되었고 (AUC=0.80) 예측의 불확실 정도가 개선됨.

- 특성 추출 고려
- 예를 들어, cell signaling, cellular processes 등

Study 4-2

Results

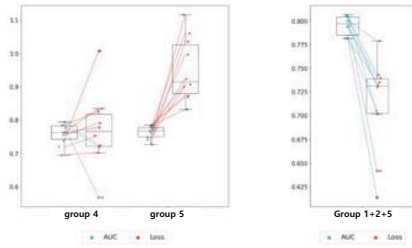
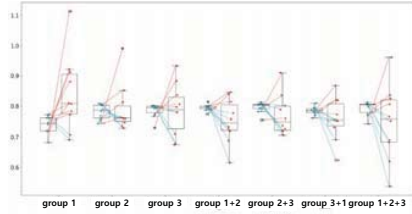
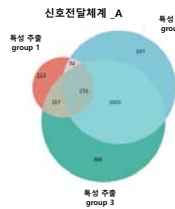
학습 결과

개선점

학습 모델 한계 및 고려사항

Study_4-2 결과

- 암세포의 신호전달체계 중 항암제 관련 신호전달체계 선택 후, 특성 추출
- 추출된 특성으로 예측모델 수행



- 항암제 관련 신호전달체계에서 특성 추출 후 예측모델 수행 시 학습 능력 향상되었고 (AUC=0.80) 예측의 불확실 정도가 개선됨.

- 특성 추출 고려
- 예를 들어, cell signaling, cellular processes 등

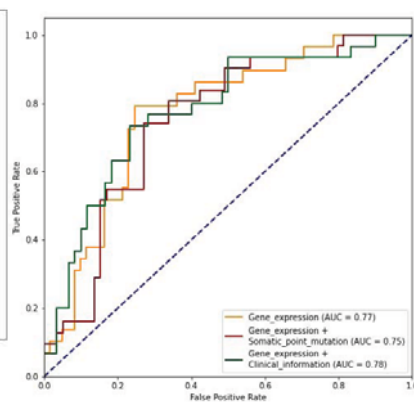
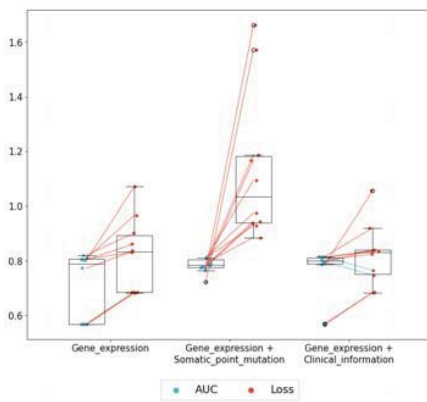
Study 4-2

Results

학습 결과

개선점

학습 모델 한계 및 고려사항



- 특성 중 항암제 관련성이 반영되지 않는 임상정보 및 돌연변이를 사용했을 경우 학습 능력 (+돌연변이: 0.78, +임상정보: AUC=0.75)이 떨어지고 예측의 불확실 정도가 나쁨.
- Layer 마다 특성을 고려해야 함.

- Layer 병합을 이용한 실험 모델에서는 각각 특성 추출 고려
- Layer와 layer 사이 관계 이해

Study 4-3,4

알파폴드를 이용한 단백질 구조 예측 및 평가 I

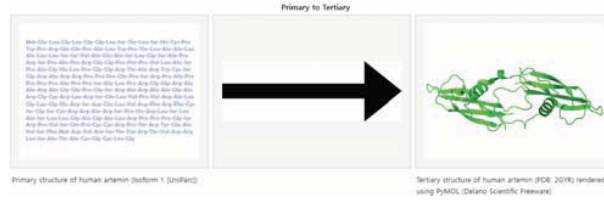
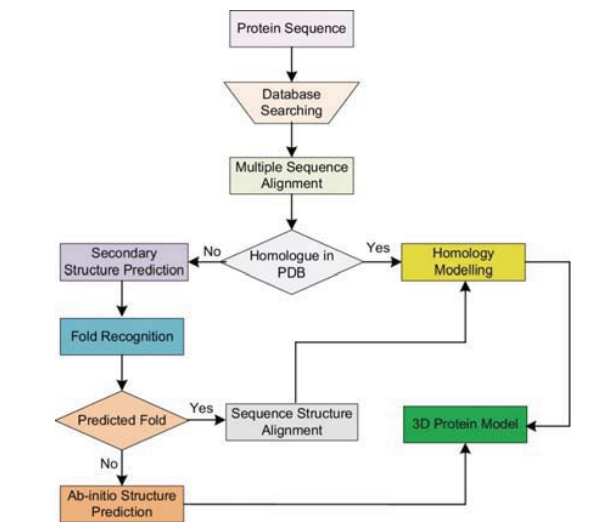
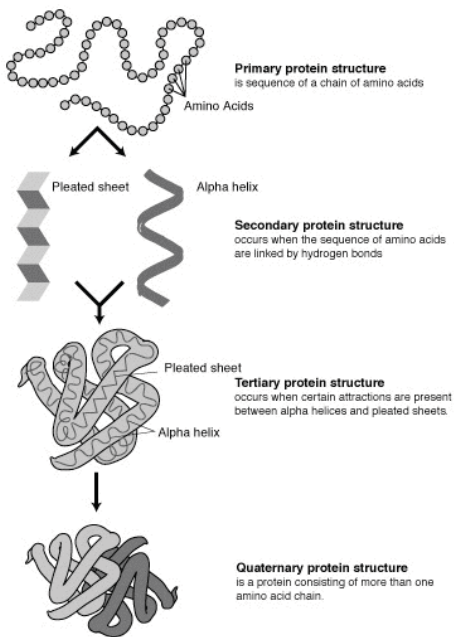
Contents

- 1 **인공 지능 및 AlphaFold의 소개**
- 2 **Google Colab 환경에서의 AlphaFold 사용법**
- 3 **AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측**

1

인공지능 및 AlphaFold의 소개

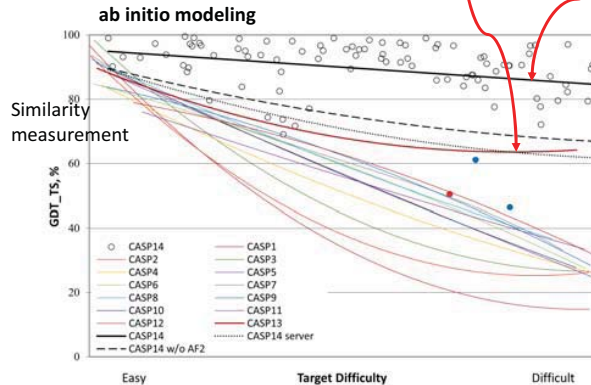
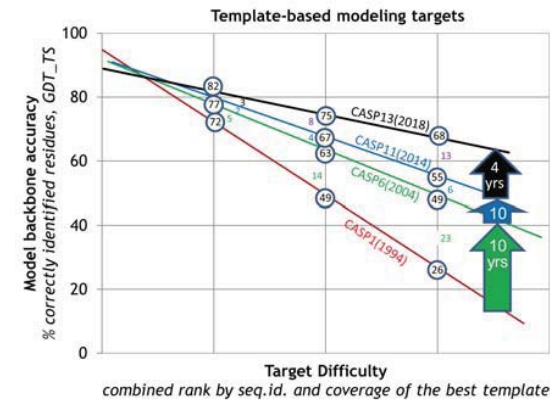
1 AlphaFold



1 AlphaFold

DeepMind AlphaFold & AlphaFold 2

1994
Critical Assessment of
protein Structure Prediction
(CASP)



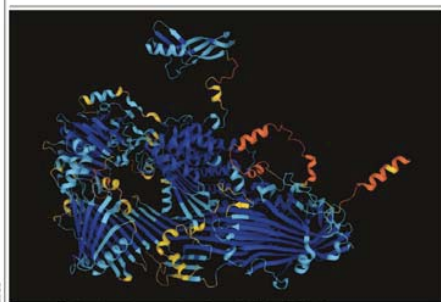
2018 CASP13 AlphaFold1
2020 CASP14 AlphaFold2

GDT; Global distance test



1 AlphaFold

The world this week
News in focus



The structure of the vitellin protein – a precursor of egg yolk – as predicted by the AlphaFold tool.

'THE ENTIRE PROTEIN UNIVERSE': AI PREDICTS SHAPE OF NEARLY EVERY KNOWN PROTEIN

DeepMind's AlphaFold tool has determined around 300 million protein structures, which are now available to scientists in database.

By Ewen Callaway

Determining the 3D shape of almost any protein known to science is now as simple as typing in a Google search. Researchers have used AlphaFold, the revolutionary artificial intelligence (AI) network – to predict the structure of more than 200 million proteins from some 1.6 billion species, covering almost every known protein on the planet. On 28 July, the data dump was made available for free in a database set up by DeepMind – that London-based AI company, owned by Google, that developed AlphaFold – and the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), an intergovernmental organization near Cambridge, UK. "Essentially you can think of it covering the entire protein universe," DeepMind chief executive Demis Hassabis said at a press briefing. "We're at the beginning of a new era of digital biology."

The 3D shape, or structure, of a protein is what determines its function in cells. Most drugs are designed using structural information, and the creation of accurate maps of proteins' amino acid arrangement is often the first step in making discoveries about how proteins work. DeepMind developed the AlphaFold network using an AI technique called deep learning, and the AlphaFold database was launched a year ago with more than 350,000 structure predictions covering nearly every

Nature | Vol 608 | 4 August 2022

News in focus

proteins made by humans, mice and 19 other widely studied organisms. Over the months that followed, the catalogue grew to more than 1 million structures.

"We're tracing our way back for the release of this huge trove," says Christine Orengo, a computational biologist at University College London, who has used the AlphaFold database to identify new families of proteins. "Having all the data predicted for us is just fantastic."

The release of AlphaFold last year made a splash in the life-sciences community, whose members have since been scrambling to use the tool. The network produces highly accurate predictions of many proteins' structures. It also provides information about the accuracy of its predictions, so researchers know whether they can be relied on. Conventionally, scientists have needed to use time-consuming and costly experimental methods such as X-ray crystallography and cryo-electron microscopy to solve protein structures.

According to EMBL-EBI, around 30% of the more than 214 million predictions are deemed to be highly accurate, which means they are as good as experimentally determined structures. Another 45% are considered to be accurate enough for many applications.

Many AlphaFold structures are good enough to replace experimental structures for some applications. In other cases, researchers use AlphaFold predictions to guide and make sense of experimental data. Poor predictions are often obvious, and some of them are caused by intrinsic disorder in the protein itself that means it has no defined shape – at least, not without other molecules present.

The 200 million predictions released last week are based on the sequences in another database, called UniProt. It's likely that scientists will have already had an idea about the shapes of some of these proteins, because they are included in databases of experimental structures or resemble other proteins in such repositories, says Edward Porto Parada, a computational biologist at Institut Ciències de la Salut i Recerca Innovadora (ICIS) in Barcelona, Spain. But such entries tend to be limited to well-studied human, mouse and other mammalian proteins. Porto Parada says that the AlphaFold dump will add to their knowledge. Because it includes a diverse set of organisms, "it's like having an awesome reservoir," says Porto Parada. Because AlphaFold's software has been available for a year, researchers have already had the capacity to predict the structure of any protein they wish. In many ways, that availability of information in a single database will save researchers time, money – and, in fact, "it's another barrier of entry that you remove," says Porto Parada. "I've used a lot of AlphaFold models. I have not ever run AlphaFold myself." Jan Kosinski, a structural modeller at EMBL,



Demis Hassabis, CEO of DeepMind

such as the ability to consume plastic – or worrying ones, like those that can drive cancer. The identification of disease-related structures of these proteins in the database can pinpoint the basis for their properties.

Marin Holmberg, a computational biologist at Seoul National University who helped to develop a cloud-based version of AlphaFold, is excited about seeing the database expand. But he says that researchers are still likely to need to use the AI network themselves. Increasingly, people are using AlphaFold to determine how proteins interact, and such predictions are not in the database. Other predictions that are not there include microbial proteins identified by sequencing genetic material from soil, ocean water and other "metagenomic" sources.

Some sophisticated applications of the expanded AlphaFold database might also depend on downloading to entire 23 terabyte contents, which won't be feasible for many teams. Steingger says. Cloud-based storage could also prove costly. Steingger has co-developed a software tool called FoldSeek that can quickly find structurally similar proteins and which should also be able to search the AlphaFold data down.

Even with almost every known protein included, the AlphaFold database will need updating as new organisms are discovered. AlphaFold's predictions can also be improved as new structural information becomes available. Hassabis says DeepMind hopes to update the database annually. He says that the repository will have a lasting impact on the life sciences. "It's going to require quite a big change in thinking."

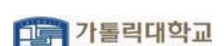
HOW LONG IS COVID INFECTIOUS? WHAT SCIENTISTS KNOW SO FAR

People with SARS-CoV-2 are told to isolate for a few days. But some can pass on the virus for much longer.

By David Adam

When the US Centers for Disease Control and Prevention (CDC) halved its recommended isolation time for people with COVID-19 to five days back in December, it said that the change was motivated by science. Specifically, the CDC said that most SARS-CoV-2 transmission occurs early in the course of the illness, in the one to two days before the onset of symptoms and for two to three days after. Many scientists disputed that decision then

and they continue to do so. Such dissent is bolstered by a series of studies confirming that more people with COVID-19 remain infectious well into the second week after they first experience symptoms. Reductions in the length of the recommended isolation period – now that the change was motivated by science – say they, rather than any reassuring new data. "The fact of how long people are infectious for has not really changed," says Amy Barakat, an infectious disease specialist at Massachusetts General Hospital in Boston. There is no data to support five days or anything



GOLD (PROTEIN) RUSH !



What can I do for a medical research ... ?

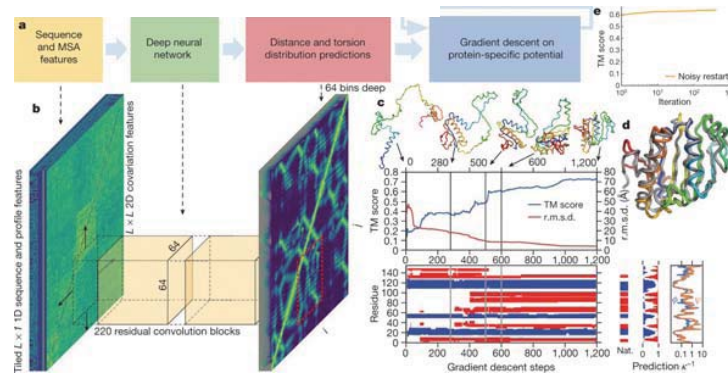
Article



Improved protein structure prediction using potentials from deep learning

<https://doi.org/10.1038/s41586-019-1923-7> Andrew W. Senior^{1,2*}, Richard Evans^{1,2}, John Jumper^{1,2}, James Kirkpatrick^{1,2}, Laurent Sifre^{1,2}, Tim Green¹, Chongli Qin¹, Augustin Zidek¹, Alexander W. R. Nelson¹, Alex Bridgland¹, Hugo Penedones¹, Stig Petersen¹, Karen Simonyan¹, Steve Crossan¹, Pushmeet Kohli¹, David T. Jones^{1,2}, David Silver¹, Koray Kavukcuoglu¹ & Demis Hassabis¹

Received: 2 April 2019
Accepted: 10 December 2019
Published online: 15 January 2020



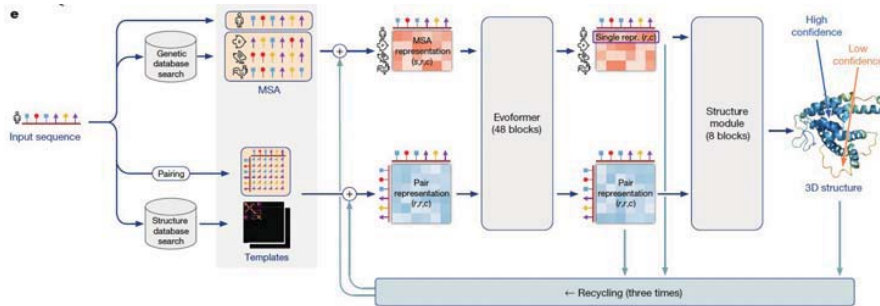
1 AlphaFold

DeepMind **AlphaFold2** Nature. 2021 Aug;596(7873):583-589

Highly accurate protein structure prediction with AlphaFold

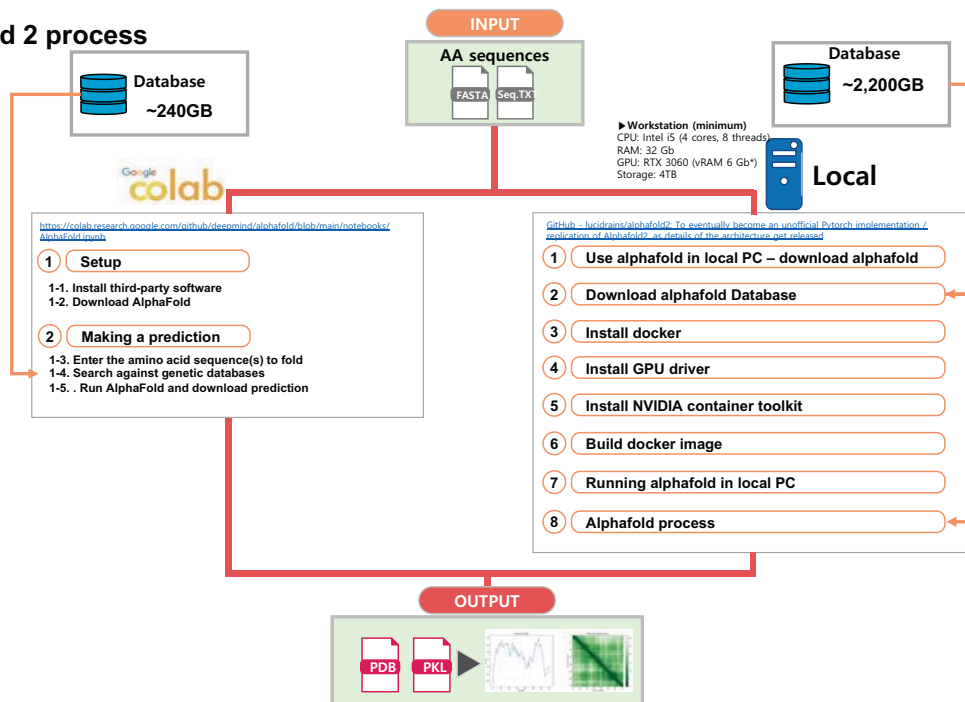
https://doi.org/10.1038/s41586-021-03819-2
 Received: 11 May 2021
 Accepted: 12 July 2021
 Published online: 15 July 2021
 Open access
 Check for updates

John Jumper^{1,2,3*}, Richard Evans^{1,4}, Alexander Pritzel^{1,4}, Tim Green^{1,4}, Michael Figurnov^{1,4}, Olaf Ronneberger^{1,4}, Kathryn Tunyasuvunakool^{1,4}, Russ Bates^{1,4}, Augustin Židek^{1,4}, Anna Potapenko^{1,4}, Alex Bridgland^{1,4}, Clemens Meyer^{1,4}, Simon A. A. Kohl^{1,4}, Andrew J. Ballard^{1,4}, Andrew Cowie^{1,4}, Bernardino Romera-Paredes^{1,4}, Stanislaw Nikolov^{1,4}, Rishub Jain^{1,4}, Jonas Adler^{1,4}, Trevor Back^{1,4}, Stig Petersen^{1,4}, David Reiman^{1,4}, Ellen Clancy^{1,4}, Michal Zietliński^{1,4}, Martin Steinegger^{1,4}, Michalina Pacholska^{1,4}, Tamas Berghammer^{1,4}, Sebastian Bodenstein^{1,4}, David Silver^{1,4}, Oriol Vinyals^{1,4}, Andrew W. Senior^{1,4}, Koray Kavukcuoglu^{1,4}, uShmeel Kohli^{1,4} & Demis Hassabis^{1,2,3*}



2 Colab 환경에서의 AlphaFold 사용법

AlphaFold 2 process



2 Colab 환경에서의 AlphaFold 사용법

AlphaFold database usage

```

DB list
$DOWNLOAD_DIR/
bfd/ # Total: ~ 2.2 TB (download: 438 GB)
# ~ 1.7 TB (download: 271.6 GB)
# 6 files.
mgnify/ # ~ 64 GB (download: 32.9 GB)
mgy_clusters_2018_12.fa
params/ # ~ 3.5 GB (download: 3.5 GB)
# 5 CASP14 models,
# 5 pTM models,
# 5 AlphaFold-Multimer models,
# LICENSE,
# = 16 files.
pdb70/ # ~ 56 GB (download: 19.5 GB)
# 9 files.
pdb_mmcif/ # ~ 206 GB (download: 46 GB)
mmcif_files/
# About 180,000 .cif files.
obsolete.dat
pdb_seqres/ # ~ 0.2 GB (download: 0.2 GB)
pdb_seqres.txt
small_bfd/ # ~ 17 GB (download: 9.6 GB)
bfd-first_non_consensus_sequences.fasta
uniclust30/ # ~ 86 GB (download: 24.9 GB)
uniclust30_2018_08/
# 13 files.
uniprot/ # ~ 98.3 GB (download: 49 GB)
uniprot.fasta
uniref90/ # ~ 58 GB (download: 29.7 GB)
uniref90.fasta
  
```

Local

Get MSA

Search templates

AF multimer

Get MSA



2 Colab 환경에서의 AlphaFold 사용법 : AlphaFold2

README.md

AlphaFold

This package provides an implementation of the inference pipeline of AlphaFold v2.0. This is a completely new model that was entered in CASP14 and published in Nature. For simplicity, we refer to this model as AlphaFold throughout the rest of this document.


We also provide an implementation of AlphaFold-Multimer. This represents a work in progress and AlphaFold-Multimer isn't expected to be as stable as our monomer AlphaFold system. Read the guide for how to upgrade and update code.

Any publication that discloses findings arising from using this source code or the model parameters should cite the AlphaFold paper and, if applicable, the AlphaFold-Multimer paper.


Please also refer to the Supplementary information for a detailed description of the method.

★ you can use a slightly simplified version of AlphaFold with this Colab notebook or community-supported versions (see below). →

If you have any questions, please contact the AlphaFold team at alphafold@deepmind.com.



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction



2 Colab 환경에서의 AlphaFold 사용법 : AlphaFold2

The screenshot shows a Google Colab notebook titled 'get PDB.ipynb'. The 'Edit' menu is open, and the 'Notebook settings' dialog is displayed. In the dialog, the 'Hardware accelerator' dropdown is set to 'GPU'. Red annotations include a box around 'Edit' in the menu, a box around the 'Notebook settings' dialog title, and a box around the 'GPU' option in the dropdown. An arrow points from the 'Notebook settings' dialog to the 'GPU' option.

2 Colab 환경에서의 AlphaFold 사용법 : AlphaFold2

Five steps to run AlphaFold2

1. Install third-party software
Please execute this cell by pressing the Play button on the left to download and import third-party software in this Colab notebook. (See the [acknowledgements](#) in our readme.)
Note: This installs the software on the Colab notebook in the cloud and not on your computer.
코드 표시
2. Download AlphaFold
Please execute this cell by pressing the Play button on the left.
코드 표시

Making a prediction

Please paste the sequence of your protein in the text box below, then run the remaining cells via **Runtime > Run** after. You can also run the cells individually by pressing the Play button on the left.

Note that the search against databases and the actual prediction can take some time, from minutes to hours, depending on the length of the protein and what type of GPU you are allocated by Colab (see FAQ below).

3. Enter the amino acid sequence(s) to fold

Enter the amino acid sequence(s) to fold:

- if you enter only a single sequence, the monomer model will be used.
- if you enter multiple sequences, the multimer model will be used.

sequence_1: "여기에 text 입력"

sequence_2: "여기에 text 입력"

sequence_3: "여기에 text 입력"

sequence_4: "여기에 text 입력"

sequence_5: "여기에 text 입력"

sequence_6: "여기에 text 입력"

input length range: min: 16, max: 2500

3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

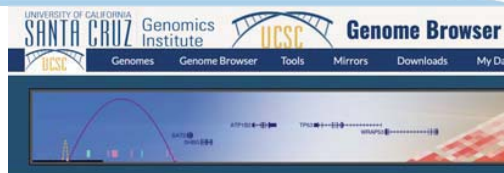
Google

검색결과 약 4,740,000개 (0.45초)

https://genome.ucsc.edu

UCSC Genome Browser Home

Tools · Genome Browser - Interactively visualize genomic data · BLAT - Rapidly align sequences to the genome · In-Silico PCR - Rapidly align PCR primer pairs to ...



- Tools**
- Genome Browser** - Interactively visualize genomic data
 - BLAT** - Rapidly align sequences to the genome
 - In-Silico PCR** - Rapidly align PCR primer pairs to the genome
 - Table Browser** - Download and filter data from the Genome Browser
 - LiftOver** - Convert genome coordinates between assemblies
 - REST API** - Returns data requested in JSON format
 - Variant Annotation Integrator** - Annotate genomic variants
 - More tools...**

Browse/Select Species

POPULAR SPECIES

Search through thousands of genome browsers
Enter species, common name or assembly ID

Unable to find a genome? Send us a request.

UCSC SPECIES TREE AND CONNECTED ASSEMBLY HUBS

Find Position

SARS-CoV-2 Assembly
Jan. 2020 (NC_045512.2)

Position/Search Term
Enter position, gene symbol or search terms
Current position: NC_045512:1-29,903

SARS-CoV-2 Genome Browser - wuhCor1 assembly [view sequences](#)

The SARS-CoV-2 coronavirus emerged in December 2019 as a novel human pathogen causing a severe acute respiratory syndrome (COVID-19). In the following months the disease spread internationally and was declared a pandemic by the World Health Organization on March 11, 2020.

This genome browser is based on sequence obtained from the sample Wuhan-Hu-1, obtained in December 2019 in the city of Wuhan (Hubei province, China), submitted to the Shanghai Public Health Clinical Center on January 5, 2020 and to the US National Center for Biotechnology Information on January 17, 2020.

A manuscript describing this work, The UCSC SARS-CoV-2 Genome Browser, was published in the September 9, 2020 issue of Nature Genetics.

Additional information can be found in our UCSC COVID-19 Resources page. See also our web interface to USHER, a tool for placing new SARS-CoV-2 sequences in a global phylogenetic tree.

Download sequence and annotation data:

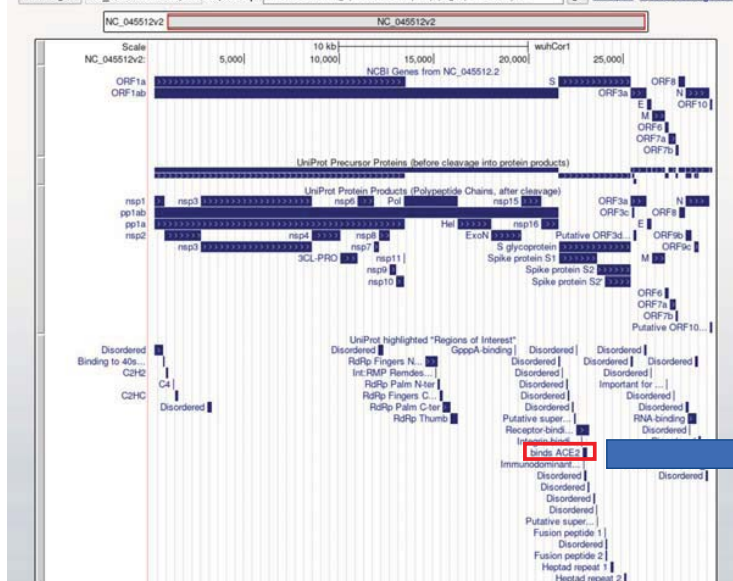
- Using HTTP
- Using FTP



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

UCSC Genome Browser on SARS-CoV-2 Jan. 2020 (NC_045512.2) (wuhCor1)

multi-region NC_045512:1-29,903 29,903 bp. [chromosome range](#), [search terms](#), [help pages](#), see examples [Quick start guide](#)



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Google search results for 'cov lineage'. The search results show 'Cov-Lineages' as the top result. A red box highlights the 'Lineage List' link, with a blue arrow pointing to the right.

Lineage List

All Fields search for lineages...

Lineage	Most common countries	Earliest date	# designated	# assigned	Description	WHO Name
A	United States of America 32.0%, United_Arab_Emirates 11.0%, China 8.0%, Germany 7.0%, Canada 4.0%	2019-12-30	1697	2547	One of the two original haplotypes of the pandemic (A and B). Many sequences originating from China and many global exports, including to South East Asia Japan South Korea Australia the USA and Europe represented in this lineage	
B	United States of America 37.0%, United Kingdom 19.0%, China 7.0%, Mexico 6.0%, Germany 4.0%	2019-12-24	4001	9688	One of the two original haplotypes of the pandemic (and first to be discovered)	
B.1	United States of America 46.0%, Turkey 12.0%, United Kingdom 7.0%, Canada 4.0%, France 3.0%	2020-01-01	46228	109623	A large European lineage the origin of which roughly corresponds to the Northern Italian outbreak early	



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

All Fields delta

Lineage	Most common countries	Earliest date	# designated	# assigned	Description	WHO Name
B.1.617.2	United States of America 22.0%, India 19.0%, United Kingdom 13.0%, Turkey 7.0%, Germany 5.0%	2020-03-27	8274	177047	Predominantly India lineage with several spike mutations, pango-designation issue	Delta

Lineage B.1.617.2

→Go to parent lineage: B.1.617

Predominantly India lineage with several spike mutations, pango-designation issue #49

Most Common Countries: United States of America 22.0%, India 19.0%, United Kingdom 13.0%, Turkey 7.0%, Germany 5.0%

Earliest Date: 2020-03-27

Number Designated: 8274

Number Assigned: 177047

[View more information at OutbreakInfo](#)

WHO Name: Delta

PHE Name(s): VOC-21APR-02

Characteristic mutations in lineage

Mutations in at least 75% of B.1.617.2 sequences (read more)

Compare to other lineages View S-gene mutations

Characteristic mutations of B.1.617.2

gene	amino acid
ORF1b	P314L
ORF1b	G662S
ORF1b	P1000L
S	T19R
S	E156G
S	del157/158
S	L452R
S	T478K
S	D614G
S	P681R
S	D950N



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Setup

Start by running the 2 cells below to set up AlphaFold and all required software.

[1] 1. Install third-party software

4분

Please execute this cell by pressing the *Play* button on the left to download and import third-party software in this Colab notebook. (See the [acknowledgements](#) in our readme.)

Note: This installs the software on the Colab notebook in the cloud and not on your computer.

[코드 표시](#)

100%  100/100 [elapsed: 04:37 remaining: 00:00]

[2] 2. Download AlphaFold

1분

Please execute this cell by pressing the *Play* button on the left.

[코드 표시](#)

100%  100/100 [elapsed: 01:39 remaining: 00:00]

Running with Tesla T4 GPU



3. Enter the amino acid sequence(s) to fold

Enter the amino acid sequence(s) to fold:

- If you enter only a single sequence, the monomer model will be used.
- If you enter multiple sequences, the multimer model will be used.

sequence_1: "AATTCTAACAACTCTTGATTCTAAGGTTGGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTGAACGAAATCTATCAGGCGGGTAGCACACCTTGTAAATGGTGTGAAGGTTTAAATTGTTACTTCCTTTACAATCATATGTTTCCAACCCACTAATGGTGTGGTTACCAACCATAC

sequence_2: "여기에 text 입력"

sequence_3: "여기에 text 입력"

sequence_4: "여기에 text 입력"

sequence_5: "여기에 text 입력"

sequence_6: "여기에 text 입력"

sequence_7: "여기에 text 입력"

sequence_8: "여기에 text 입력"

[코드 표시](#)

Using the single-chain model.

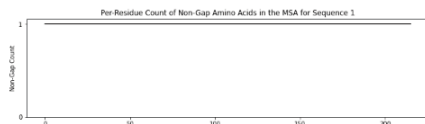
4. Search against genetic databases

Once this cell has been executed, you will see statistics about the multiple sequence alignment (MSA) that will be used by AlphaFold. In particular, you'll see how well each residue is covered by similar sequences in the MSA.

[코드 표시](#)

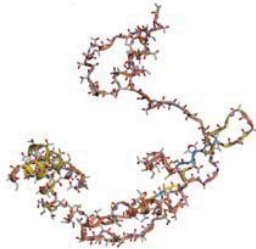
Getting MSA for sequence 1
Searching mgnify: 100%  147/147 [elapsed: 36.06 remaining: 00:00]

1 unique sequences found in uniref100 for sequence 1
1 unique sequences found in smal100 for sequence 1
1 unique sequences found in mgnify for sequence 1
1 unique sequences found in total for sequence 1

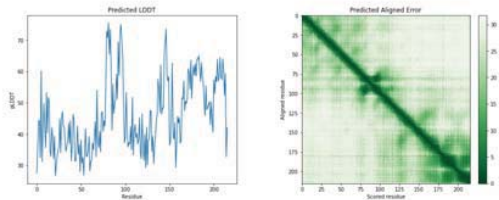


5. Run AlphaFold and download prediction

Once this cell has been executed, a zip-archive with the obtained prediction will be automatically downloaded to your computer.
 In case you are having issues with the relaxation stage, you can disable it below. Warning: This means that the prediction might have distracting small stereochemical violations.
 relax_relax:
 Relaxation is faster with a GPU, but we have found it to be less stable. You may wish to enable GPU for higher performance, but if it doesn't converge we suggested reverting to using without GPU.
 relax_use_gpu:



Model Confidence
 Very low (pLDDT < 50)
 Low (70 > pLDDT > 50)
 Confident (90 > pLDDT > 70)
 Very high (pLDDT > 90)



MODEL	1										
ATOM	1	N	ALA	A	1	11.310	45.250	12.074	1.00	27.48	N
ATOM	2	CA	ALA	A	1	10.784	44.580	13.260	1.00	27.48	C
ATOM	3	C	ALA	A	1	10.026	43.310	12.882	1.00	27.48	C
ATOM	4	CB	ALA	A	1	9.877	45.523	14.047	1.00	27.48	C
ATOM	5	O	ALA	A	1	9.012	43.370	12.183	1.00	27.48	O
ATOM	6	N	ALA	A	2	10.663	42.210	12.688	1.00	35.26	N
ATOM	7	CA	ALA	A	2	9.972	40.934	12.528	1.00	35.26	C
ATOM	8	C	ALA	A	2	10.841	39.775	13.009	1.00	35.26	C
ATOM	9	CB	ALA	A	2	9.571	40.725	11.069	1.00	35.26	C
ATOM	10	O	ALA	A	2	11.836	39.430	12.368	1.00	35.26	O
ATOM	11	N	THR	A	3	11.091	39.784	14.373	1.00	44.04	N
ATOM	12	CA	THR	A	3	11.033	38.576	15.189	1.00	44.04	C
ATOM	13	C	THR	A	3	9.720	38.510	15.964	1.00	44.04	C
ATOM	14	CB	THR	A	3	12.217	38.509	16.171	1.00	44.04	C
ATOM	15	O	THR	A	3	9.443	39.372	16.800	1.00	44.04	O
ATOM	16	CG2	THR	A	3	12.379	37.103	16.740	1.00	44.04	C
ATOM	17	OG1	THR	A	3	13.420	38.877	15.486	1.00	44.04	O
ATOM	18	N	THR	A	4	8.656	37.925	15.358	1.00	44.39	N
ATOM	19	CA	THR	A	4	7.840	37.025	16.165	1.00	44.39	C
ATOM	20	C	THR	A	4	7.094	36.031	15.279	1.00	44.39	C
ATOM	21	CB	THR	A	4	6.831	37.807	17.027	1.00	44.39	C
ATOM	22	O	THR	A	4	5.959	35.655	15.579	1.00	44.39	O
ATOM	23	CG2	THR	A	4	7.475	38.293	18.321	1.00	44.39	C
ATOM	24	OG1	THR	A	4	6.357	38.939	16.287	1.00	44.39	O
ATOM	25	N	CYS	A	5	7.714	35.364	14.361	1.00	32.47	N
ATOM	26	CA	CYS	A	5	7.085	34.230	13.693	1.00	32.47	C
ATOM	27	C	CYS	A	5	7.618	32.911	14.240	1.00	32.47	C
ATOM	28	CB	CYS	A	5	7.321	34.298	12.184	1.00	32.47	C
ATOM	29	O	CYS	A	5	7.032	31.853	14.001	1.00	32.47	O
ATOM	30	SG	CYS	A	5	5.952	35.039	11.267	1.00	32.47	S
ATOM	31	N	THR	A	6	7.951	32.750	15.488	1.00	60.19	N
ATOM	32	CA	THR	A	6	8.169	31.370	15.906	1.00	60.19	C
ATOM	33	C	THR	A	6	7.118	30.941	16.926	1.00	60.19	C
ATOM	34	CB	THR	A	6	9.576	31.183	16.504	1.00	60.19	C
ATOM	35	O	THR	A	6	6.667	29.794	16.916	1.00	60.19	O
ATOM	36	CG2	THR	A	6	10.014	29.724	16.433	1.00	60.19	C



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

3. Enter the amino acid sequence(s) to fold

Enter the amino acid sequence(s) to fold:

- If you enter only a single sequence, the monomer model will be used.
- If you enter multiple sequences, the multimer model will be used.

Sequence of Receptor Binding Domain (RBD)

```
sequence_1: RVPQTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPFTKLDLCLFVNYADSFVIRGDEVRIQAPGQTGIADYNYKLPDDFTGCVIAWNSNHLDSKVGNYNYLYRLFRKSNLKKPFERDISTEYIQAQSTPCNGVEGFNCYFPLQSYGFQPTVGVGYPYI
sequence_2: 여기에 text 입력
sequence_3: 여기에 text 입력
sequence_4: 여기에 text 입력
sequence_5: 여기에 text 입력
sequence_6: 여기에 text 입력
sequence_7: 여기에 text 입력
sequence_8: 여기에 text 입력
```

코드 표시

4. Search against genetic databases

Once this cell has been executed, you will see statistics about the multiple sequence alignment (MSA) that will be used by AlphaFold. In particular, you'll see how well each residue is covered by similar sequences in the MSA.

코드 표시

5. Run AlphaFold and download prediction

Once this cell has been executed, a zip-archive with the obtained prediction will be automatically downloaded to your computer.

In case you are having issues with the relaxation stage, you can disable it below. Warning: This means that the prediction might have distracting small stereochemical violations.

run_relax:

코드 표시

+ 코드 + 텍스트



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

[4] 4. Search against genetic databases

Once this cell has been executed, you will see statistics about the multiple sequence alignment (MSA) that will be used by AlphaFold. In particular, you'll see how well each residue is covered by similar sequences in the MSA.

코드 표시

Getting MSA for sequence 1

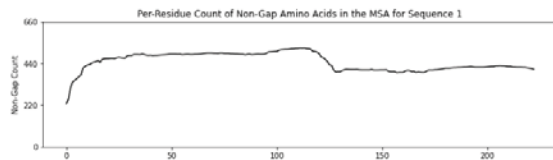
Searching mgnify: 100% 147/147 (elapsed: 17:09 remaining: 00:00)

658 unique sequences found in uniref90 for sequence 1

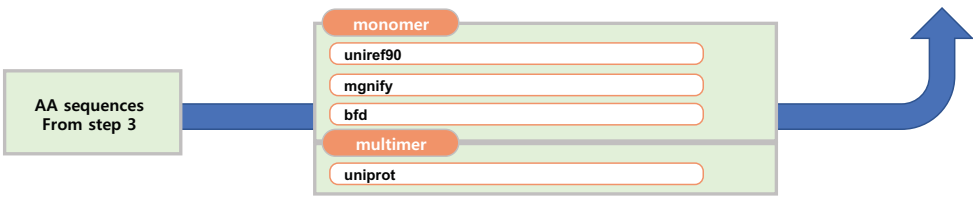
4 unique sequences found in smallbfd for sequence 1

1 unique sequences found in mgnify for sequence 1

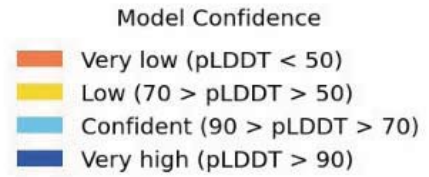
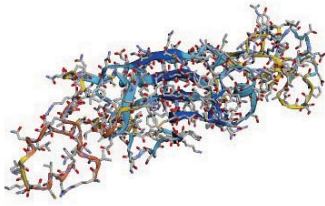
660 unique sequences found in total for sequence 1



```
Q58940_BOVIN -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_HUMAN -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_MOUSE -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_HAT -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_CLICK -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_RABBY -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
Q7ZUG3_BRAKE -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_TCTFU -----MREDRATWRSYFLKIQGLDQKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_DROME -----MYREKAKAKAQYIKVVELEDFKCFIVGADVGSKQMQDMSIRG-AVVLGKSTHSHRAIRGRLHNM--PALE 76
RLAO_DICDI -----MSLAE-SKKRLPIEKRTKLTTPKQKIVLADYVSDIQLDCKEYSLGI-SAVLHGKSTHSHRAIRGRLHNM--PALE 75
Q54LP0_DICDI -----MSLAE-SKKRLPIEKRTKLTTPKQKIVLADYVSDIQLDCKEYSLGI-SAVLHGKSTHSHRAIRGRLHNM--PALE 75
RLAO_PLATF -----MAKLSFGQKQHWYKESLQQKSLIVYVGVGSHMAYVYSIRG-KETLHGKSTHSHRAIRGRLHNM--PALE 76
RLAO_SULAC -----HLSAVYTKKIKRNVYFAEIKLACTIKLILAMISQVADLIEIKKIRG-ADIVYKSTHSHRAIRGRLHNM--PALE 79
RLAO_SULTO -----HLSAVYTKKIKRNVYFAEIKLACTIKLILAMISQVADLIEIKKIRG-ADIVYKSTHSHRAIRGRLHNM--PALE 79
RLAO_SULSO -----HLSAVYTKKIKRNVYFAEIKLACTIKLILAMISQVADLIEIKKIRG-ADIVYKSTHSHRAIRGRLHNM--PALE 80
RLAO_SULSO -----HKLALAKQKRYASWLELYKELTLEKSNITLQNGKFPADKMEIKKIRG-AEIVYKSTHSHRAIRGRLHNM--PALE 80
RLAO_HERPE -----HSPGIVYQWYKKEKYSFSEILMGSEELSESEVLEFLADYDQYFYVYKSKKQK-SYHWFALGILISLHMGANGLE---IDM 86
RLAO_DYRAE -----HMLAGKRAVYTRDPAKLVYERATLQKLVYFLPOLHSGSRIGHEVYRLRY-GYIILIDPLFKITAVYKSG---IDM 85
RLAO_METAC -----MKREHHTTEIFQWQKDETEKXLOSKEVYGVSLGSLATKQKIDEDLDQV-AVLYVYKSTHSHRAIRGRLHNM--PALE 78
RLAO_METHA -----MKREHHTTEIFQWQKDETEKXLOSKEVYGVSLGSLATKQKIDEDLDQV-AVLYVYKSTHSHRAIRGRLHNM--PALE 78
RLAO_ABCUF -----MAVAVS-----PFLKRVAVSEKMHISEVVAIVSRVFAQKQKIRAFKQ-ATIKVYKSTHSHRAIRGRLHNM--PALE 75
RLAO_METKA -----MAVAVS-----PFLKRVAVSEKMHISEVVAIVSRVFAQKQKIRAFKQ-ATIKVYKSTHSHRAIRGRLHNM--PALE 75
RLAO_METHH -----MAVAVS-----PFLKRVAVSEKMHISEVVAIVSRVFAQKQKIRAFKQ-ATIKVYKSTHSHRAIRGRLHNM--PALE 74
RLAO_METTL -----MITAREHKKIAPKIEYKELKELKQVIALVQMSVPAIDQITQKIR-SHMLKMSHTLILKELVAVLQVTFEFA 82
RLAO_METVA -----HIDAKSHKIAQWILEYVALKELKSNVIALVQMSVPAIDQITQKIR-SHMLKMSHTLILKELVAVLQVTFEFA 82
RLAO_METJA -----MEKVKVAFKIEKIEYKELKELKQVIALVQMSVPAIDQITQKIR-SHMLKMSHTLILKELVAVLQVTFEFA 81
RLAO_DYRAE -----MAVAVS-----PFLKRVAVSEKMHISEVVAIVSRVFAQKQKIRAFKQ-ATIKVYKSTHSHRAIRGRLHNM--PALE 77
RLAO_PYRHO -----MAVAVS-----PFLKRVAVSEKMHISEVVAIVSRVFAQKQKIRAFKQ-ATIKVYKSTHSHRAIRGRLHNM--PALE 77
RLAO_PYRHO -----MAVAVS-----PFLKRVAVSEKMHISEVVAIVSRVFAQKQKIRAFKQ-ATIKVYKSTHSHRAIRGRLHNM--PALE 77
RLAO_PYRHO -----MAVAVS-----PFLKRVAVSEKMHISEVVAIVSRVFAQKQKIRAFKQ-ATIKVYKSTHSHRAIRGRLHNM--PALE 77
RLAO_HALMA -----HLSERKRTETIPIHKEVDAVYVHICHSVGVVVALSFLSKQDMSHDSHT-ANLVYKSTHSHRAIRGRLHNM--PALE 79
RLAO_HALVO -----HLSERKRTETIPIHKEVDAVYVHICHSVGVVVALSFLSKQDMSHDSHT-ANLVYKSTHSHRAIRGRLHNM--PALE 79
RLAO_HALSA -----HLSERKRTETIPIHKEVDAVYVHICHSVGVVVALSFLSKQDMSHDSHT-ANLVYKSTHSHRAIRGRLHNM--PALE 79
RLAO_THEAC -----HKVYQCKRLEKIEKESASVATVDEKIEKIDQDQKMDK-EMLVYKSTHSHRAIRGRLHNM--PALE 72
RLAO_THEVO -----MKEDKPKKCLYKELKQVETKSKAVIYDQVYKESQDLEAKMDK-FLKLVYKSTHSHRAIRGRLHNM--PALE 72
RLAO_PICTO -----MTEPKWIDFKWHEHHEKSKVATYSKGNHHEKIDMSIDQK-ARLYVYKSTHSHRAIRGRLHNM--PALE 72
Tuler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90
```



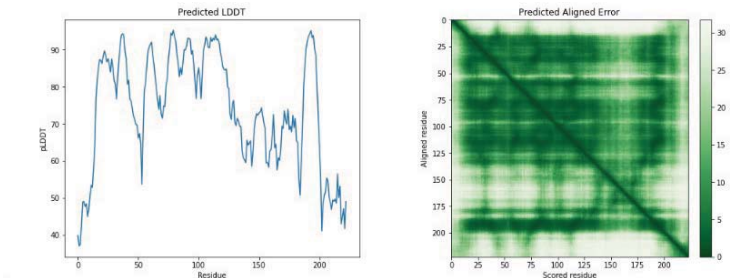
3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측



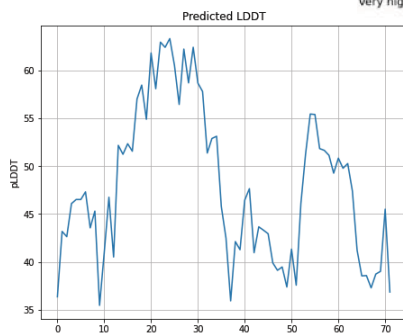
Local Distance Difference Test (LDDT)

pFAM uses a quality score (LDDT)

>= 0.6 : considered a reasonable model
> 0.8 : great model



Model Confidence
 Very low (pLDDT < 50)
 Low (70 > pLDDT > 50)
 Confident (90 > pLDDT > 70)
 Very high (pLDDT > 90)



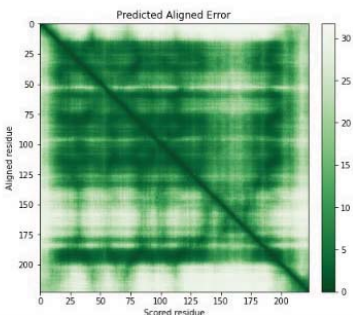
pLDDT(predicted Local Distance Difference Test)

pLDDT는 LDDT값을 예측한 수치로 예측 모델의 각 residue에서의 신뢰도를 의미합니다.

실제 모델의 residue와 얼마나 일치하는지 예측한 값인 동시에 해당 위치에서 단백질 구조가 folding이 잘 되는지를 나타냅니다.

pLDDT값이 낮음은 신뢰도가 낮음을 의미하며 해당 위치의 residue들이 무질서한 구조를 가짐을 의미합니다.

LDDT : 예측된 모델이 실제 단백질과 얼마나 일치하는지를 나타내는 [0,1]의 범위를 가진 값으로 1에 가까울 수록 실제 단백질과 일치함을 의미합니다.

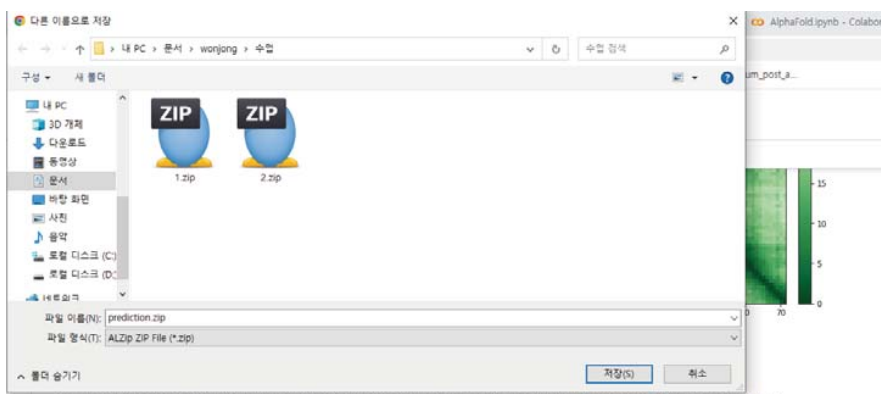


PAE(Predicted Aligned Error)

PAE는 모델의 residue간의 위치 에너지가 실제 모델과 얼마나 차이를 가지는지 예측한 값으로 해당 값이 낮으면 두 residue간 위치의 정확도가 높음을 의미합니다.



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측



In general predicted LDDT (pLDDT) is best used for intra-domain confidence, whereas Predicted Aligned Error (PAE) is best used for determining between domain or between chain confidence.

Please see the [AlphaFold methods paper](#), the [AlphaFold predictions of the human proteome paper](#), and the [AlphaFold-Multimer paper](#) as well as our [FAQ](#) on how to interpret AlphaFold predictions.

사용자 > User > 문서 > wonjong > 수업 > prediction > prediction

이름	수정한 날짜	유형	크기
predicted_aligned_error.json	2022-05-03 오전 9:05	JSON 파일	54KB
selected_prediction.pdb	2022-05-03 오전 9:05	Protein Data Bank...	89KB



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측



Record Name	Atom serial number	Atom name	Residue Name	Chain ID	Residue sequence number	X	Y	Z	Occupancy	temperature factor (B-factor)	Element symbol
ATOM	1	N	MET	A	1	-6.594	-16.75	58.3	1	23.69	N
ATOM	2	H	MET	A	1	-6.533	-15.777	58.035	1	23.69	H
ATOM	3	H2	MET	A	1	-7.551	-16.959	58.547	1	23.69	H
ATOM	4	H3	MET	A	1	-6.011	-16.91	59.109	1	23.69	H
ATOM	5	CA	MET	A	1	-6.158	-17.624	57.184	1	23.69	C
ATOM	6	HA	MET	A	1	-6.298	-18.659	57.494	1	23.69	H
ATOM	7	C	MET	A	1	-4.656	-17.501	56.924	1	23.69	C
ATOM	8	CB	MET	A	1	-6.99	-17.429	55.907	1	23.69	C
ATOM	9	HB2	MET	A	1	-6.503	-17.961	55.09	1	23.69	H
ATOM	10	HB3	MET	A	1	-7.043	-16.372	55.649	1	23.69	H
ATOM	11	O	MET	A	1	-3.943	-18.391	57.354	1	23.69	O
ATOM	12	CG	MET	A	1	-8.406	-17.996	56.06	1	23.69	C
ATOM	13	HG2	MET	A	1	-8.337	-19.033	56.39	1	23.69	H
ATOM	14	HG3	MET	A	1	-8.953	-17.425	56.81	1	23.69	H
ATOM	15	SD	MET	A	1	-9.333	-17.958	54.511	1	23.69	S
ATOM	16	CE	MET	A	1	-10.835	-18.857	54.988	1	23.69	C
ATOM	17	HE1	MET	A	1	-11.339	-18.332	55.8	1	23.69	H
ATOM	18	HE2	MET	A	1	-10.578	-19.867	55.307	1	23.69	H
ATOM	19	HE3	MET	A	1	-11.507	-18.914	54.132	1	23.69	H
ATOM	20	N	PHE	A	2	-4.184	-16.412	56.3	1	22.95	N
ATOM	21	H	PHE	A	2	-4.856	-15.744	55.949	1	22.95	H
ATOM	22	CA	PHE	A	2	-2.816	-16.168	55.783	1	22.95	C
ATOM	23	HA	PHE	A	2	-2.798	-16.448	54.73	1	22.95	H
ATOM	24	C	PHE	A	2	-1.599	-16.889	56.41	1	22.95	C
ATOM	25	CB	PHE	A	2	-2.557	-14.651	55.862	1	22.95	C
ATOM	26	HB2	PHE	A	2	-2.966	-14.246	56.788	1	22.95	H
ATOM	27	HB3	PHE	A	2	-1.482	-14.476	55.896	1	22.95	H
ATOM	28	O	PHE	A	2	-0.756	-17.388	55.674	1	22.95	O
ATOM	29	CG	PHE	A	2	-3.099	-13.873	54.68	1	22.95	C
ATOM	30	CD1	PHE	A	2	-2.249	-13.598	53.592	1	22.95	C
ATOM	31	HD1	PHE	A	2	-1.221	-13.931	53.602	1	22.95	H
ATOM	32	CD2	PHE	A	2	-4.434	-13.423	54.656	1	22.95	C

pLDDT



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측



AlphaFold v.2.2.3(2022.08.25 released)

●Change the Colab PAE json output to new format that matches the format used in the new release of the AlphaFold Protein Structure Database (AFDB).

Old version

```
[
  {
    "residue1": [1, 1, 1, 1, 1, ...],
    "residue2": [1, 2, 3, 4, 5, ...],
    "predicted_aligned_error": [0, 1, 4, 7, 9, ..., ...],
    "max_predicted_aligned_error": 31.75
  }
]
```

Latest version

```
[
  {
    "predicted_aligned_error": [[0, 1, 4, 7, 9, ...], ...],
    "max_predicted_aligned_error": 31.75
  }
]
```



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

<http://honglab.catholic.ac.kr>



[2023 KSBI-BIML] 강의 자료

작성자	Dongwan Hong	조회수	1
첨부파일	BIML_2.zip[30588954 byte] BIML_1.zip[178277 byte]		

- 2023 KSBI-BIML 강의자료
- AlphaFold를 활용한 실습

LECTURE

- [2022-11-04] 핵심과학기술실험
- [2022-09-24] [22학년도2학기]의학과3학년_맞춤의학
- [2022-09-16] [22학년도2학기]BIT융합정밀의학협동&...
- [2022-08-30] [22학년도 2학기] 의료 빅데이터, 생명정...
- [2022-06-14] [의학과 2학년] (22학년도 1학기 선택과정)

NEWS

- [2023-01-27] [2023 KSBI-BIML] 강의 자료
- [2023-01-02] [환정] (신입생) 정대선, 송유석
- [2022-12-27] [BK21_4TH 미래인재형 의과학자 교...
- [2022-12-06] AI 기반 디지털 바이오 심포지엄 발표
- [2022-11-30] LAIDD (LECTURES ON AI-DRIVE...

3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Google colab을 이용한 예측 구조의 해석

The image shows a Google search interface. On the left, the Google logo is displayed above a search bar containing 'google colab'. Below the search bar, a dropdown menu shows 'google colab' and 'google colab - Google 검색'. On the right, the search results page is shown. The search bar contains 'google colab'. Below it, the search results are listed. The first result is 'https://colab.research.google.com' with the title 'Google Colab'. Below this, there is a description of Colab notebooks and a 'Pro' section. The 'Filter notebooks' and 'Connect' sections are also visible.



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

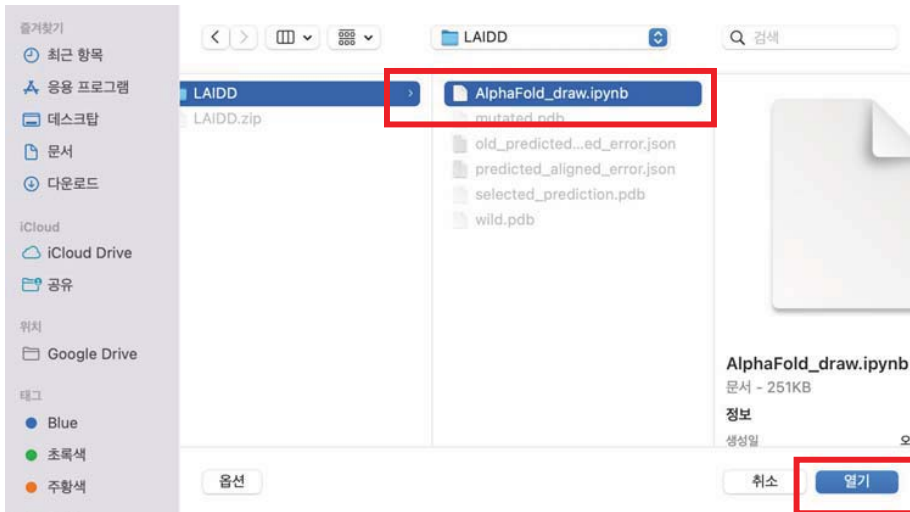
Google colab을 이용한 예측 구조의 해석

The image shows the Google Colab interface. At the top, there is a navigation bar with tabs for '예' (Home), '최근 사용' (Recent), 'Google Drive', 'GitHub', and '업로드' (Upload). The '업로드' tab is highlighted with a red box. Below the navigation bar, there is a large dashed box representing the workspace. Inside this box, there is a button labeled '파일 선택' (Select file) and '선택된 파일 없음' (No files selected). This button is also highlighted with a red box. At the bottom right of the workspace, there is a '취소' (Cancel) button.



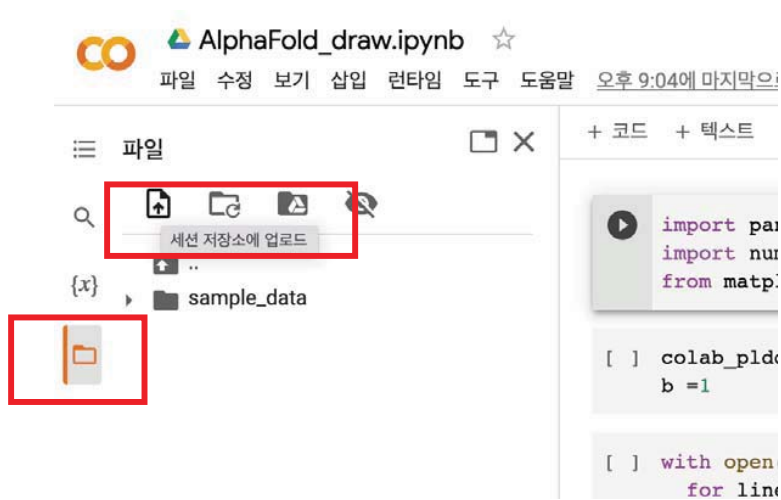
3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Google colab을 이용한 예측 구조의 해석



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Google colab을 이용한 예측 구조의 해석



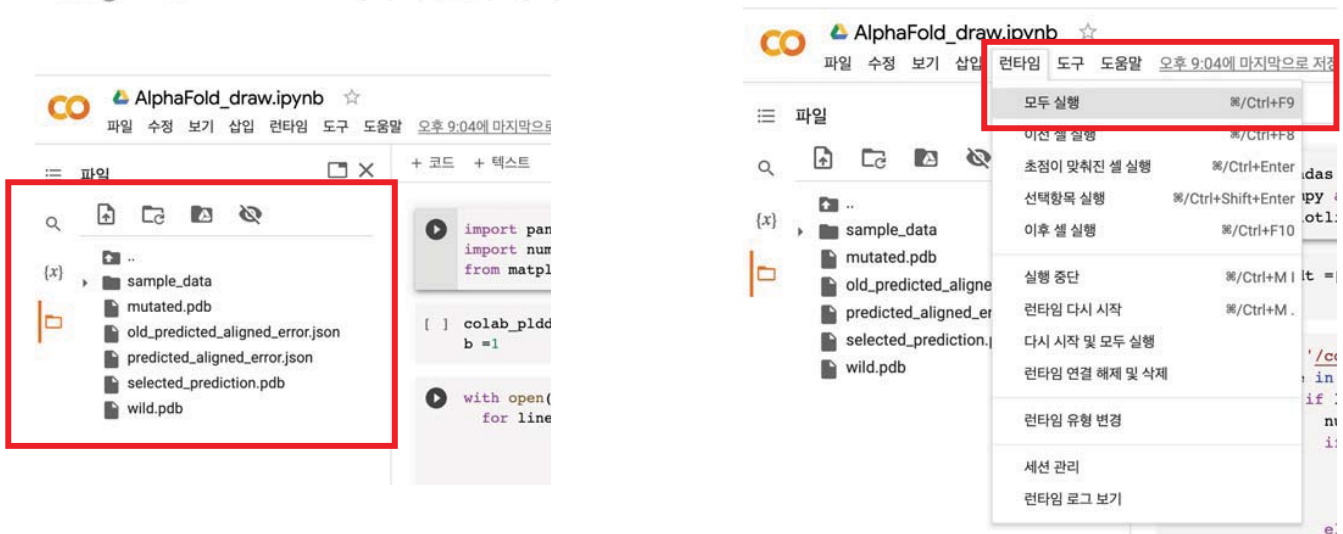
3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Google colab을 이용한 예측 구조의 해석



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Google colab을 이용한 예측 구조의 해석



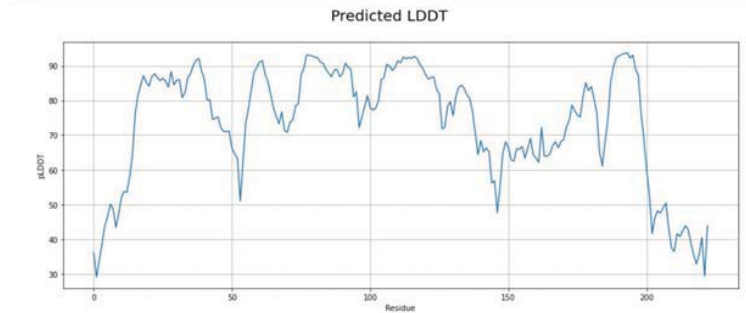
3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측 Get information from PDB & JSON file – draw pLDDT & PAE

```
[1] import json
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt

[2] colab_plddt = []
b = 1

[3] with open('/content/selected_prediction.pdb', 'r') as file:
    for line in file:
        if line.startswith('ATOM'):
            number = int(line[22:26])
            if number == b:
                b+=1
                score = line[61:66].strip()
                colab_plddt.append(float(score))
            elif number == 1 and b > 2:
                b=1
                b+=1
                score = line[61:66].strip()
                colab_plddt.append(float(score))

plt.figure(figsize=[16, 6])
plt.plot(colab_plddt, linestyle='solid')
plt.grid('True')
plt.xlabel('Residue')
plt.ylabel('pLDDT')
plt.suptitle('Predicted LDDT', fontsize=20)
plt.savefig('/content/pLDDT.png', dpi=300)
```



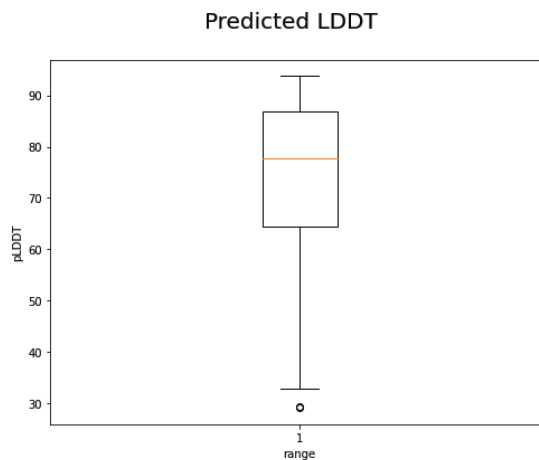
3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측 Get information from PDB & JSON file – draw pLDDT & PAE

```
[1] import json
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt

[2] colab_plddt = []
b = 1

[3] with open('/content/selected_prediction.pdb', 'r') as file:
    for line in file:
        if line.startswith('ATOM'):
            number = int(line[22:26])
            if number == b:
                b+=1
                score = line[61:66].strip()
                colab_plddt.append(float(score))
            elif number == 1 and b > 2:
                b=1
                b+=1
                score = line[61:66].strip()
                colab_plddt.append(float(score))

[9] plt.figure(figsize=[8, 6])
plt.boxplot([colab_plddt])
plt.xlabel('range')
plt.ylabel('pLDDT')
plt.suptitle('Predicted LDDT', fontsize=20)
plt.savefig('/content/pLDDT.png', dpi=300)
```



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

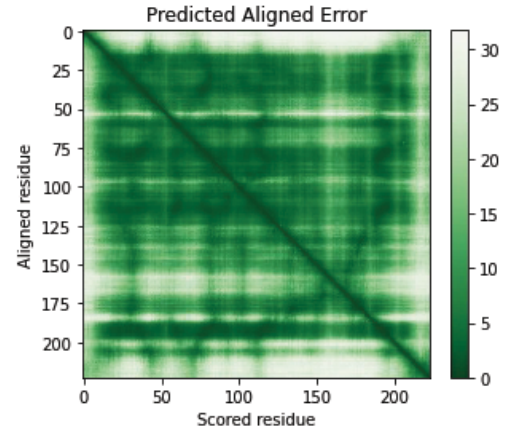
Get information from PDB & JSON file – draw pLDDT & PAE

```
#colab AlphaFold latest version result file
```

```
import json
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
```

```
with open('/content/predicted_aligned_error.json', 'r') as pae:
    info = json.load(pae)
    pae = info[0]['predicted_aligned_error']
    max_pae = info[0]['max_predicted_aligned_error']
```

```
plt.imshow(pae, vmin=0., vmax=max_pae, cmap='Greens_r')
plt.colorbar(fraction=0.046, pad=0.04)
plt.title('Predicted Aligned Error')
plt.xlabel('Scored residue')
plt.ylabel('Aligned residue')
plt.savefig('/content/PAE.png', dpi = 300)
```



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Get information from PDB & JSON file – draw pLDDT & PAE

```
#colab AlphaFold old version result file
```

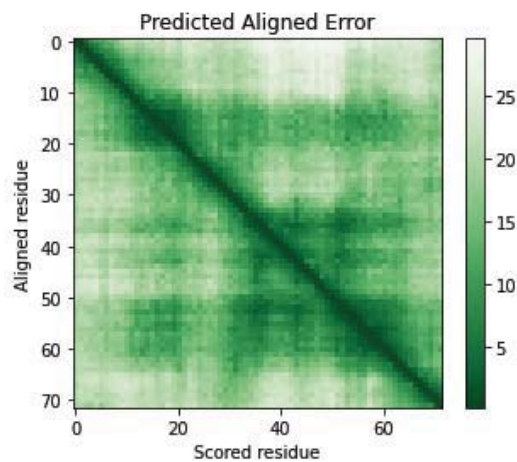
```
import json
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
```

```
with open('/content/predicted_aligned_error.json', 'r') as pae:
    info = json.load(pae)
    residue1 = info[0]['residue1']
    residue2 = info[0]['residue2']
    pae_score = info[0]['distance']
```

```
pae_array = np.ones((max(residue1), max(residue2)))
```

```
for x, y, z in zip(residue1, residue2, pae_score):
    pae_array[int(x-1), int(y-1)] = z
```

```
plt.subplot(1, 1, 1)
plt.imshow(pae_array, alpha=1, cmap='Greens_r')
plt.colorbar(fraction=0.046, pad=0.04)
plt.title('Predicted Aligned Error')
plt.xlabel('Scored residue')
plt.ylabel('Aligned residue')
plt.savefig('/content/PAE.png', dpi = 300)
```



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측

Get information from PDB & JSON file – draw pLDDT & PAE

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
```

#1. Wuhan-hu-1의 pLDDT 값 추출

```
wild_plddt = []
b = 1
```

```
with open('/content/wild.pdb', 'r') as file:
    for line in file:
        if line.startswith('ATOM'):
            number = int(line[22:26])
            if number == b:
                b+=1
                score = line[61:66].strip()
                wild_plddt.append(float(score))
            elif number == 1 and b > 2:
                b=1
                b+=1
                score = line[61:66].strip()
                wild_plddt.append(float(score))
```

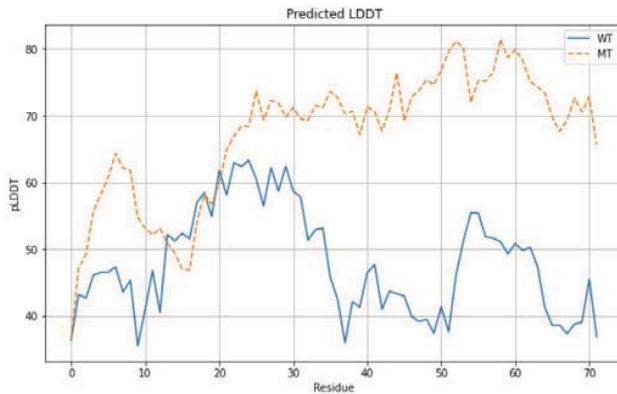
#2. delta의 pLDDT 값 추출

```
mutated_plddt = []
b = 1
```

```
with open('/content/mutated.pdb', 'r') as file:
    for line in file:
        if line.startswith('ATOM'):
            number = int(line[22:26])
            if number == b:
                b+=1
                score = line[61:66].strip()
                mutated_plddt.append(float(score))
            elif number == 1 and b > 2:
                b=1
                b+=1
                score = line[61:66].strip()
                mutated_plddt.append(float(score))
```

#3. 비교 그래프 그리기

```
plt.figure(figsize=[10, 6])
plt.plot(wild_plddt, linestyle='solid', label='WT')
plt.plot(mutated_plddt, linestyle='dashed', label='MT')
plt.legend()
plt.grid(True)
plt.title('Predicted LDDT')
plt.xlabel('Residue')
plt.ylabel('pLDDT')
plt.savefig('/content/compare_plddt.png', dpi=300)
```



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측 Get information from PDB & JSON file – view 3D structure



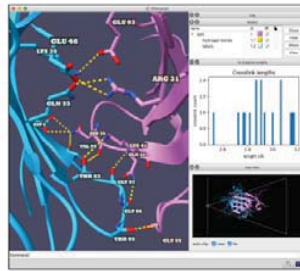
- Home
- ChimeraX Home
- FAQ
- Documentation
- User Guide
- Tutorials and Videos
- Presentations
- Download**
- Toolshed
- Citing ChimeraX
- Contact Us
- Related Software
- Featured Citations

UCSF ChimeraX

UCSF ChimeraX (or simply ChimeraX) is the next-generation molecular visualization program from the [Resource for Biocomputing, Visualization, and Informatics](#) (RBVI), following [UCSF Chimera](#). ChimeraX can be downloaded free of charge for academic, government, nonprofit, and personal use. Commercial users, please see [ChimeraX commercial licensing](#).

ChimeraX is developed with support from [National Institutes of Health](#) R01-GM129225, [Chan Zuckerberg Initiative](#) grant E0554-0000000439, and the Office of Cyber Infrastructure and Computational Biology, [National Institute of Allergy and Infectious Diseases](#).

Feature Highlight



Interactive H-Bond Histogram

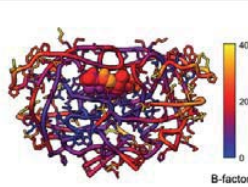
Hydrogen bonds (H-bonds) can be identified with the [H-Bonds](#) tool, [hbonds](#) command, or the [Molecule Display](#) icon and plotted as an interactive histogram with the command [crosslinks.histogram](#).

The ChimeraX graphics window shows the complex between a natural killer cell receptor 2B4 and its ligand CD48 (PDB [2qtl](#)). The receptor protein is blue; the ligand protein pink, and H-bonds between them dashed yellow, with H-bonding residues labeled. Although not done here, the H-bonds could also be labeled by distance.

The histogram of H-bond distances on the top right is interactive: when the cursor is placed over a bar in the histogram, the corresponding H-bonds are temporarily enlarged in the 3D view and the others hidden. For image setup other than orientation, see the command file [hb2.csg](#).

[More features...](#)

Example Image



B-factor Coloring

Atomic B-factor values are read from PDB and mmCIF input files and assigned as [attributes](#) that can be shown with [coloring](#) and used in [atom specification](#). This example shows B-factor variation within a structure of the HIV-1 protease bound to an inhibitor (PDB [4hrp](#)). For complete image setup, including positioning, [color key](#), and label, see the command file [bfactor.csg](#).

Additional color key examples can be found in tutorials: [Coloring by Electrostatic Potential](#), [Coloring by Sequence Conservation](#)

[More images...](#)



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측 Get information from PDB & JSON file – view 3D structure

Download UCSF ChimeraX

ChimeraX is the state-of-the-art visualization program from the [Resource for Biocomputing, Visualization, and Informatics](#) at UC San Francisco. It is free for academic, government, nonprofit, and personal use; commercial users, please see [commercial licensing](#). Please [cite ChimeraX](#) in publications.

See also: [Features and Missing Features](#), [Change Log](#), [System Requirements](#), [Older Releases](#), [Common Platform Problems](#), [Download & Citation Counts](#)

ChimeraX version 1.5

Production releases are stable versions for [ChimeraX Toolshed](#) bundles to work with. You may need to use an [older release](#) if a bundle you wish to use has not been updated yet. Showing releases for Windows 10.

Operating System	Distribution	Date	Notes
Windows	ChimeraX-1.5.exe	2022년 11월 23일	Download is a Windows installer. Tested on Windows 10 and Windows 11. ▶ More Info...

▶ Other releases

Daily Build

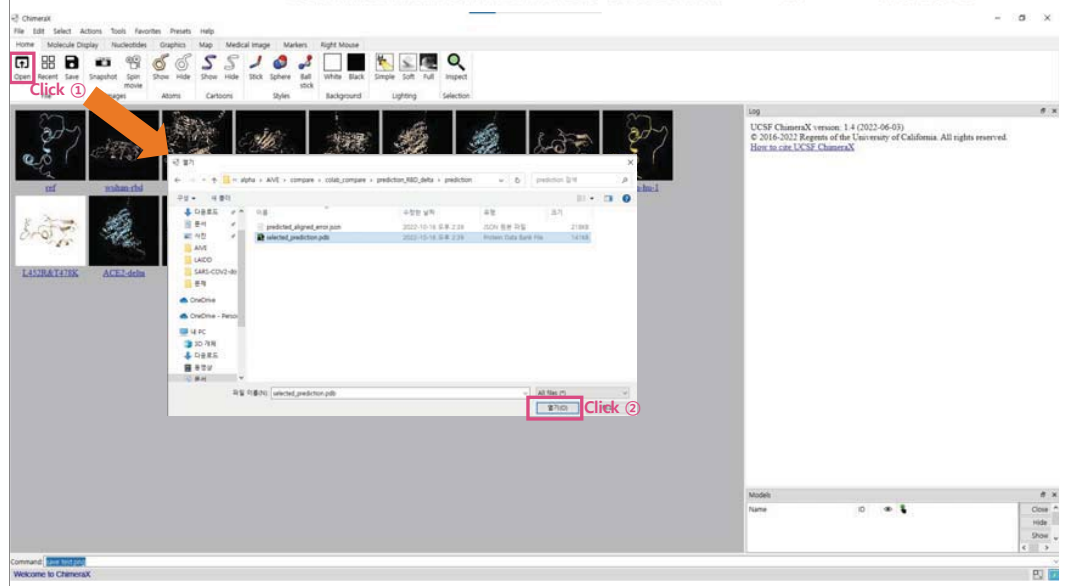
Daily builds are generated automatically each night from the [development source code](#) (see the [change log](#)). While a given build may have unforeseen problems, these are often fixed by the next day. Showing releases for Windows 10.

Operating System	Distribution	Date	Notes
Windows	chimeraX-daily.exe	2022년 11월 28일	Download is a Windows installer. Tested on Windows 10 and Windows 11. ▶ More Info...

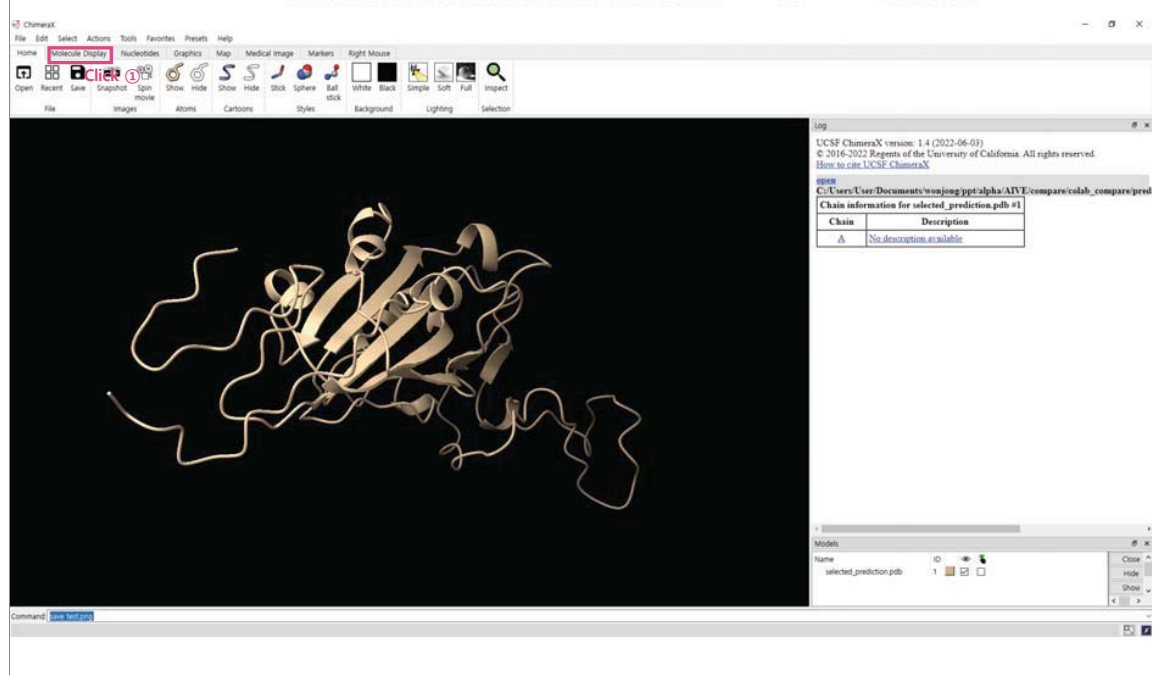
▶ Other releases



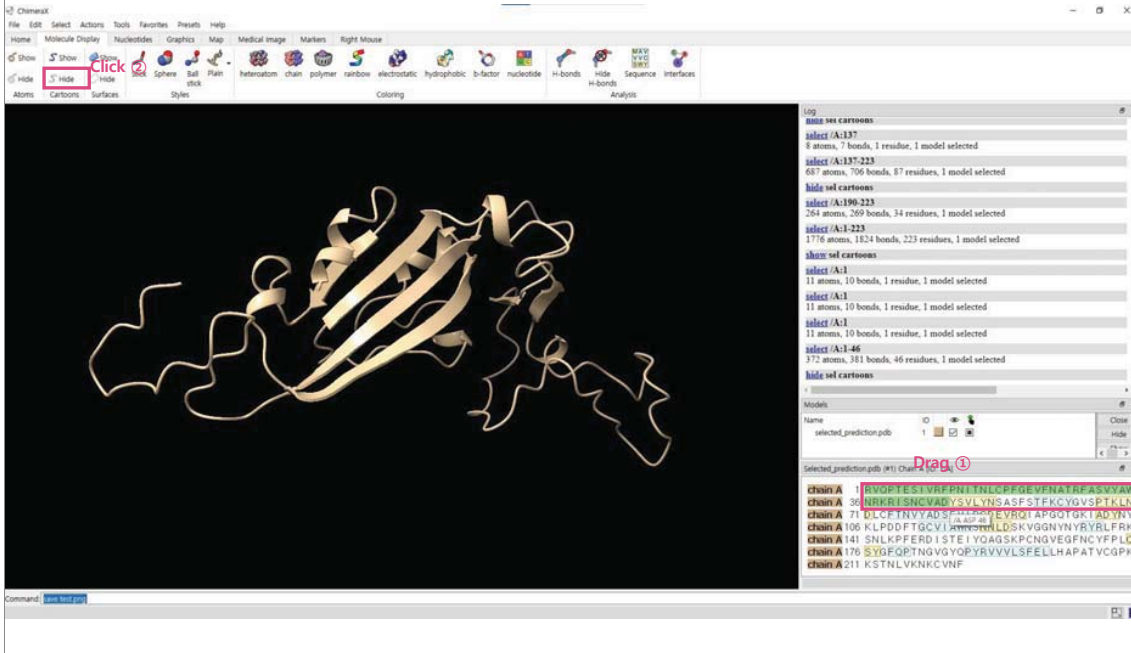
3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측 Get information from PDB & JSON file – view 3D structure



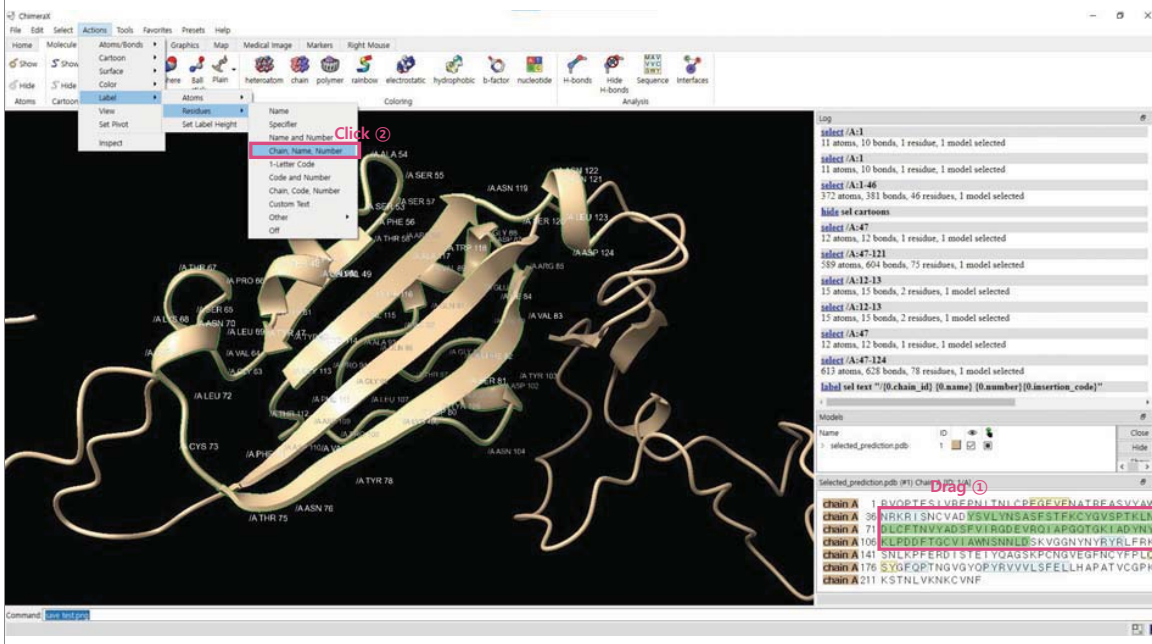
3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측 Get information from PDB & JSON file – view 3D structure



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측
Get information from PDB & JSON file – view 3D structure



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측
Get information from PDB & JSON file – view 3D structure



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측
Get information from PDB & JSON file – view 3D structure

ChimeraX interface showing a 3D protein structure. The toolbar includes a button for 'H-bonds' (highlighted with a red box and 'Click ②'). The command line at the bottom left displays '111 hydrogen bonds found'. The right-hand panel shows a list of selected residues and a sequence viewer with a red box around a specific sequence segment.



3 AlphaFold를 이용한 SARS-CoV-2의 단백질 구조 예측
Get information from PDB & JSON file – view 3D structure

ChimeraX interface showing a zoomed-in view of a protein structure. A red box highlights a specific residue 'ASP 149' with a 'Click ②' label. The right-hand panel shows the sequence viewer with a red box around the corresponding sequence segment.

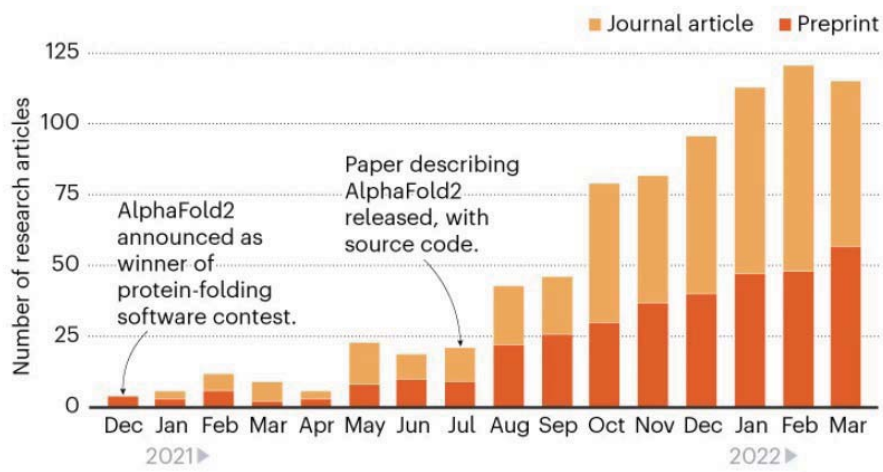


Contents

- 1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측
- 2 AlphaFold를 이용한 Protein- target 간의 상호작용의 친화도 측정 방법
- 3 Local computer에서 AlphaFold 사용 방법
- 4 AIVE 사용 설명

- 1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측

1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측



*Nature analysis using Dimensions database; removing duplicate preprints and papers/R. Van Noorden, E. Callaway.

©nature



1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측

nature reviews drug discovery

Explore content About the journal Publish with us Subscribe

nature > nature reviews drug discovery > news > article

NEWS | 14 September 2021

What does AlphaFold mean for drug discovery?

AlphaFold and RoseTTAFold have delivered a revolutionary advance for protein structure predictions, but the implications for drug discovery are more incremental. For now.

Archer Mullard



Source: Nature 596, 583–589 (2021) | Springer Nature Limited

arXiv:2201.09647v2 13 Feb 2022

Non-uniform motions → Uniform motions

Labroots: Drug Discovery & Development

Model Confidence:
 ■ Very high (pLDDT > 90)
 ■ Confident (pLDDT 80–90)
 ■ Low (pLDDT 50–80)
 ■ No prediction (pLDDT < 50)

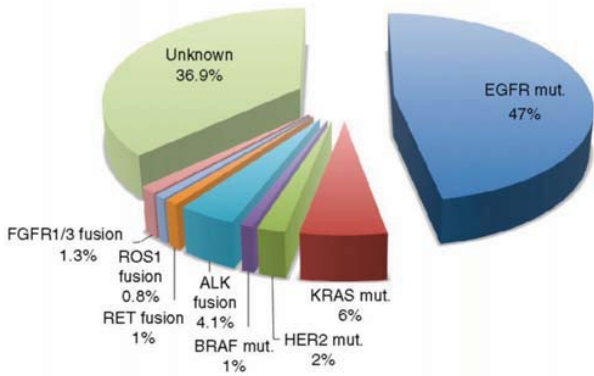
② Moving parts

③ The longer wish list

<https://pdb101.rcsb.org/motm/motm-about>



1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측

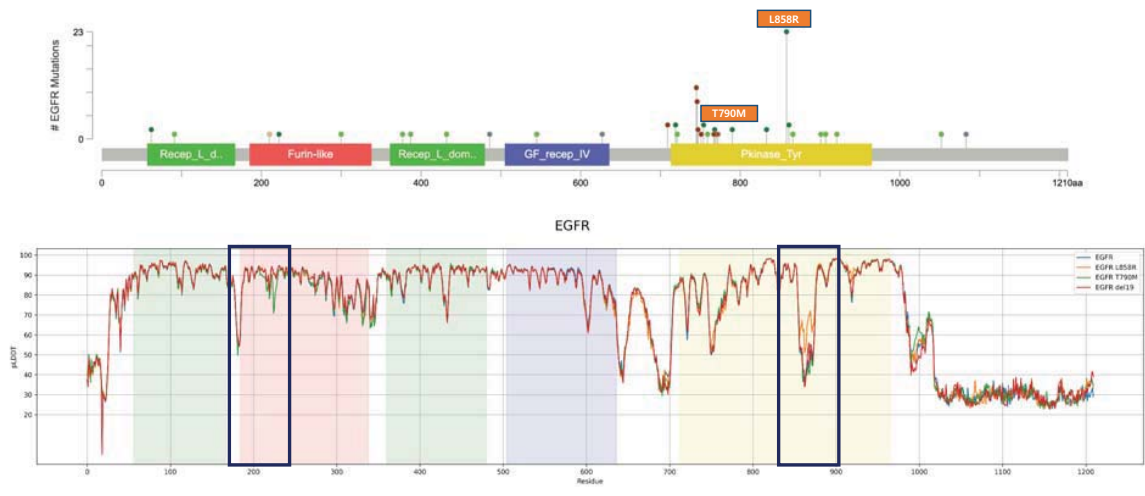


Cancer drugs

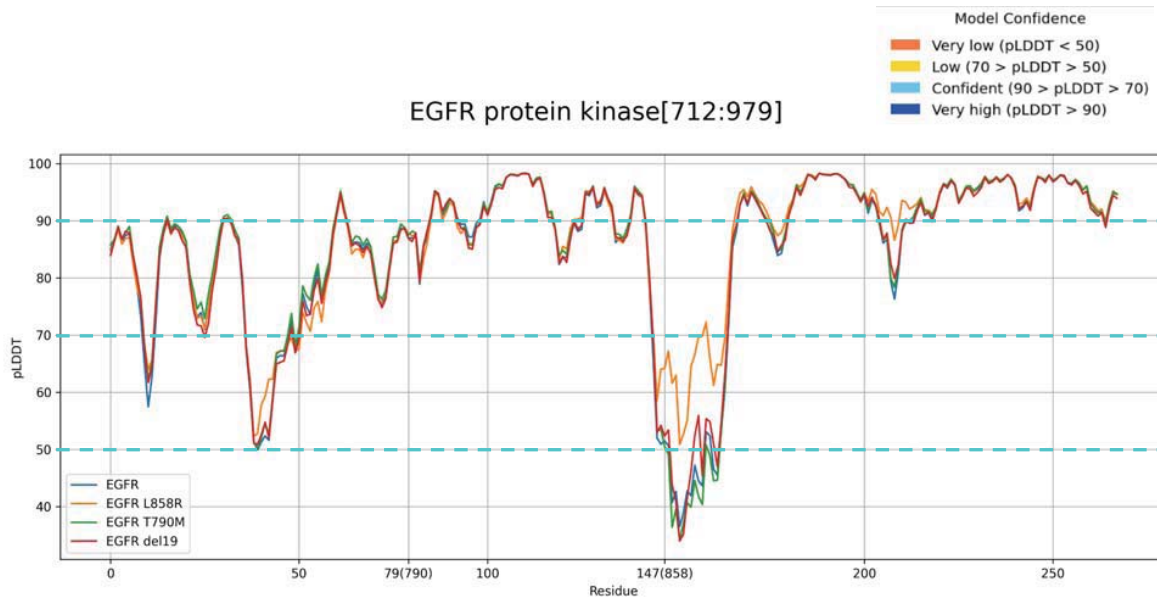
- Vemurafenib and trametinib
BRAF V600E mutations in melanoma
- Erlotinib and osimertinib
EGFR mutations in NSCLC; L858R, Del (19), T790M
- Pembrolizumab (immune checkpoint inhibitor): FDA approved
SOLID tumors from any tissue type
Mismatch repair deficiency (dMMR)
- Nivolumab, ipilimumab and atezolizumab
High tumor mutation burden

1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측

Lung Adenocarcinoma
(TCGA, PanCancer Atlas, 566 patients; EGFR is altered in 12% of patients)



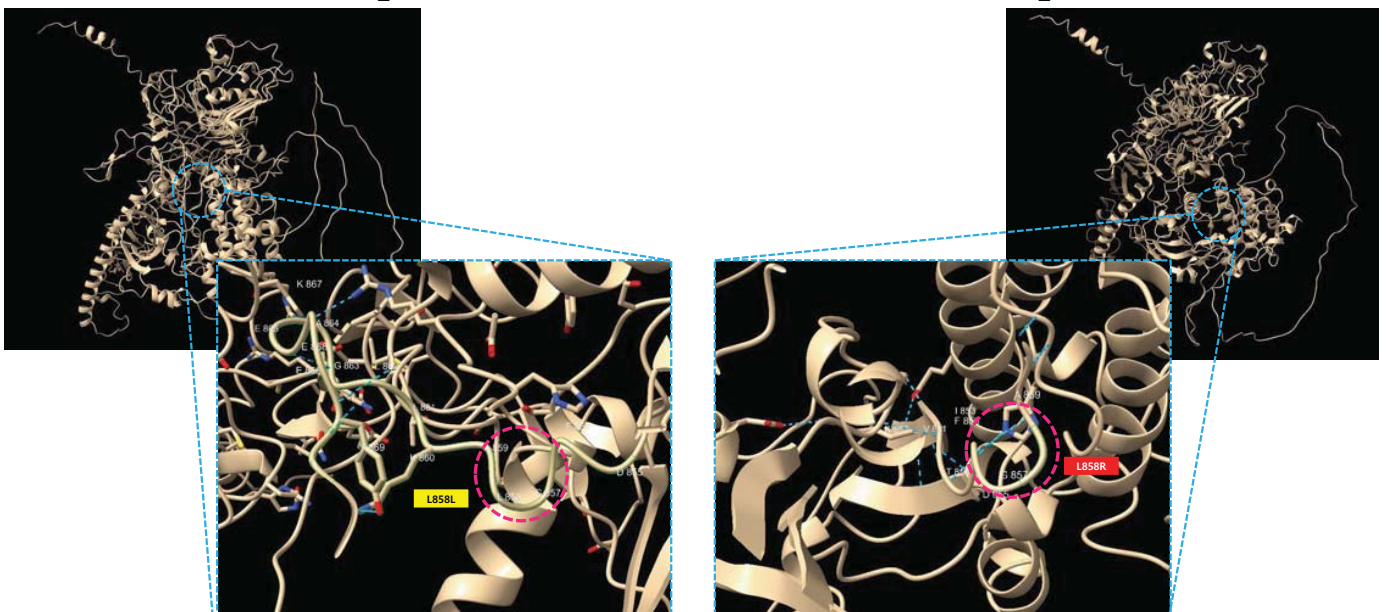
1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측



1 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측

EGFR_Wild

EGFR_L858R



2

AlphaFold를 이용한 Protein- target 간의 상호작용의 친화도 측정 방법

2 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측

Article | [Open Access](#) | [Published: 10 March 2022](#)

Improved prediction of protein-protein interactions using AlphaFold2

[Patrick Bryant](#) , [Gabriele Pozzati](#) & [Arne Elofsson](#) 

[Nature Communications](#) **13**, Article number: 1265 (2022) | [Cite this article](#)

32k Accesses | **31** Citations | **48** Altmetric | [Metrics](#)

FoldDock

This repository contains the simultaneous folding and docking protocol **FoldDock**.

The protocol has been developed on 216 heterodimeric complexes from [Dockground](#) and tested on [1481 heterodimeric complexes extracted from the PDB](#).

The protocol uses the recently published state-of-the-art end-to-end protein structure predictor [AlphaFold2](#) to predict the structure of heterodimeric complexes.

AlphaFold2 is available under the [Apache License, Version 2.0](#) and so is FoldDock, which is a derivative thereof.

The AlphaFold2 parameters are made available under the terms of the [CC BY 4.0 license](#) and have not been modified.

You may not use these files except in compliance with the licenses.

The success rate of the final protocol is 63% on the test set. By analyzing the predicted interfaces, we are able to distinguish accurate models with an AUC of 0.94 on the test set. For more information on this pipeline and its performance see [Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments](#)

2 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측

```

2분 [1] !git clone https://gitlab.com/ElofssonLab/FoldDock.git

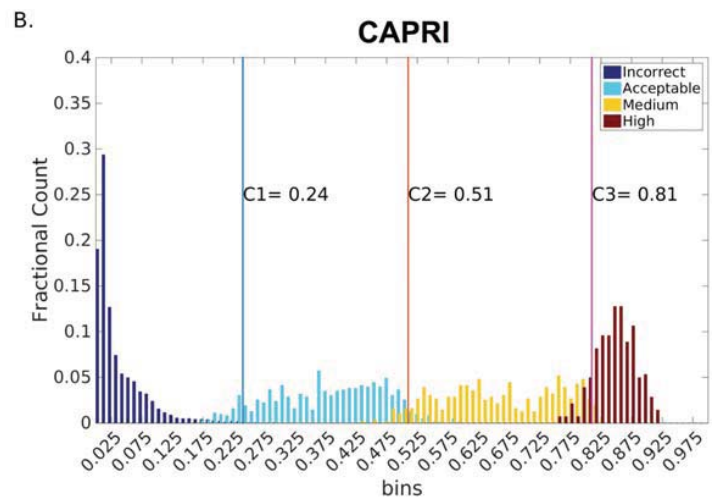
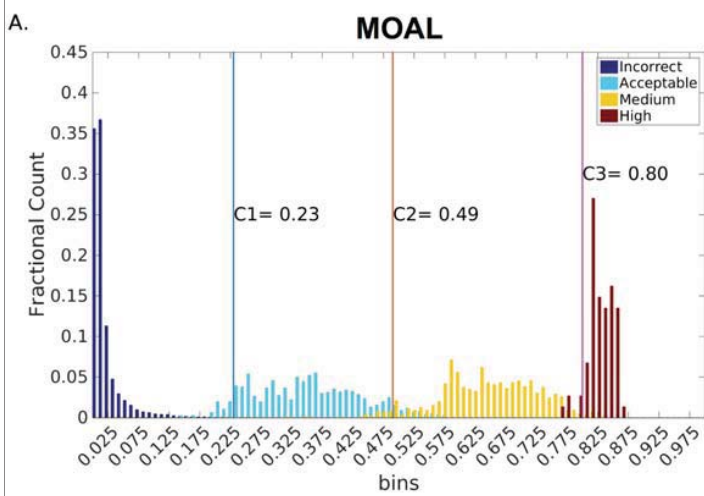
Cloning into 'FoldDock'...
remote: Enumerating objects: 30083, done.
remote: Counting objects: 100% (198/198), done.
remote: Compressing objects: 100% (190/190), done.
remote: Total 30083 (delta 89), reused 61 (delta 8), pack-reused 29885
Receiving objects: 100% (30083/30083), 2.94 GiB | 18.00 MiB/s, done.
Resolving deltas: 100% (5586/5586), done.
Checking out files: 100% (19182/19182), done.

0초 !python3 /content/FoldDock/src/pdockq.py --pdbfile /content/selected_prediction.pdb

pdockQ = 0.154 for /content/selected_prediction.pdb
This corresponds to a PPV of at least 0.71391784
    
```



2 Cancer에서 AlphaFold를 이용한 protein 3차원 구조 예측



3

Local computer에서 AlphaFold 사용 방법

3 Local computer에서 AlphaFold 사용 방법

Use alphafold in local PC – download alphafold

● Downloaded file list

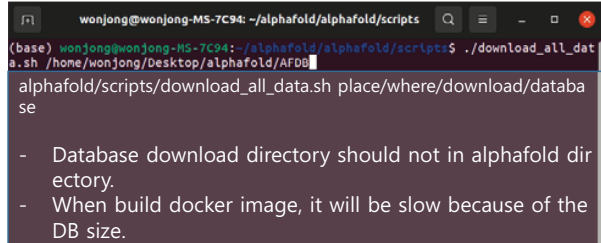
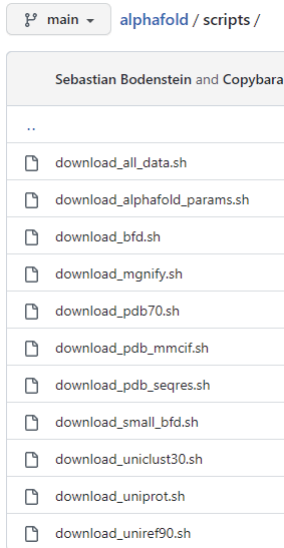
File Name	Description	Last Commit
afdb	Internal change	last month
alphafold	Internal change	last month
docker	Fix Dockerfile breakage due to google/jax#11142.	2 months ago
imgs	Initial release of AlphaFold.	14 months ago
notebooks	Add info for getting in touch via email to colab and readme.	6 months ago
scripts	Update multimer models to v2.	6 months ago
.dockernignore	Update .dockernignore	8 months ago
CONTRIBUTING.md	Initial release of AlphaFold.	14 months ago
LICENSE	Initial release of AlphaFold.	14 months ago
README.md	Clarify that AlphaFold works only under Linux.	3 months ago
requirements.txt	Pin protobuf version to 3.20.1.	3 months ago
run_alphafold.py	Remove unused type hint from run_alphafold script.	3 months ago
run_alphafold_test.py	Do not reuse the temporary output directory in run_alphafold_test.	6 months ago
setup.py	Add matplotlib test dependency so we can run notebook_utils_test.	3 months ago

<https://github.com/deepmind/alphafold>

```
wonjong@wonjong-MS-7C94: ~/Desktop/alphafold
(base) wonjong@wonjong-MS-7C94: ~/Desktop$ mkdir alphafold
(base) wonjong@wonjong-MS-7C94: ~/Desktop$ cd alphafold/
(base) wonjong@wonjong-MS-7C94: ~/Desktop/alphafold$ git clone https://github.com/deepmind/alphafold.git
```

3 Local computer에서 AlphaFold 사용 방법

Download alphafold Database



3 Local computer에서 AlphaFold 사용 방법

Install docker

```
#update advance packaging tool(apt) list
$ sudo apt update

#install apt-transport-https, ca-certificates, curl, gnupg, and lsb-release to setting d
ocker repository
$ sudo apt install apt-transport-https ca-certificates curl gnupg lsb-release

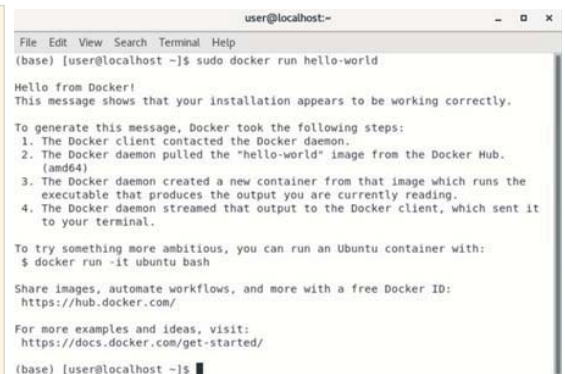
#generate docker GPG(GNU Privacy Guard) key
$ curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -
o /usr/share/keyrings/docker-archive-keyring.gpg

#setting docker repository to install docker. Arch=amd64 or arm64 (users env)
$ echo "deb [arch=amd64 signed-by=/usr/share/keyrings/docker-archive-keyring.g
pg] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable" | sudo tee /etc
/apt/sources.list.d/docker.list > /dev/null

#update apt list
$ sudo apt update

# install docker-ce, docker-ce-cli, and containerd.io
$ sudo apt install docker-ce docker-ce-cli containerd.io

#test docker
$ sudo docker run hello-world
```



3 Local computer에서 AlphaFold 사용 방법

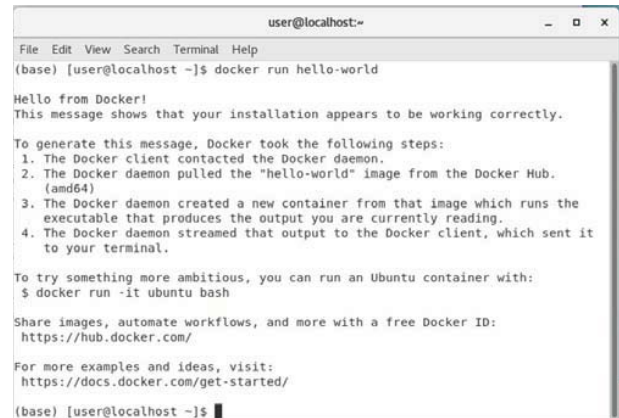
Install docker – add user

```
# add docker user group
$ sudo groupadd docker

#add user in docker user group
$ sudo usermod -aG docker $USERNAME

#update group infomation
$ newgrp docker

#test docker
$ docker run hello-world
```



```
user@localhost:~
File Edit View Search Terminal Help
(base) [user@localhost ~]$ docker run hello-world

Hello from Docker!
This message shows that your installation appears to be working correctly.

To generate this message, Docker took the following steps:
 1. The Docker client contacted the Docker daemon.
 2. The Docker daemon pulled the "hello-world" image from the Docker Hub.
    (amd64)
 3. The Docker daemon created a new container from that image which runs the
    executable that produces the output you are currently reading.
 4. The Docker daemon streamed that output to the Docker client, which sent it
    to your terminal.

To try something more ambitious, you can run an Ubuntu container with:
$ docker run -it ubuntu bash

Share images, automate workflows, and more with a free Docker ID:
https://hub.docker.com/

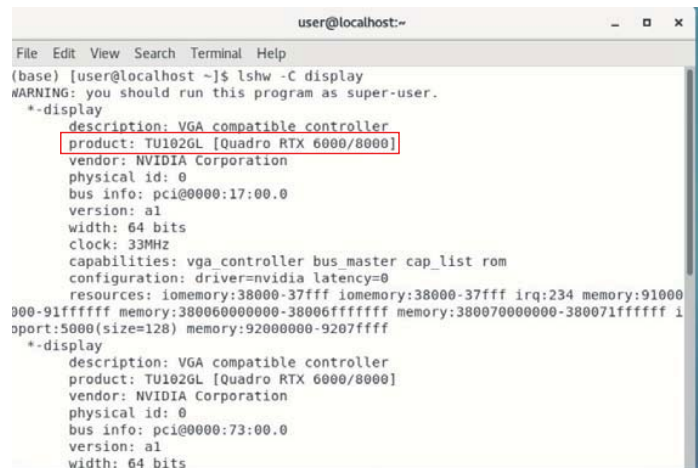
For more examples and ideas, visit:
https://docs.docker.com/get-started/

(base) [user@localhost ~]$
```

3 Local computer에서 AlphaFold 사용 방법

Install GPU driver

```
#check PC GPU product
$ lshw -C display
```



```
user@localhost:~
File Edit View Search Terminal Help
(base) [user@localhost ~]$ lshw -C display
WARNING: you should run this program as super-user.
*-display
  description: VGA compatible controller
  product: TU102GL [Quadro RTX 6000/8000]
  vendor: NVIDIA Corporation
  physical id: 0
  bus info: pci@0000:17:00.0
  version: a1
  width: 64 bits
  clock: 33MHz
  capabilities: vga_controller bus_master cap_list rom
  configuration: driver=nvidia latency=0
  resources: iomemory:38000-37fff iomemory:38000-37fff irq:234 memory:91000
            900-91ffffff memory:380060000000-38006ffffff memory:380070000000-380071ffffff i
            nport:5000(size=128) memory:92000000-9207ffff
*-display
  description: VGA compatible controller
  product: TU102GL [Quadro RTX 6000/8000]
  vendor: NVIDIA Corporation
  physical id: 0
  bus info: pci@0000:73:00.0
  version: a1
  width: 64 bits
```

3 Local computer에서 AlphaFold 사용 방법

Install GPU driver

```
#in ubuntu environment, ubuntu can check and auto install recommended GPU driver
$ ubuntu-drivers devices
$ ubuntu-drivers autoinstall
```

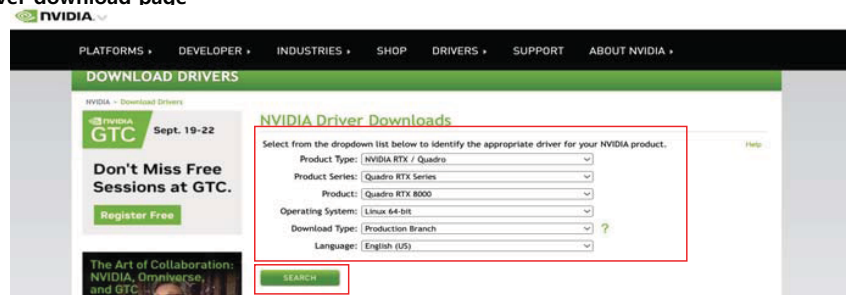
```
wonjong@wonjong-MS-7C94: ~/Desktop
(base) wonjong@wonjong-MS-7C94:~/Desktop$ ubuntu-drivers devices
WARNING:root: pkg_get_support nvidia-driver-390: package has invalid Support Legacyheader, cannot determine support level
== /sys/devices/pci0000:00/0000:00:03.1/0000:2b:00.0 ==
modalias : pci:v000010DEd0000128Bsv00001043sd00008770bc03sc00i00
vendor : NVIDIA Corporation
model : GK208B [GeForce GT 710] (GT710-4H-SL-2GD5)
driver : nvidia-driver-390 - distro non-free
driver : nvidia-driver-470 - distro non-free recommended
driver : nvidia-driver-470-server - distro non-free
driver : nvidia-driver-418-server - distro non-free
driver : nvidia-driver-460-server - distro non-free
driver : nvidia-driver-450-server - distro non-free
driver : nvidia-driver-460 - distro non-free
driver : xserver-xorg-video-nouveau - distro free builtin

(base) wonjong@wonjong-MS-7C94:~/Desktop$ ubuntu-drivers autoinstall
```

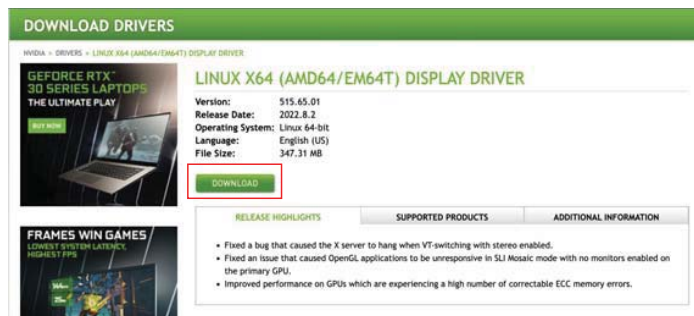
3 Local server에서 AlphaFold 사용 방법

Install GPU driver from nvidia driver download page

Step ①



Step ②



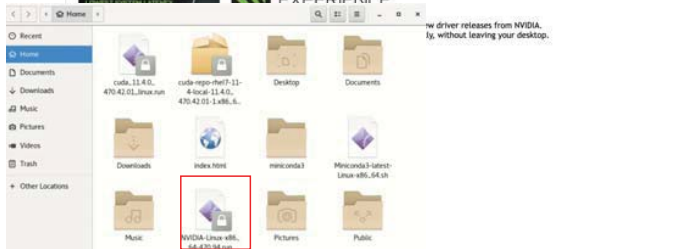
3 Local computer에서 AlphaFold 사용 방법

Install GPU driver from nvidia driver download page

Step ③



Step ④



Step ⑤

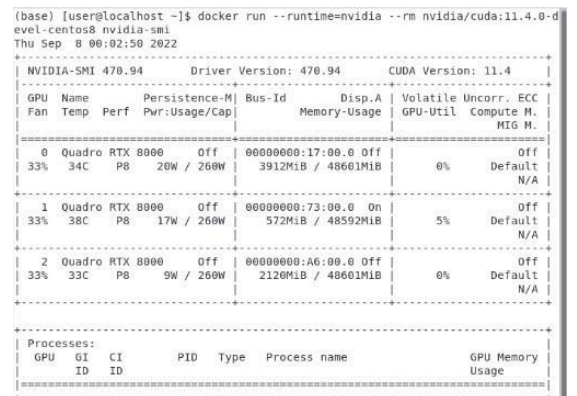


3 Local computer에서 AlphaFold 사용 방법

Install Nvidia container toolkit for using gpu in docker container

Install nvidia container toolkit

```
$ distribution=$(. /etc/os-release;echo $ID$VERSION_ID) && \
curl -s -L https://nvidia.github.io/nvidia-docker/gpgkey | sudo apt-key add - && \
\
curl -s -L https://nvidia.github.io/nvidia-docker/$distribution/nvidia-docker.list | \
sudo tee /etc/apt/sources.list.d/nvidia-docker.list
$ curl -s -L https://nvidia.github.io/nvidia-container-runtime/experimental/$distrib \
ution/nvidia-container-runtime.list | sudo tee /etc/apt/sources.list.d/nvidia-contain \
er-runtime.list
$ sudo apt update
$ sudo apt install -y nvidia-docker2
$ sudo service docker restart
$ docker run --rm --gpus all nvidia/cuda:11.1-base nvidia-smi
```



3 Local computer에서 AlphaFold 사용 방법

Build docker image

```
Build docker image and install docker require packages
#docker build -f (use dockerfile) -t (docker image name)
$ docker build -f docker/Dockerfile -t alphafold .
#install requirement packages to use alphafold
$ pip3 install -r docker/requirements.txt
#check installed docker image
$ docker images
```

```
main | alphafold / docker /
Augustin-Zidek and Copybara-Service Pin Conda version to 4.13.0. ...
...
Dockerfile Pin Conda version to 4.13.0.
openmm.patch Initial release of AlphaFold.
requirements.txt Initial release of AlphaFold.
run_docker.py No longer use ruetiae-nvidia to expose GPUs to the container.

(base) [user@localhost docker]$ docker images
REPOSITORY TAG IMAGE ID CREATED
alphafold latest e8d8550291c8 13 hours ago
12.6GB
```

3 Local computer에서 AlphaFold 사용 방법

Running alphafold in local pc

Running alphafold code

```
python3 /data/AF/alphafold/docker/run_docker.py
--fasta_paths=input_fasta_path_path ₩
--max_template_date=used_template_date ₩
--model_preset=(monomer/monomer_tpm/multimer) ₩
--num_multimer_predictions_per_model=num_of_seeds_per_model (only multimer, default = 5) ₩
--data_dir=Database_path ₩
--db_preset=reduced_dbs (if use reduced database) ₩
--docker_user=0 ₩
--gpu_devices=id_of_using_gpu ₩
--output_dir=result_saved_path
```

-> result file = .pdb/.pkl file

3 Local computer에서 AlphaFold 사용 방법

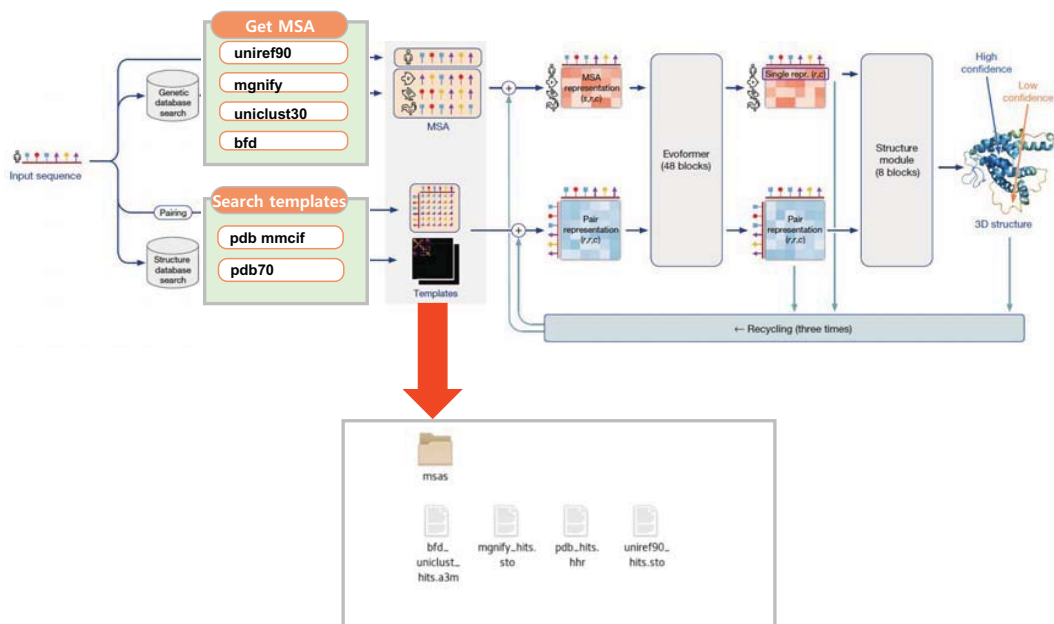
Running alphafold in local pc

Input file : monomer seq fasta file & multimer seq fasta file

```
Example Monomer file  
> sars-cov-2 ace2 receptor bind domain  
NSNNLDSKVGGNYNLYRLFRKSNLKPFRDISTEIYQAGSKPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
```

```
Example Multimer file  
> sars-cov-2 ace2 receptor bind domain  
NSNNLDSKVGGNYNLYRLFRKSNLKPFRDISTEIYQAGSKPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY  
> sp|Q9BYF1|ACE2_HUMAN Angiotensin-converting enzyme 2  
MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNAGDKWSAFLKEQSTLA  
QMYPLQEIQLTVKL...
```

3 Local computer에서 AlphaFold 사용 방법



3 Local computer에서 AlphaFold 사용 방법



format of bfd_uniclust_hits.a3m

Input AA sequence

```
> sp|P00533|EGFR_HUMAN Epidermal growth factor receptor OS=Homo sapiens OX=9606 GN=EGFR PE=1 SV=2
MRPSTGAGALLLALLAALCPASRALEEKVCQGTNSNKLTLQFTFEDHFLSLQRMFNNECVLGNLEITYVQRNYDLSFLKTIQEVAGYVLIANTVERIPLENLQIRGNMYEENSYALAVLSNDYANKTKGLKELPMRNLQELHGVAFVSNPALCNVSIQWRDIVSDFLSNM
SMDFQNLHLSGSCQKCDPSPNGSCWGAEGEENCKLTKIICQQSCGRGKSPSDCCNQAAGCTGPRESDCLVCRKFRDEATCKDTCPLMLYNTPTTYQMDVNEPVGKYSFGATCVKCKPRNYVTDHGSVCVRACGADSYEMEEDGVRKCKCKEGPCRVKCVNGIGIEFK
DLSINATNIKHFKNCTSIISGDLHLPVAFRQDFTHTPLDPQELDKTKVKEITGFLLIQAWPENRTDLHAFENLEIIRGRKQHGQFSLAVVSLNITSLGLRSLKEISDGDVVISGNKNCYANTINWKKLFGTSGQTKIISNRGENSCKATGQVCHALCSPGECWGPEDRDCVSCR
NVSRGREDVCNKDLLEGEPRFVENSIECQHPCELPQAMNITCTGRGPNDCIQCAHYIDGPHCVKTCAGVGMENNTLVWYKADAGHVCHLCPNCTYGTGPGLEGPTNGPKPSIATGMVGLALLLVLALGILGFMRRRHIVRKRTRLLRLLQERELVEPTPSGEAPNQ
ALLRLIKETEFKIKVLGSGAFGTYYKGLWIPEGEKVIPIVAIKELREATSPKANKELDEAYVMASVDNPHVCRLLGICLTSTVQLTQLMPPGCLLDYVREHKDNIQSQYLLNWCVQIAKGMNYLEDRLVHRDLAARNVLTQPQHVKITDFGLAKLLGAEKEYHAEGKVPKW
MALESILRHRYTHQSDVSWYGVTVWELMTFGSKPYDGPASEISSLKGERLPQPPICTIDVYIMVVKWIMIDADSRPKFRELIEFSKMARDPQRYLVIQGDERMHLPSPTDSNFYRALMDEEDMDVDAVEYLPQQGFSSPSTRPILLSLATSANNSTVACIDRNLGQS
CPKIEDFSRLQRYSSDPTGALTEDSIDTFLVPEVINQVSPKRPAGSVQNVVYHNQPLNAPSRDPHYQDPHSTAVGNPEVLYNTQPTCVNSTFDSPAHWAKGQSHQISLONDPYQQDFFPKPEAKPNIGFKGSAENAELRVAPQSEFIGA
```

MSA

```
> tr|A0A218UMP6|A0A218UMP6_9PASE Receptor protein-tyrosine kinase OS=Lonchura striata domestica OX=299123 GN=ERBB2 PE=3 SV=1
-----AAAEVCTGDMKLLRPSSPESHYETLRHLRYQGCVVQGNLELYLPADADTSFLKDIKEVQGYVLIENQVSGALQLSLRIRGTQLFQDRYALAVVGNAGpT-gTPGLRQLGMRHLTELKGGVRIERNPELFCQETILWSDILH----
RQNEFRAEiqvesarsrsc-----PDCRALCAEGHCWGEKQDCQLTNSICH-GC-
PRCKGTPTDCCHEQCAAGCTGPKHSDCLACLHFNRSIGELHCPPLMNVYNSDTFESMPNPDGRYTFGASCVSQCPYNYLATEVGSCTLVCPQNSQEVTVNNVQKCEKCKPCPEVCYGLGVDFLKGVRVAVNASNIHQSFQKCTIGFSLAFETPAGDPSTNTPPLDKLLRF
ESLEELTGFLYIAAWPPDMKDLGVFNLRVIRGRVLRHNGAYSLLREI-
AVQALGLRALQEISSGMVLVHHPNLCFLQKVPWHSIFRNPRQLRQFTHNPKPEQCESEGLVCFHLCAGHGWCPGPTQVCACERFLRQECVASCNLLDGVAREHANGTRCLPCHPECQPngt---
ETCFGSDPDQVACAHYKDAQCVRRCPSPGVKADAsFVWVWYKYPDEFQVCLCPTNCTHSTIRDEDCPVQKpSQVTSIIAGVVGALLVIVLLITVICVRRRQQRKHMTMRLLQETELVEPLTPSGALPNQAQMRILKTELKVKVVLGSGAFGTYYKGIWIPDGESVKIPV
AIKVLRENTSPKANKELDEAYVMAGVSPYVSRLLGICLTSTVQLVQLMPPYGLLDYVRENKDRIGSQDLNWCVQIAKGMVSYLEEVRLVHRDLAARNVLTQSPNHVKITDFGLARLLDIDETEHADGGKVPKWMALESILRRRFTHQSDVSWYGVTVWELMTFGAKPYDG
IPAREPILDEKGERLPQPPICTIDVYIMVVKWIMIDSECRPKFRELVEFSRMARDPQRVFIQNDLVGLPGS-MDSTFYRALLDEEDMDVDAEYLVPHGFFSTDTSTYRSRISMRSTAsPaKVE-egglaaafpaaglaegpegpvevpgdkVA--LQSP-
GREPGTLPRYSEDPGLTA-kdgedpecFTVPAPHSAMPEYVYVQAGE---RPPRAPPSPDKPKGHQKngliKDKNSFPGFgha--VENPEYLAPHA--P-----APFSQAFDNPPYYVWQPAK--aggPEGGPGTPTTAENPEYLAGLGPDTTAV--
> tr|A0A0F7Z597|A0A0F7Z597_CROAD Receptor protein-tyrosine kinase OS=Crotalus adamanteus OX=8729 PE=2 SV=1
-----STTEVCTGDMKLLQPPSPENHPDTRRLRYEQCVVQGNLELYLPADVTSFLKDIKIREVQYVLIKNNHMHHLNLSLIIRGTQLYEYKVALVDNADsS-gDVGLEQLGMLKTELKGSVIERNPQLCFQETIHWEAIFHKYVW----
QKYteisshrnrc-----PDCSQCIPHNNHCWGEAKGSCQLTSTICASS-
PRCKGGHPDTCHEQCAAGCTGPKHSDCLACLHFNHSIGELHCPPLMNVNPDTEIMHNPNGRYTFGASCVPHCPYNYLAAEVGSCTLVCPQNSQEVSVIGAMQKCEKCDSSCEVCYGLGMDFLKGVRVAVNASNIHHTGTCTKIFGSLAFKETFAGDPATNTPPLQLEQL
VFWHLKDLTGLFVIESWPETLLDLSFQNLQVIRGRALYNGAYSLMLQNI-
NISSLGLRSLQEISSGMVLHHPNLCFLQKVPWDDIFRNPRQLRQFTHNPKPEQCESEGLVCFHLCAGHGWCPGPTQVCACERFLRQECVASCNLLDGVAREHANGTRCLPCHPECMPQngt---
ESCYGSEADQCIACAHYKDGKLSVCVERCPSPGVKLDGgFIPWVWYKYPDEFQVCLCPTNCTHSTIRDEDCPVQKpSHASSIAGVVSAAALALLLLITVICINRRKQQRKHMTMRLLQETELVEPLTPSGALPNQAQMRILKTELKVKVVLGSGAFGTYYKGIWIPDGENVKIPV
IKVLRENTSPKANKELDEAYVMAGVSPYVSRLLGICLTSTVQLVQLMPPYGLLDYVRENKDRIGSQDLNWCVQIAKGMVSYLEEVRLVHRDLAARNVLTQSPNHVKITDFGLARLLDIDETEHADGGKVPKWMALESILRRRFTHQSDVSWYGVTVWELMTFGAKPYDGI
PAREPILDEKGERLPQPPICTIDVYIMVVKWIMIDSECRPKFRELVEFSRMARDPQRVFIQNDLVGLPGS-MDSTFYRALLDEEDMDVDAEYLVPHGFFSTDTSTYRSRISMRSTAsPaKVE-egglaaafpaaglaegpegpvevpgdkVA--LQSP-
eteesppqpylgspsseesisesitreeedyTASKaVQNSAFLQaDCSLQRYSEDPSTASLNkenepndrksYTTPLTAVIPEYINQENNL-ppnKRLRGLSTGKQKRHMgkngliKEPKNAFYTSfsA--VENPEYLTCPSPAS-----
SPLPQAFDNLYVWQENPKCNpaefatsivppaATNGFPPTTAENLEVLGLSEPTCSKDF
```



3 Local computer에서 AlphaFold 사용 방법



bfd_uniclust_hits.a3m의 format

>db|Uniqlentifier|EntryName ProteinName OS=OrganismName OX=OrganismIdentifier [GN=GeneName]PE=ProteinExistence SV =SequenceVersion

- db : sp = UniProtKB/Swiss-Prot | tr = UniProtKB/TrEMBL
- Uniqlentifier : uniprot ID /accession number
- EntryName ProteinName : entry name of the UniProtKB entry
- OS : scientific name of the organism of the UniProtKB entry(생물명)
- OX : unique identifier of the source organism, assigned by the NCBI(생물 식별번호)
- GN : gene name

- PE : numerical value describing the evidence for the existence of the protein [1:5]
 1. Experimental evidence at protein level : The value 'Experimental evidence at protein level' indicates that there is clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein-protein interaction or detection of the protein by antibodies.
 2. Experimental evidence at transcript level : The value 'Experimental evidence at transcript level' indicates that the existence of a protein has not been strictly proven but that expression data (such as existence of cDNA(s), RT-PCR or Northern blots) indicate the existence of a transcript.
 3. Protein inferred from homology : The value 'Protein inferred by homology' indicates that the existence of a protein is probable because clear orthologs exist in closely related species.
 4. Protein predicted : The value 'Protein predicted' is used for entries without evidence at protein, transcript, or homology levels.
 5. Protein uncertain : The value 'Protein uncertain' indicates that the existence of the protein is unsure.

SV : sequence version



3 Local computer에서 AlphaFold 사용 방법

mgnify_hits.sto



mgnify_hits.sto

```
# STOCKHOLM 1.0
#=GS sp|P00533|EGFR_HUMAN DE Epidermal growth factor receptor OS=Homo sapiens OX=9606 GN=EGFR PE=1 SV=2
#=GS MGYP000983186742/226-829 DE [subseq from] PL=00 UP=0 BIOMES=000000100000
#=GS MGYP000983186742/1016-1366 DE [subseq from] PL=00 UP=0 BIOMES=000000100000
#=GS MGYP001020618282/199-789 DE [subseq from] PL=00 UP=0 BIOMES=000000100000
.
.
.
.
.
.
sp|P00533|EGFR_HUMAN MRPSGTAGAALLAALLCPASRALEEKVCGQTSNKLTQLGT-FEDHFLSLQRMFNCEVVLGNLEITYVQ----R-----N----Y----DLSF-LKT-----IQEVACY-VLI-ALNTVE--RIP--LENLQIRG-NMYY-----E-----N---SYAL
MGYP000983186742/226-829 -----LKKICIGTSHRMSSGQHN-KTDHYQNLAEYRNCTYVIGNLEITWLD----KAASGEQ-----V----DLSF-LES-----IREITGY-LLI-AYVEVE--RIR--MPNLQIRG-RDLF-----K---LN---SDQKEEF---AI
MGYP000983186742/1016-1366 -----EKICIGTANRMSSGQHN-KTDHYQNLAEYRNCTYVIGNLEITWLE----K-----GPDNKPV---DLTF-LES-----IREITGY-LLI-AYVEVE--KIR--MPNLQIRG-RDLF-----R---LN---NGQKEEF---AI
MGYP001020618282/199-789 -----
MGYP001020618282/983-1329 -----
```

uniref90_hits.sto



uniref90_hits.sto

```
# STOCKHOLM 1.0
#=GS sp|P00533|EGFR_HUMAN DE Epidermal growth factor receptor OS=Homo sapiens OX=9606 GN=EGFR PE=1 SV=2
#=GS UniRef90_P00533/1-1210 DE [subseq from] Epidermal growth factor receptor n=96 Tax=Boreoeutheria TaxID=1437010 RepID=EGFR_HUMAN
#=GS UniRef90_A0A2I3GX32/20-1206 DE [subseq from] Receptor protein-tyrosine kinase n=3 Tax=Nomascus leucogenys TaxID=61853 RepID=A0A2I3GX32_NOMLE
#=GS UniRef90_A0A2R9CTL6/11-1201 DE [subseq from] Receptor protein-tyrosine kinase n=4 Tax=Pan TaxID=9596 RepID=A0A2R9CTL6_PANPA
.
.
.
.
.
.
sp|P00533|EGFR_HUMAN MRPSG-TAG-A-A-L-L---A-L-L---A-A-L-L---C--P-A---S-----R---A-----L---E---E---K-K-V-CQG---T-S-N-K---L-T-Q---L---G--TF--E-D--HFLS-L-Q-----R-MF-N---NCEVVL---G--NLEI
UniRef90_P00533/1-1210 MRPSG-TAG-A-A-L-L---A-L-L---A-A-L-L---C--P-A---S-----R---A-----L---E---E---K-K-V-CQG---T-S-N-K---L-T-Q---L---G--TF--E-D--HFLS-L-Q-----R-MF-N---NCEVVL---G--NLEI
UniRef90_A0A2I3GX32/20-1206 -----P-V-CQG---T-S-N-K---L-T-Q---L---G--TF--E-D--HFLS-L-Q-----R-MF-N---NCEVVL---G--NLEI
UniRef90_A0A2R9CTL6/11-1201 -----L-L---S-Q---S-L-L---C--P-L---V-----R---A-----L---E---E---K-K-V-CQG---T-S-N-K---L-T-Q---L---G--TF--E-D--HFLS-L-Q-----R-MF-N---NCEVVL---G--NLEI
UniRef90_A0ABB7H412/1-1208 MRPSG-TAA-A-A-L-L---V-L-L---A-A-L-L---V--Q-A---S-----R---A-----L---E---E---K-K-V-CQG---T-S-N-K---L-T-Q---L---G--TF--E-D--HFLS-L-Q-----R-MF-N---NCEVVL---G--NL
```



3 Local computer에서 AlphaFold 사용 방법

mgnify_hits.sto
의 format



mgnify_hits.sto

```
# STOCKHOLM 1.0
#=GS sp|P00533|EGFR_HUMAN DE Epidermal growth factor receptor OS=Homo sapiens OX=9606 GN=EGFR PE=1 SV=2
#=GS MGYP000983186742/226-829 DE [subseq from] PL=00 UP=0 BIOMES=000000100000
#=GS MGYP000983186742/1016-1366 DE [subseq from] PL=00 UP=0 BIOMES=000000100000
#=GS MGYP001020618282/199-789 DE [subseq from] PL=00 UP=0 BIOMES=000000100000
```

uniref90_hits.sto
의 format



uniref90_hits.sto

```
# STOCKHOLM 1.0
#=GS sp|P00533|EGFR_HUMAN DE Epidermal growth factor receptor OS=Homo sapiens OX=9606 GN=EGFR PE=1 SV=2
#=GS UniRef90_P00533/1-1210 DE [subseq from] Epidermal growth factor receptor n=96 Tax=Boreoeutheria TaxID=1437010 RepID=EGFR_HUMAN
#=GS UniRef90_A0A2I3GX32/20-1206 DE [subseq from] Receptor protein-tyrosine kinase n=3 Tax=Nomascus leucogenys TaxID=61853 RepID=A0A2I3GX32_NOMLE
#=GS UniRef90_A0A2R9CTL6/11-1201 DE [subseq from] Receptor protein-tyrosine kinase n=4 Tax=Pan TaxID=9596 RepID=A0A2R9CTL6_PANPA
```

#=GS : Metadata make-up line. Data relating to a specific sequence in the multiple sequence alignment. Starts with #=GS followed by the sequence name followed by a feature name and data relating to the feature. Typically comes after GF metadata in a Stockholm file.
taxid : NCBI taxonomy ID currently attached to the AnnotationDb objects database.



3 Local server에서 AlphaFold 사용 방법

Get information from Pkl file

```

[0] WT_ranking0
  "aligned_confidence_probs"
  "distogram"
  "experimentally_resolved"
  "masked_lm_data"
  "max_predicted_aligned_error"
  "pLDDT"
  "predicted_aligned_error"
  "predicted_iddt"
  "pTSA"
  "ranking_confidence"
  "structure_module"

[1] WT_ranking0["oiddt"]

array([[0.98031104, 0.61947461, 0.610561773, 0.626237255, 0.676994348,
        0.706443221, 0.717886511, 0.709204344, 0.59001456, 0.529839997,
        0.49683812, 0.50530915, 0.29188408, 0.588962527, 0.00363676,
        0.15130071, 0.38821194, 0.72451150, 0.30056627, 0.50019101,
        0.07898172, 0.30567928, 0.67006151, 0.736123875, 0.72333991,
        0.0725129, 0.61556595, 0.72331492, 0.84829977, 0.73943813,
        0.74881836, 0.7640726, 0.37958201, 0.70391687, 0.22523062,
        0.700593, 0.15079237, 0.25844303, 0.43220542, 0.66830366,
        0.6608185, 0.820331, 0.69434982, 0.37247905, 0.74839131,
        0.22361482, 0.73630478, 0.00089194, 0.32067242, 0.656316617,
        0.10006421, 0.47483827, 0.66318809, 0.96000889, 0.29783215,
        0.77074893, 0.71150150, 0.73842917, 0.4512907, 0.7545888,
        0.68626206, 0.748281467, 0.8512781, 0.52242321, 0.09821692,
        0.63665824, 0.70267401, 0.679500473, 0.32296578, 0.60743298,
        0.04286846, 0.37925208])

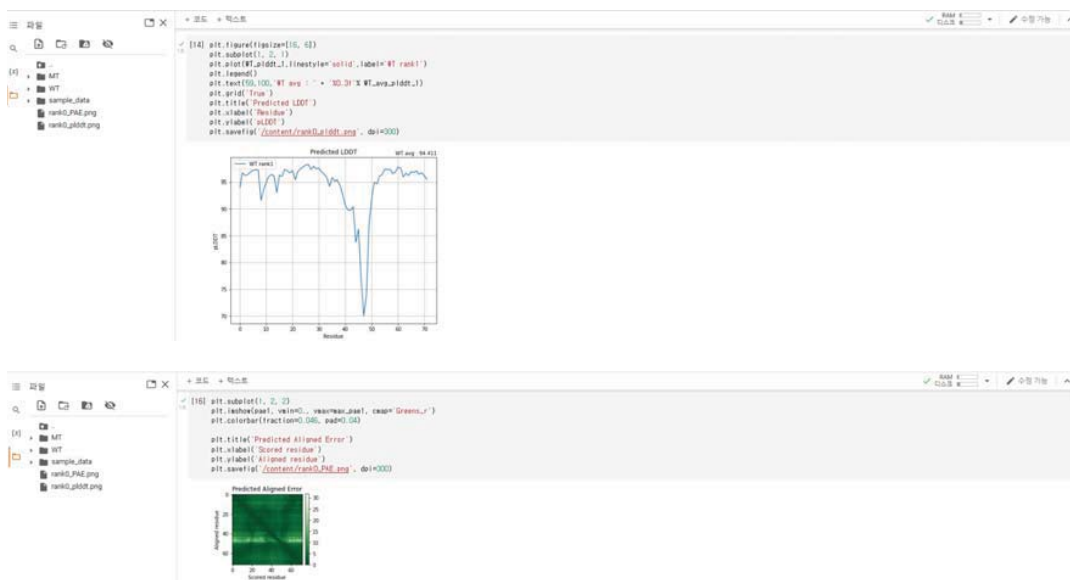
[12] WT_pLDDT_1 = WT_ranking0["oiddt"]
     WT_pLDDT_2 = WT_ranking1["oiddt"]
     WT_pLDDT_3 = WT_ranking2["oiddt"]
     WT_pLDDT_4 = WT_ranking3["oiddt"]
     WT_pLDDT_5 = WT_ranking4["oiddt"]
     WT_PAE_1 = (WT_ranking0["predicted_aligned_error"], WT_ranking0["max_predicted_aligned_error"])
     WT_PAE_2 = (WT_ranking1["predicted_aligned_error"], WT_ranking1["max_predicted_aligned_error"])
     WT_PAE_3 = (WT_ranking2["predicted_aligned_error"], WT_ranking2["max_predicted_aligned_error"])
     WT_PAE_4 = (WT_ranking3["predicted_aligned_error"], WT_ranking3["max_predicted_aligned_error"])
     WT_PAE_5 = (WT_ranking4["predicted_aligned_error"], WT_ranking4["max_predicted_aligned_error"])

[13] pae1, max_pae1 = WT_PAE_1
  
```



3 Local server에서 AlphaFold 사용 방법

Get information from Pkl file – draw pLDDT & PAE



3 Local server에서 AlphaFold 사용 방법

Compare other predictions - local

```
[ ] with open('/content/WT/ranking_debug.json', 'r') as rank:
    WT_rank = json.load(rank)
    for i in range(0,5):
        rank = WT_rank['order'][i]
        with open('/content/WT/result_'+rank+'.pkl', 'rb') as f:
            globals()['WT_ranking'+str(i)] = pickle.load(f)

[ ] with open('/content/WT/ranking_debug.json', 'r') as rank:
    MT_rank = json.load(rank)
    for i in range(0,5):
        rank = MT_rank['order'][i]
        with open('/content/WT/result_'+rank+'.pkl', 'rb') as f:
            globals()['MT_ranking'+str(i)] = pickle.load(f)

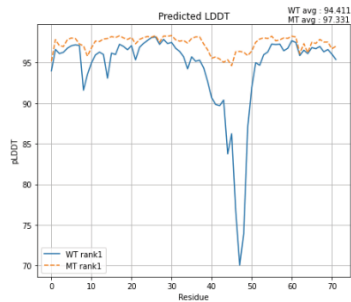
[ ] WT_plddt_1 = WT_ranking0['plddt']
    WT_plddt_2 = WT_ranking1['plddt']
    WT_plddt_3 = WT_ranking2['plddt']
    WT_plddt_4 = WT_ranking3['plddt']
    WT_plddt_5 = WT_ranking4['plddt']
    WT_PAE_1 = (WT_ranking0['predicted_aligned_error'], WT_ranking0['max_predicted_aligned_error'])
    WT_PAE_2 = (WT_ranking1['predicted_aligned_error'], WT_ranking1['max_predicted_aligned_error'])
    WT_PAE_3 = (WT_ranking2['predicted_aligned_error'], WT_ranking2['max_predicted_aligned_error'])
    WT_PAE_4 = (WT_ranking3['predicted_aligned_error'], WT_ranking3['max_predicted_aligned_error'])
    WT_PAE_5 = (WT_ranking4['predicted_aligned_error'], WT_ranking4['max_predicted_aligned_error'])
    WT_avg_plddt_1 = WT_ranking0['ranking_confidence']
    WT_avg_plddt_2 = WT_ranking1['ranking_confidence']
    WT_avg_plddt_3 = WT_ranking2['ranking_confidence']
    WT_avg_plddt_4 = WT_ranking3['ranking_confidence']
    WT_avg_plddt_5 = WT_ranking4['ranking_confidence']

[ ] MT_plddt_1 = MT_ranking0['plddt']
    MT_plddt_2 = MT_ranking1['plddt']
    MT_plddt_3 = MT_ranking2['plddt']
    MT_plddt_4 = MT_ranking3['plddt']
    MT_plddt_5 = MT_ranking4['plddt']
    MT_PAE_1 = (MT_ranking0['predicted_aligned_error'], MT_ranking0['max_predicted_aligned_error'])
    MT_PAE_2 = (MT_ranking1['predicted_aligned_error'], MT_ranking1['max_predicted_aligned_error'])
    MT_PAE_3 = (MT_ranking2['predicted_aligned_error'], MT_ranking2['max_predicted_aligned_error'])
    MT_PAE_4 = (MT_ranking3['predicted_aligned_error'], MT_ranking3['max_predicted_aligned_error'])
    MT_PAE_5 = (MT_ranking4['predicted_aligned_error'], MT_ranking4['max_predicted_aligned_error'])
    MT_avg_plddt_1 = MT_ranking0['ranking_confidence']
    MT_avg_plddt_2 = MT_ranking1['ranking_confidence']
    MT_avg_plddt_3 = MT_ranking2['ranking_confidence']
    MT_avg_plddt_4 = MT_ranking3['ranking_confidence']
    MT_avg_plddt_5 = MT_ranking4['ranking_confidence']
```

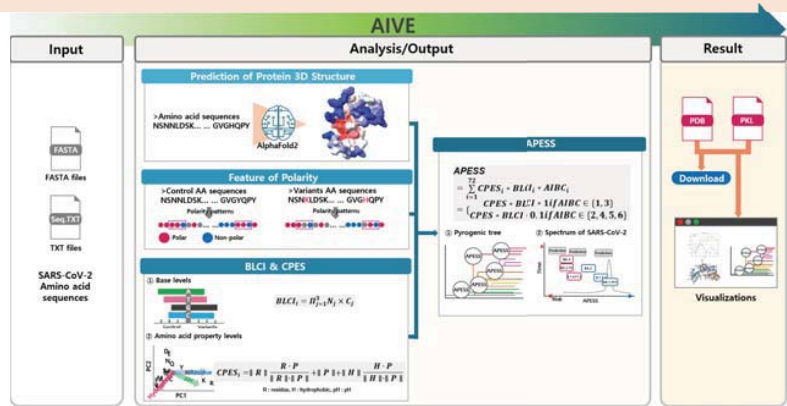
3 Local server에서 AlphaFold 사용 방법

Compare other predictions - local

```
[ ] plt.figure(figsize=(16, 6))
    plt.subplot(1, 2, 1)
    plt.plot(WT_plddt_1, linestyle='solid', label='WT rank1')
    plt.plot(MT_plddt_1, linestyle='dashed', label='MT rank1')
    plt.legend()
    plt.text(55, 100, 'WT avg : ' + '%0.3f' % WT_avg_plddt_1 + '\nMT avg : ' + '%0.3f' % MT_avg_plddt_1)
    plt.grid(True)
    plt.title('Predicted LDDT')
    plt.xlabel('Residue')
    plt.ylabel('pLDDT')
    plt.savefig('/content/11.png', dpi=300)
```



4 AIVE 사용 설명



The information and analysis tools we provide are the following:

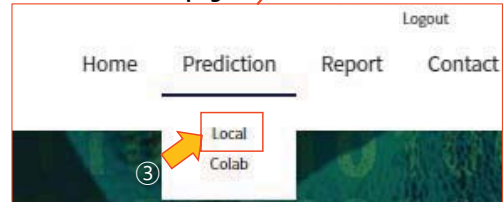
- A. Protein structure prediction from viral sequences using learning models
Prediction of folding and docking from viral mutations
Comparison of folding and docking scores
- B. Polarity changes in protein sequences
Measurement of repeated polarity changes
- C. Mathematical models based on base and amino acid property levels (BLCI & CPES)
Scoring of rate of change in base levels
Scoring of rate of change amino acid property(residue, hydrophobicity, and pH levels)
- D. Comprehensive mathematical analysis model: Protein structure prediction and base/amino acid property levels (APESS)



4 AIVE 사용 설명



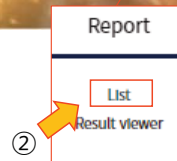
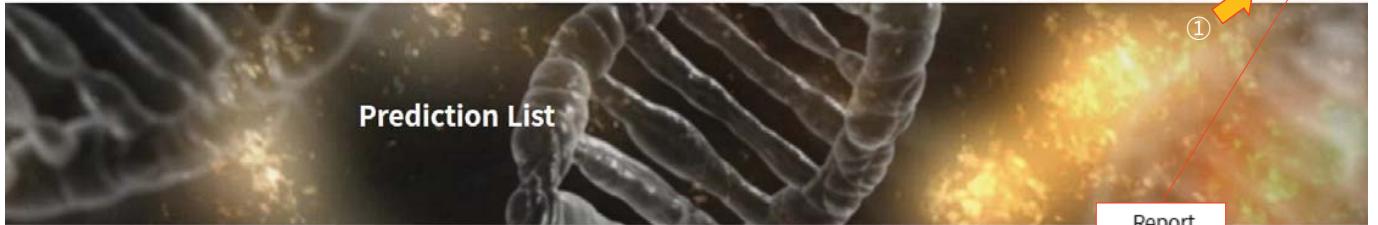
Local Prediction page



4 AIVE 사용 설명



Home Prediction Report Contact



4 AIVE 사용 설명



Sign Up Login Home Prediction Report Contact



Job List

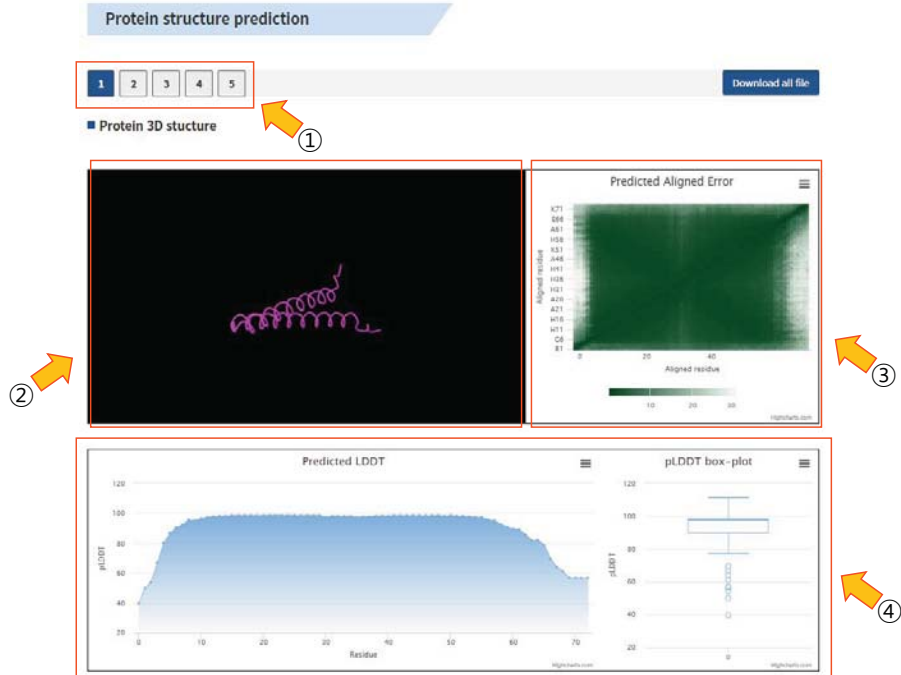
No	Project name	Prediction	Target virus	Status	Result link
1	None	Multimor	전체	Complete	Result link
2	Sars-Cov-2 Test	Monomor	SARS-CoV...	Complete	Result link

Job List

No	Project name	Prediction	Target virus	Status	Result link
1	None	Multimor	전체	Complete	Result link
2	Sars-Cov-2 Test	Monomor	SARS-CoV...	Complete	Result link
3	Test 2	Monomor	SARS-CoV...	Complete	Result link
4	Test 3	Monomor	SARS-CoV...	Complete	Result link
5	apess 오류 테스트	Monomor	SARS-CoV...	Complete	Result link
6	test	Monomor	전체	Complete	Result link
7	test2	Multimor	전체	Complete	Result link
8	test3	Monomor	SARS-CoV...	Complete	Result link
9	test1	Monomor	SARS-CoV...	Complete	Result link
10	sssss	Monomor	SARS-CoV...	Complete	Result link

<< < 1 2 > >>

4 AIVE 사용 설명 : 결과 설명



4 AIVE 사용 설명 : 결과 설명



4 AIVE 사용 설명 : 결과 설명

■ Virus amino acid info
- Wild type virus sequencing

```

R K A H K G A E H H H K A A E H H E Q A A K H H H A A A E H H E K G E H E Q A A H H A D T A Y A
B B N B B N N A B B B B N N A B B A P N N B B B N N N A B B A B N A B A P N N B B N A P N P N
E E E E S H O E E E E H H O E E O P H H E E E E H H H O E E O E S O E O P H H E E H O P H H H
    
```

① ② ③

④

```

H H K H A E E H A A Q A A K H D A E H H A P K P D
B B B B N A A B N N P N N B B A N A B B N N B N A
E E E E H O O E H H P H H E E O H O E E H S E S O
    
```

- Mutated type virus sequencing

```

R K A H K G A E H H H K A A E H H E Q A A K H H H A A A E H H E K G E H E Q A A H H A D T A Y A
B B N B B N N A B B B B N N A B B A P N N B B B N N N A B B A B N A B A P N N B B N A P N P N
E E H E E S H O E E E E H H O E E O P H H E E E E H H H O E E O E S O E O P H H E E H O P H H H
    
```

⑤

```

H H K H A E E H A A Q A A K H D A E H H A P K P W
B B B B N A A B N N P N N B B A N A B B N N B N N
E E E E H O O E H H P H H E E O H O E E H S E S H
    
```

	Polarity feature		Amino acid	
	Polarity features	Non Polar	■	Ala (A), Val (V), Leu (L), Gly (G), Ile (I), Met (M), Trp (W), Phe (F), Pro (P)
Polar		■	Ser (S), Cys (C), Asn (N), Gln (Q), Thr (T), Tyr (Y)	
Acidic		■	Asp (D), Glu (E)	
Basic		■	Lys (K), Arg (R), His (H)	
Five amino acid properties	Amino acids with electrically charged side chain _negative	1	Lys (K), Arg (R), His (H)	
	Amino acids with electrically charged side chain _positive	2	Asp (D), Glu (E)	
	Amino acids with Polar uncharged side chain	3	Ser (S), Asn (N), Gln (Q), Thr (T)	
	Special cases	4	Cys (C), Gly (G), Pro (P)	
	Amino acids with hydrophobic side chain	5	Ala (A), Val (V), Leu (L), Ile (I), Met (M), Trp (W), Phe (F)	

⑥



QUIZ

- <http://honglab.catholic.ac.kr> 의 NEWS 링크를 참조하세요. (실습 데이터, 향후 솔루션을 제공합니다)

HONG LAB

OF CATHOLIC UNIVERSITY OF KOREA, COLLEGE OF MEDICINE

We are **DATA SCIENTIST**!!

We are studying Cancer Multi-Omics Data Analysis and Artificial Intelligence Research.

You Can **Come With Us**.

LECTURE

- [2021-11-05] [090053] 핵심의과학실험
- [2021-09-11] [과3학년] 정밀의학
- [2021-09-02] [090072] (성밀외료빅데이터) 생명정보학...
- [2021-09-02] [BMED173] 의료데이터베이스시스템
- [2021-07-23] [BMED217] 의료빅데이터

NEWS

- [2022-02-16] 제8회 KSBI-BIML 2022 생물정보학...
- [2022-02-09] [THE 16TH KUGU WINTER SY...
- [2022-01-21] 제 14회 대한암학회 동계 워크숍 발표
- [2021-12-17] [질병청 연구 세미나]
- [2021-10-27] '21 ANNUAL CONFERENCE OF...

Q1.

- Mutalisk를 수행 후 확인할 수 있는 결과입니다. 돌연변이가 한 곳에 집중적으로 모여 있는 현상을 무엇이라고 합니까?

Q2.

- Mutalisk 웹 사이트 (<http://mutalisk.org>)에서 제공하는 melanoma 환자의 돌연변이 데이터 (sample_melanoma.vcf)를 다운로드 받아서 분석을 진행하고 다음 물음에 답하십시오.
- Version 2와 Version 3을 이용한 시그니처 분석 후 1st, 2nd signature번호 (타입)는 각각 무엇인가?
- Replication timing에서 돌연변이가 가장 많은 구간과 유의미한 돌연변이는 무엇인가? (Version 2, 3 각각에 대해 답하십시오.)

Q3.

- AlphaFoldI에서는 () 방식을 이용하여 구조를 예측하고, AlphaFoldII에서는 () 방식을 예측하게 된다. 이로 인하여 단백질 구조 예측 개수에서 약 1,000배 가까운 성능 향상을 보이게 되었다.

Q4.

- 오미크론 세부 계통 변이 BQ 1.1의 pLDDT와 PAE를 제시하고, Alpha 변이와 어느 정도의 위험도 (전파도) 차이를 보이는지? 보이시오.

Q5.

- BRAF 유전자 V600E 변이에 대해서 Vemurafenib의 표적 치료를 할 수 있다. AlphaFold를 이용하여 변이 유/무에 따른 구조 차이를 보이시오.

경청해 주셔서 감사합니다.

Big Data Analysis Lab. (<http://honglab.catholic.ac.kr>)
교수: 홍동완 (dwhong@catholic.ac.kr, 02-3147-8424)
연구실: 가톨릭 성의교정 옴니버스파크 의과대학 7113호

연구 주제

- Cancer Multi-Omics (WGS, WES, Panel, RNA-Seq, Proteome, Hi-C 등) 분석
- 분석 소프트웨어 및 통합 데이터베이스 개발
- 인공지능 기반 약물반응성, 적정 치료제, 치료법 개발
- 빅데이터 분석 알고리즘 개발
- NFT & Metaverse



I Dear CMC | #2. 유전자? 내 손안에 있소이다! | 홍동완 교수
조회수 577회 · 2주 전

가톨릭대학교 산학협력단

이런 연구를 하고 있습니다!

<https://www.youtube.com/watch?v=H9fFnAwEcyQ>

'구독'과 '좋아요'는 생계형.과학자에게 도움이 됩니다