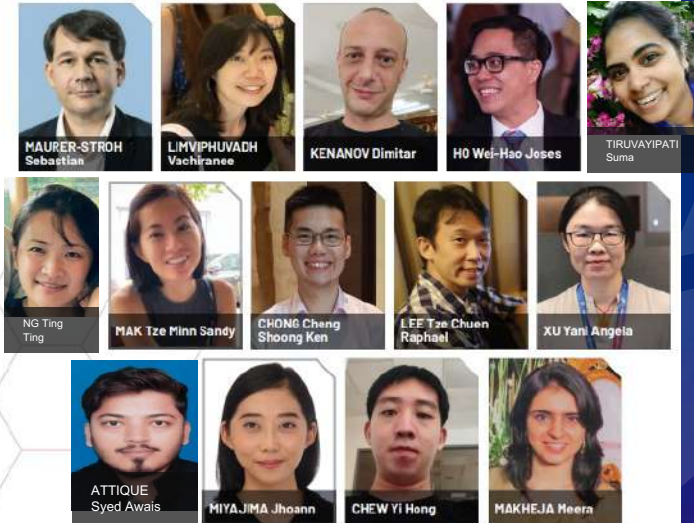


# Bioinformatics Institute (BII)

## Protein Sequence Analysis



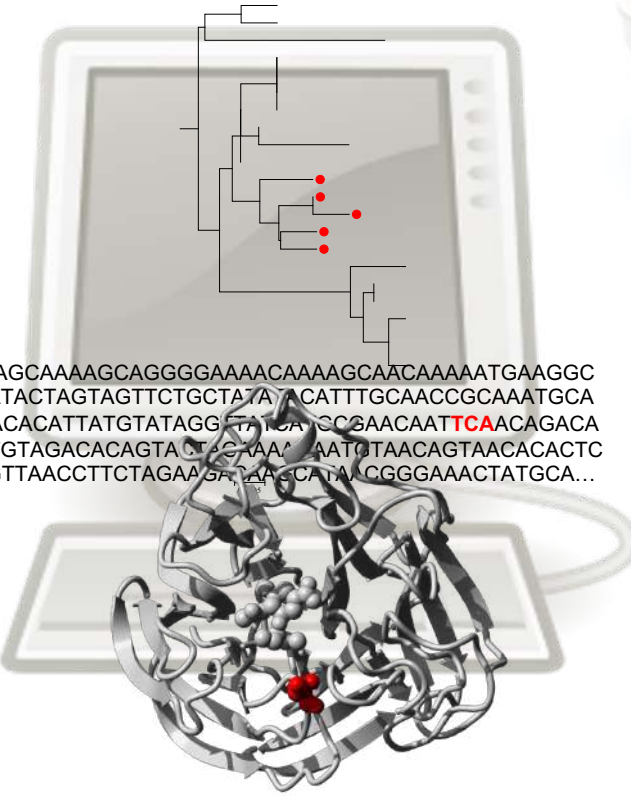
## Viruses



Nature Methods. 2023 Apr;20(4):512-522.  
 Emerg Infect Dis. 2023 Apr;29(4):778-781.  
 Nature Commun. 2022 Nov 16;13(1):7003.  
 JAMA Netw Open. 2022 Aug 1;5(8):e2228900.  
 Clin Infect Dis. 2022 Aug 24;75(1):e1128-e1136.  
 EMBO Mol Med. 2022 Mar 7;14(3):e15227.  
 ... +10 other publications

## Restricted

### Computational Sequence and Structure Analysis



```

...AGCAAAAGCAGGGGAAAACAAAAGCAATCAAAAATGAAGGC
AATACTAGTAGTTCTGCTATAACATTTGCAACCGCAAATGCA
GACACATTATGTATAGGATATCAACCCBAACAATTCAACAGACA
CTGTAGACACAGTAACTGAAATAATGTAACAGTAACACACTC
TGTTAACCTTCTAGAAATTAACATAACGGGAAACTATGCA...
  
```

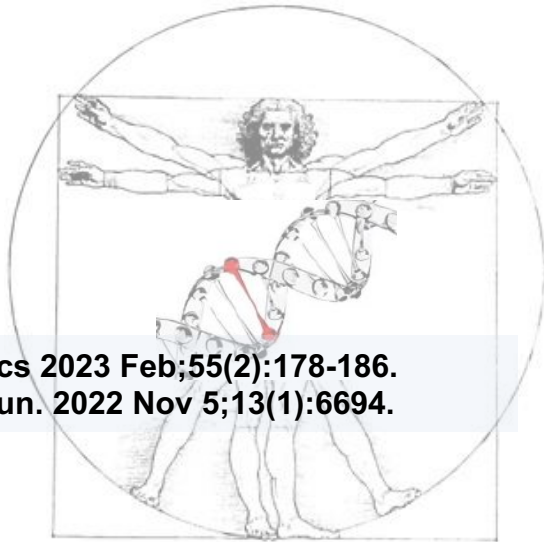
## Food flavors



## Drugs

Patent application: BII/P/13305/00/SG  
 Patent application: ISCE2/P/13296/00/SG

## Human variation



Nature Genetics 2023 Feb;55(2):178-186.  
 Nature Commun. 2022 Nov 5;13(1):6694.

We are working at the  
 interface between  
 sequence and structure

## Product Safety



J Proteomics. 2022 Oct 30;269:104724.  
 Nucleic Acids Res. 2022 Jul 5;50(W1):W36-W43.



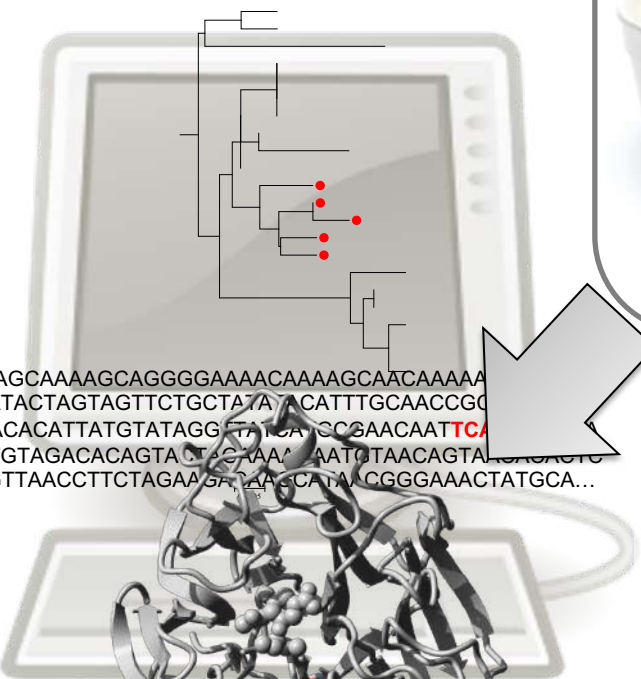


Viruses



Restricted

Computational  
Sequence and  
Structure Analysis



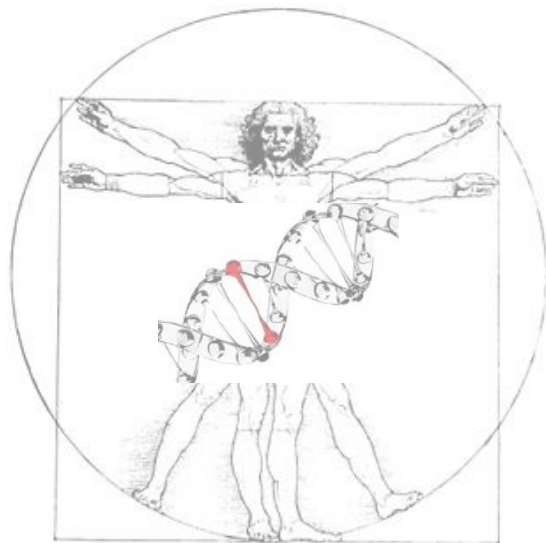
...AGCAAAAGCAGGGGAAAACAAAAGCAATCAAAAA  
AATACTAGTAGTTCTGCTATAACATTGCAACCGC  
GACACATTATGTATAGGATACACCCBAACAATCA  
CTGTAGACACAGTACTGAAATAATCTAACAGTAAAGGATG  
TGTTAACCTTCTAGAAATCAATACGCGGAAACTATGCA...

Food flavors



Drugs

Human variation



Product Safety



Enzymes can be used to naturally produce food flavours and fragrances but also drugs!





# AI (CNN) prediction of mutation effects on enzyme function

## Kinetics

EVB Model

## Binding

Docking scores  
Binding free energy

r=0.9

## Stability

Yasara + FoldX

12 of 18 correct predictions of increased stability (incl. top 3)

PHARMA INNOVATION  
PROGRAMME SINGAPORE (PIPS)



## Aggregation

Tango  
Waltz

1 predicted mutation set increases solubility 2-fold

## Allostery

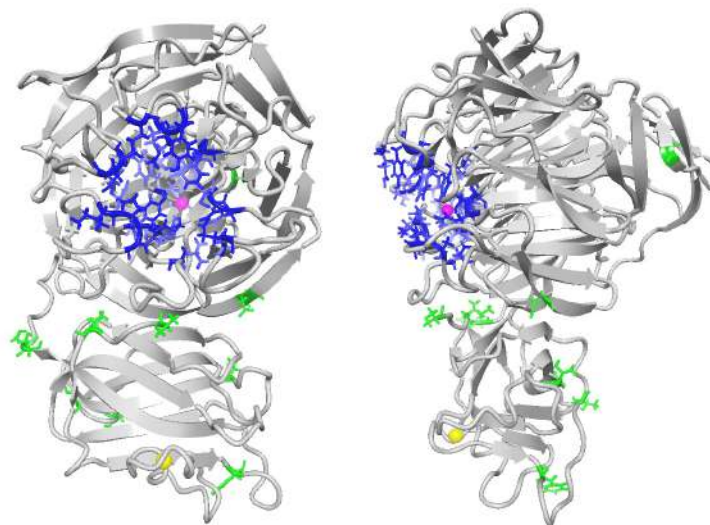
Allosigma

7 of 17 correct predictions of long range binding effect

## Sequence Conservation

SIFT  
Shannon entropy

Jump from 12% good sites to 63% good sites



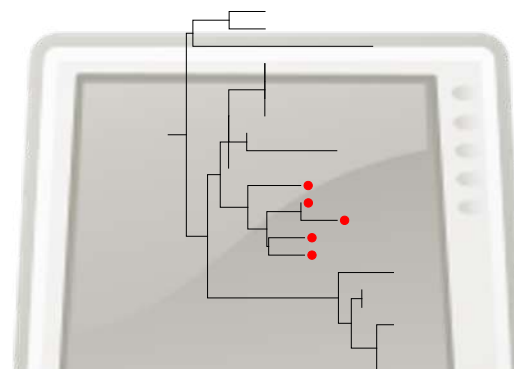


Viruses



Restricted

### Computational Sequence and Structure Analysis

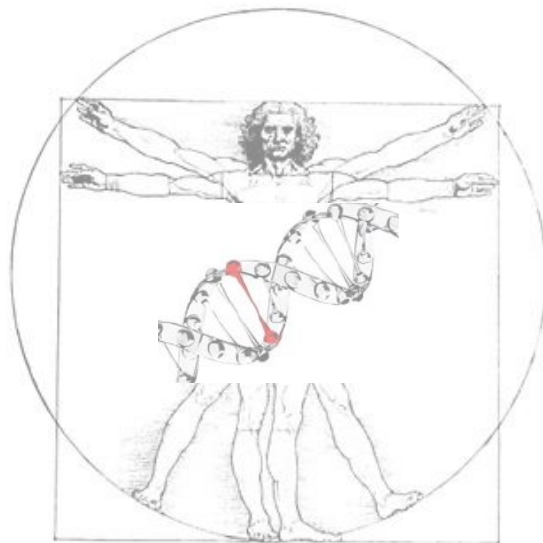


Food flavors

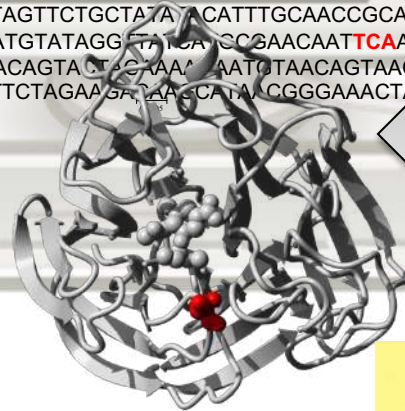


Drugs

Human variation



...AGCAAAAGCAGGGGAAAACAAAAGCAATCAAAAATGAAGGC  
AATACTAGTAGTTCTGCTATAACATTGCAACCGCAAATGCA  
GACACATTATGTATAGGATACACCCBAACAATTCAACAGACA  
CTGTAGACACAGTACTGAAATCAATCTAACAGTAACACCTC  
TGTTAACCTTCTAGAAATCAACATAACGGGAAACTAT



Product safety



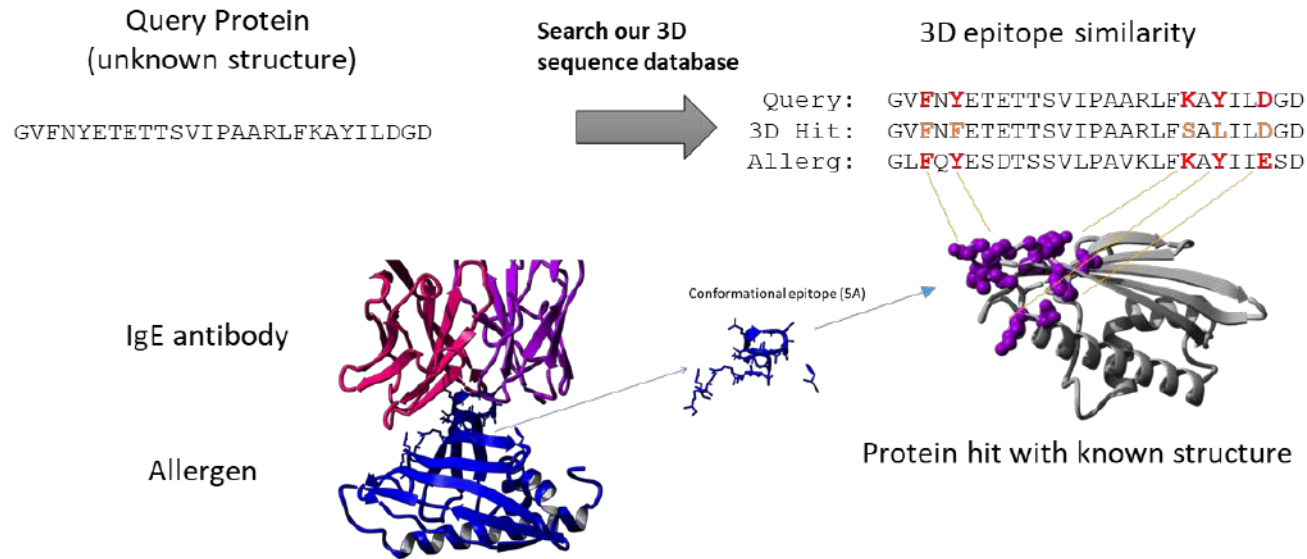
Proteins can cause allergy if similar to other allergens.





Vachi, Minh

# Computational prediction of protein allergenicity potential: AllerCatPro



**AllerCatPro** predicts if a protein is similar to known allergens





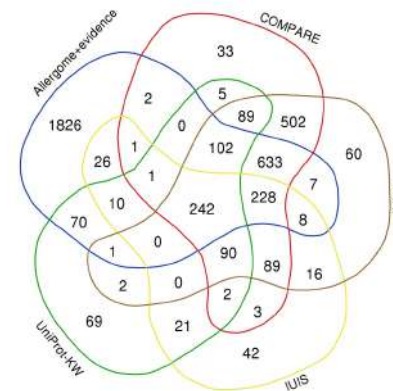
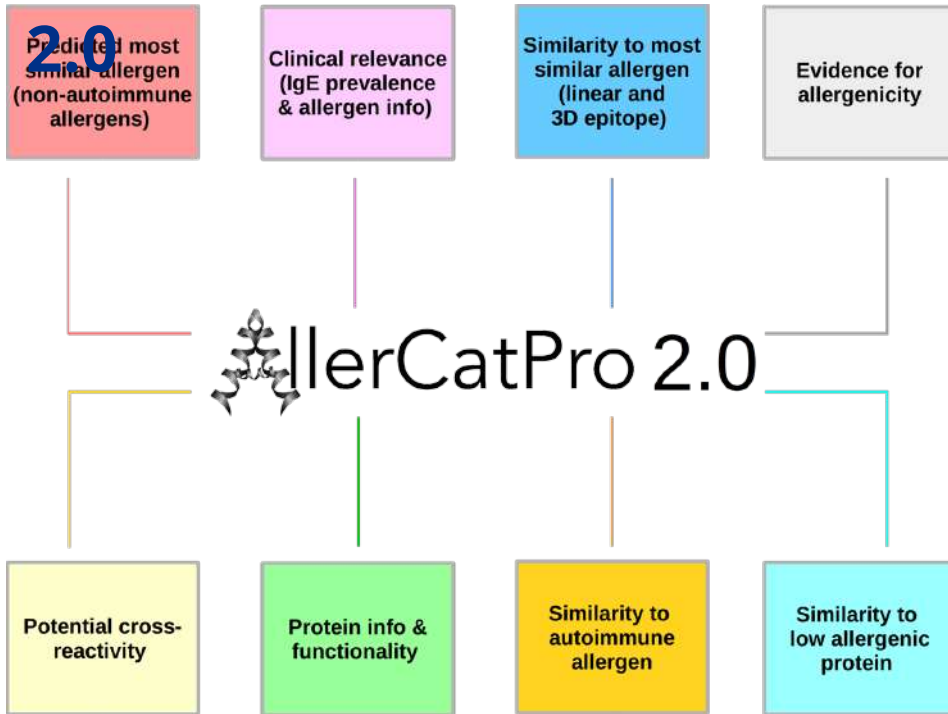


Minh Vachi Andreas

- Using the most comprehensive datasets of reliable proteins associated with allergenicity i.e. **4,979 protein allergens**, **162 low allergenic proteins** and **165 autoimmune allergens**.

## What's new in AllerCatPro

RSC Partnership Agreement on Scientific Services since 28 Feb.2023



**Merged database**  
(WHO/IUIS, COMPARE, FARRP, UniProtKB, Allergome)

Home > Tech Platforms

**Technology Platforms**

- National Shared Platform +
- Tech Access Initiative
- Mass Spectrometry Cluster (A\*STAR) +
- Flow Cytometry Cluster (A\*STAR)
- Animal Research and Testing +
- Bioinformatics +
- Bioimaging +
- Bioprocess Engineering
- Cell and Gene Therapy +
- Drug Discovery +
- Food and Consumer Care +
- Genomics +
- Histopathology +
- Immunomonitoring
- Mechanical Testing +
- Microbiology +
- Molecular Engineering

**Allergenicity and Toxicity Platform by Bioinformatics Institute (BII)**

In silico safety assessment using AllerCatPro 2.0

By providing us protein/nucleotide sequences of interest, we provide detailed results of protein allergenicity potential using AllerCatPro 2.0. Our service can be used as a first step for protein safety assessment to find potential allergens for further evaluation.

**Key Techniques**

- AllerCatPro 2.0 predicts protein allergenicity potential based on the most comprehensive databases of reliable proteins associated with allergenicity from the WHO/IUIS, COMPARE, FARRP, UniProtKB and Allergome.
- AllerCatPro 2.0 combines the similarity of both amino acid sequences and 3D structures for the prediction while other computational methods use only sequence features. The results on our benchmark datasets indicated that 3D structure similarity contributes significantly to the performance of AllerCatPro 2.0.
- We provide an excel file containing AllerCatPro 2.0 result and a summary result table. The result includes potential cross-reactivity based on similarity in sequences and protein 3D structure to known allergens, protein information (UniProt/NCBI), functionality (Pfam, InterPro, SUPFAM), as well as clinical relevance with regards to IgE prevalence and allergen information related to the most similar allergen.
- We also have a consultancy package if customers need more guidance and/or interpretation on results from AllerCatPro 2.0.

<https://allercatpro.bii.a-star.edu.sg>

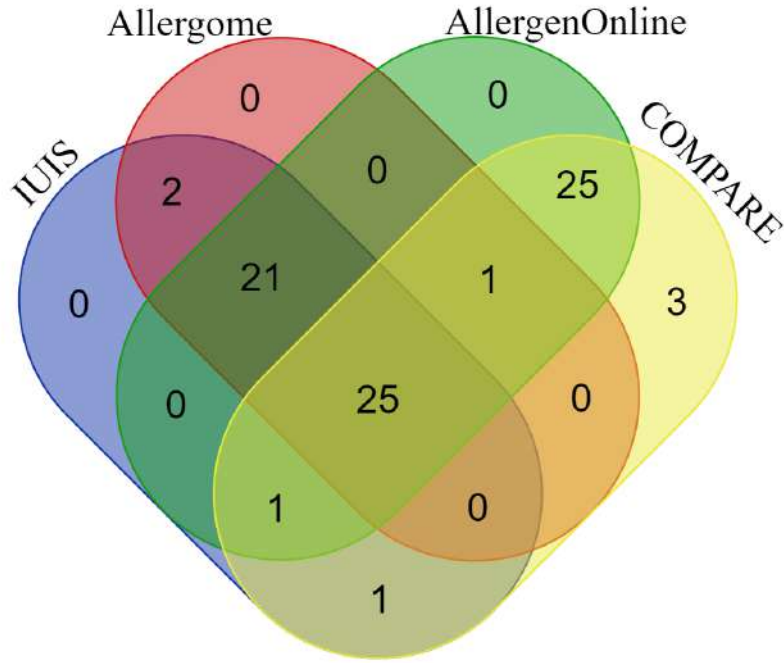
Nguyen MN, Krutz NL, Limviphuvadh V, Lopata AL, Gerberick GF, Maurer-Stroh S. (2022). AllerCatPro 2.0: a web server for predicting protein allergenicity potential. *Nucleic Acids Res.*



Vachi Thimo Andreas

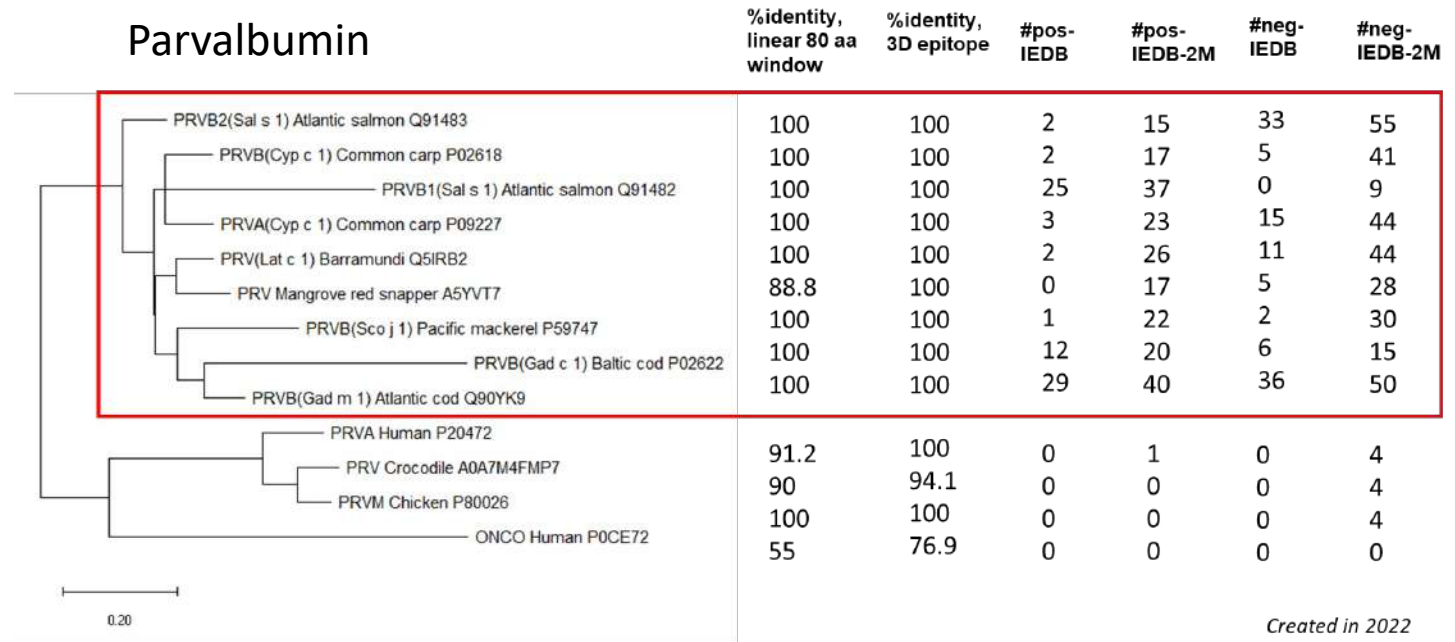
# Understanding allergens in alternative food from urban aquaculture fish to underpin food safety

Two of the most consumed fish in Singapore



Compiled all known fish allergens from the four most common and well-known allergen databases (n=79)

## Parvalbumin



Created in 2022

- In the future, %relative intensity of IgE binding of parvalbumin of 12 species of fish can be added to phylogenetic tree in order to predict cross-reactivity.
- Family-specific threshold optimization will be implemented to improve AllerCatPro allergenicity assessment and predict cross-reactivity of putative proteins.

The National Research Foundation, Singapore and A\*STAR under the Singapore Food Story R&D Programme [W22W3D0003]





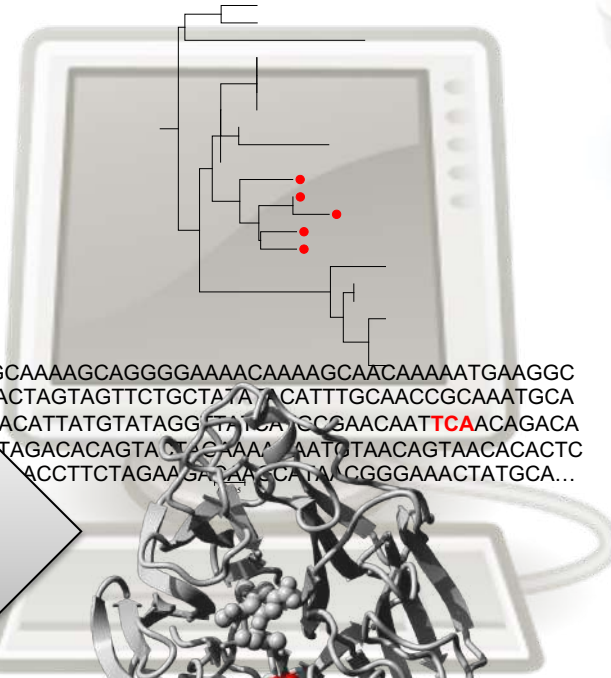


Viruses



Restricted

### Computational Sequence and Structure Analysis



Food flavors



Human variation

gsk  
MSD  
Pfizer

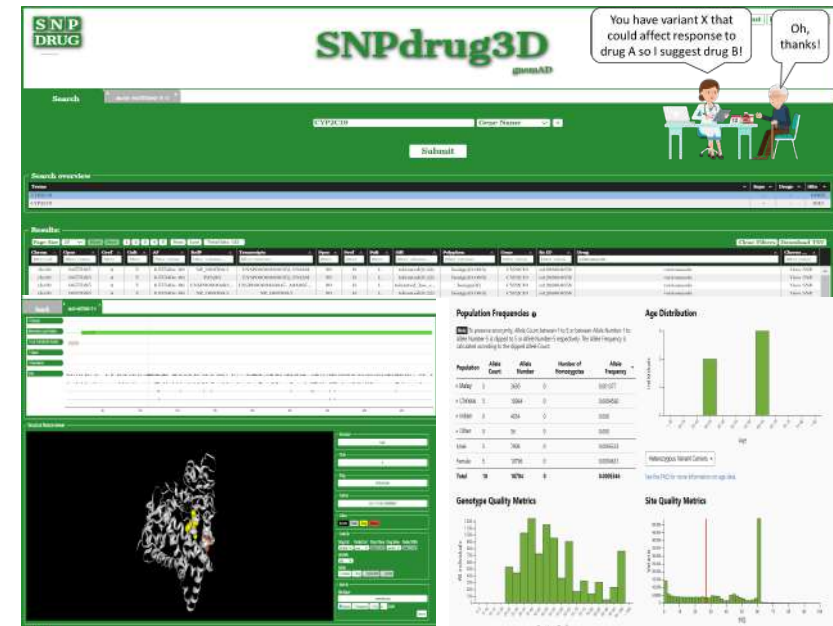
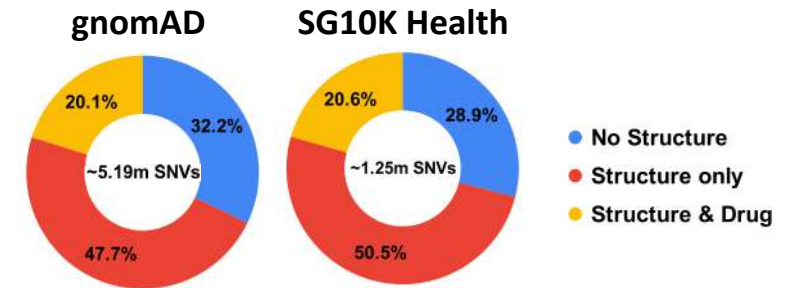
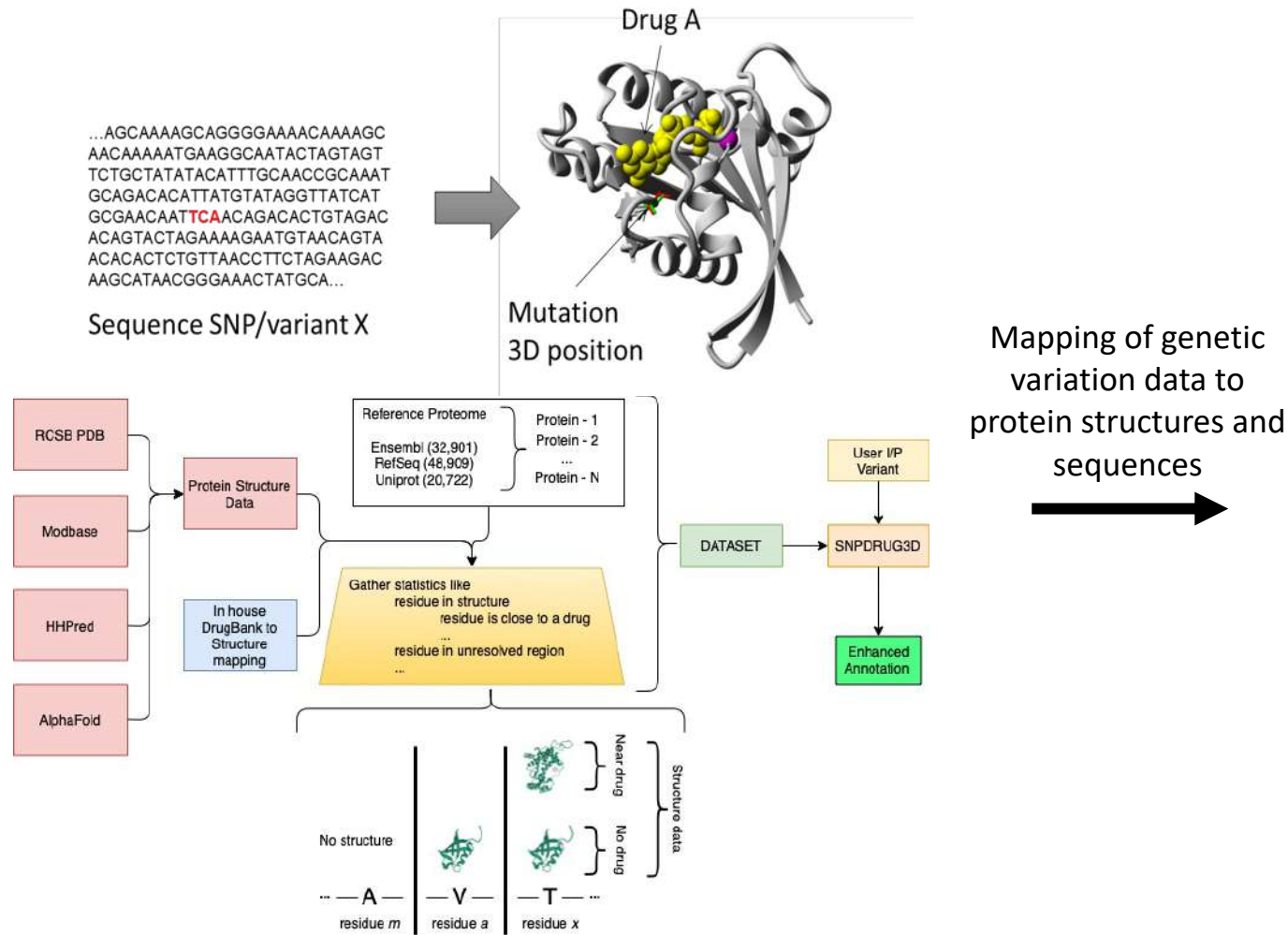
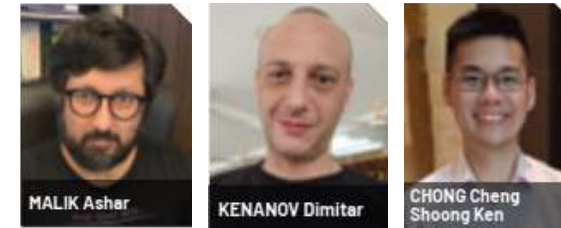
GIS

Individual mutations can cause disease or affect drug efficacy

Product Safety



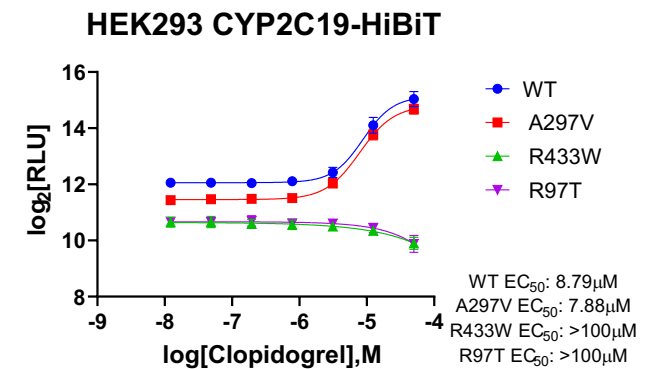
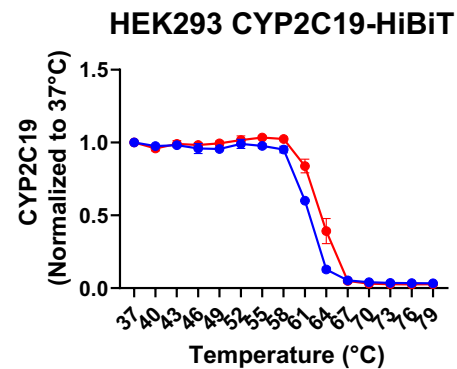
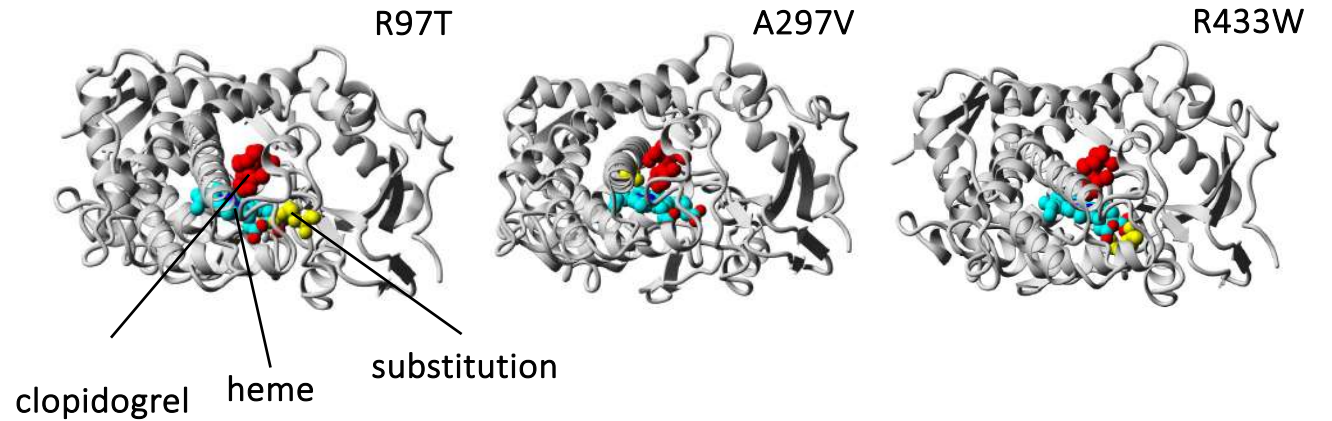
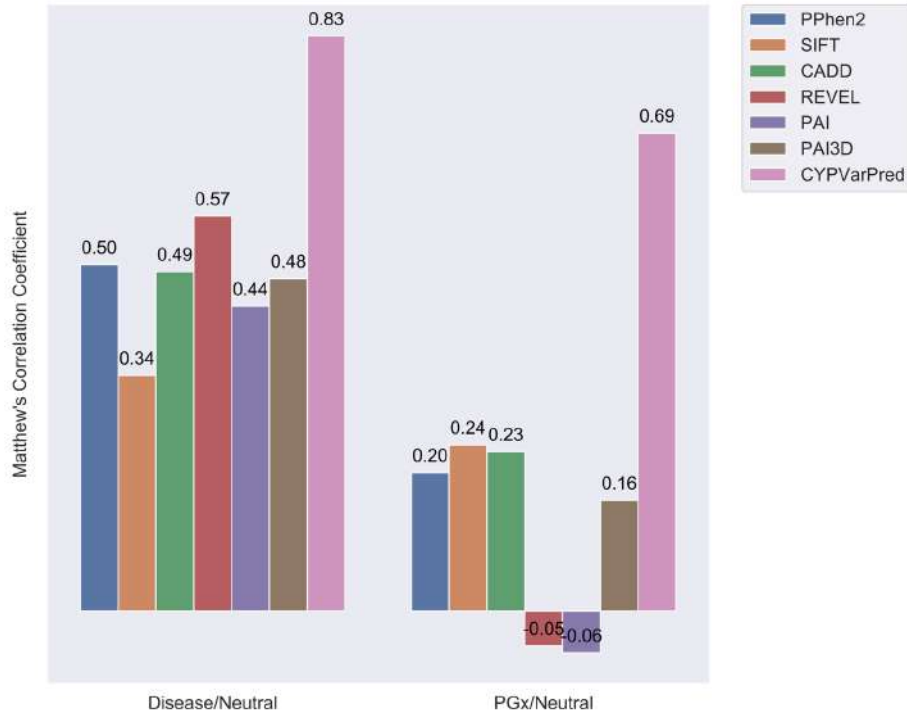
# SNPdrug3D – placing genetic variants in a protein structural context



5.8 million SNVs in 20,442 genes from SG10K Health and gnomAD populations mapped to 202,299 protein structures (experimental or AI-generated) containing 5962 drugs

# SNPdrug3D – applications

## Drug metabolizing enzyme (CYP2C19)



1) Derive features using SNPdrug3D data to build PGx-inclusive variant pathogenicity predictors. Predictor ('CYPVarPred') outperformed all other tested predictors in prediction of PGx CYP variants.

2) Identify variants in protein targets that may affect protein-drug binding. Here, R433W and R97T (novel) but not A297V disrupted CYP2C19-clopidogrel binding according to cellular thermal shift assay results.



# SNPdrug3D – clinical use case

"Several patients I treated with **clopidogrel** did not respond to it, are there any **common mutations** in the population that could affect drug response hence allowing me to **stratify** who should get this drug?"

## Search

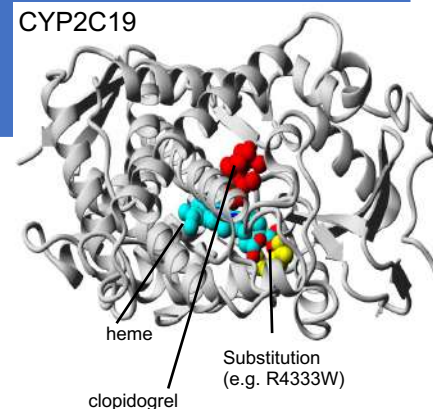
- Searching for “clopidogrel” shows multiple CYP genes with mutations near the drug
- Filtering for the common drug metabolizing gene CYP2C19 lists possible relevant mutations

## SNPdrug3D

Pos	Ref	Alt	Sift	Polyphen	Gene	Rs ID
433	B	W	deleterious	possibly_damaging	CYP2C19	rs56
112	I	S	deleterious	probably_damaging	CYP2C19	
297	A	V	deleterious	benign	CYP2C19	rs1399

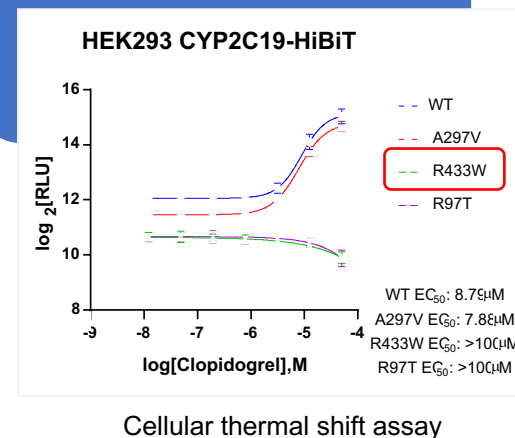
## Identify

- Sorting by Allele Frequency shows R433W as top variant
- R433W is predicted to impact CYP2C19-clopidogrel binding. More common in the Malay population (AF = 0.2%)



## Validate

- Search if reported in literature
- R433W impaired binding of clopidogrel to CYP2C19 directly.
- Other functional assays can also be performed



## Clinical implications

- Stratify patients  
Patients with one normal copy and defective copy of the defective CYP2C19 allele (i.e. with R433W mutation) are intermediate metabolizers (IM), producing less active metabolites of the drug.
- Dose adjustments  
Avoid standard dose of 75mg, use alternative antiplatelet agents (e.g. prasugrel) if possible<sup>1</sup>.

<sup>1</sup>CPIC guidelines. Lee et al., 2022. Clin Pharmacol Ther.



GIS



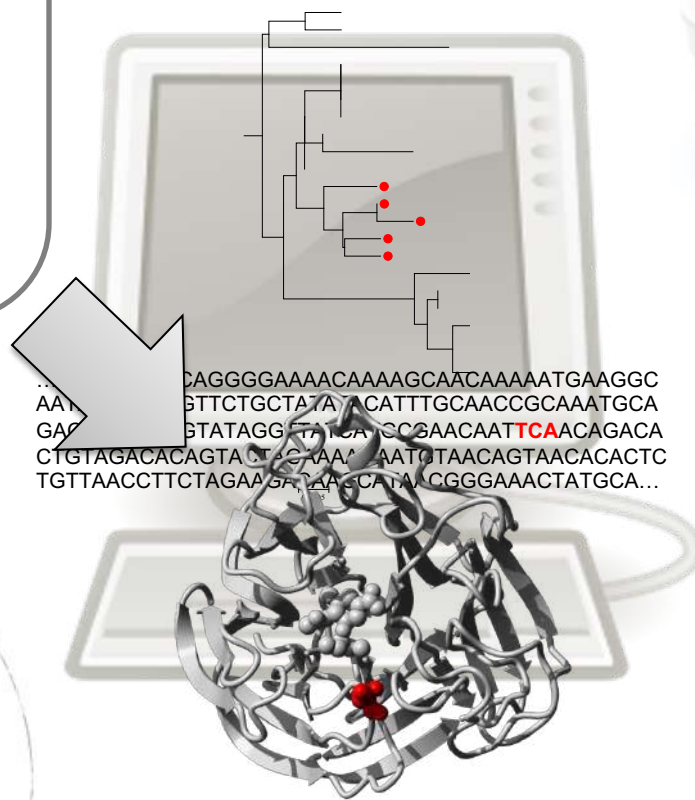
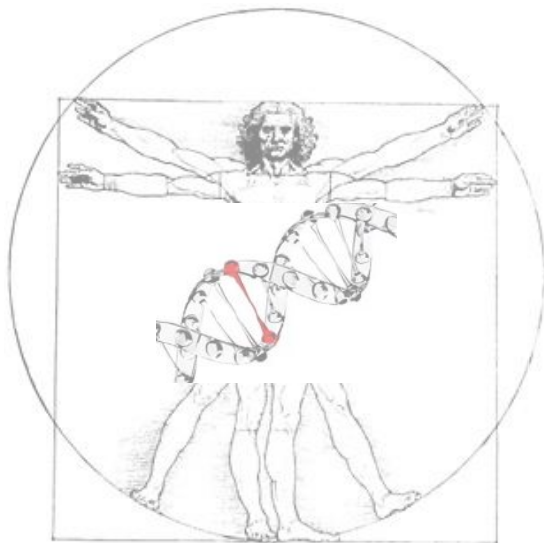
Restricted

Computational  
Sequence and  
Structure Analysis

Food flavors



Human variation



Product Safety

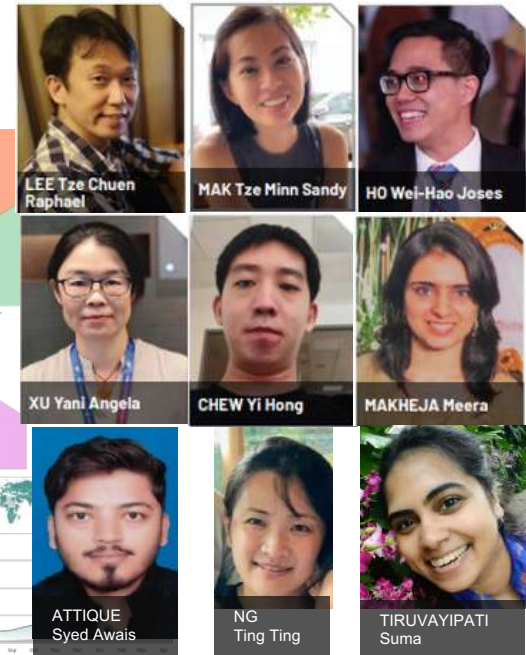


Long history of flu  
work but COVID put  
us into the spotlight

# Example COVID-19: data + analysis = impact

One A\*STAR: BII works with ID Labs, SlgN, GIS, EDDC, DxDhub, IHPC, I2R

3. Comparison with other strains' sequences to trace global and local transmission



Hospital, GP etc.

Lab

Bioinformatics

Sample

Sequence

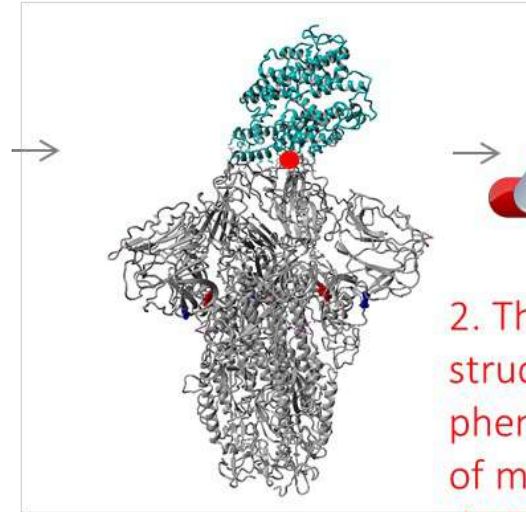
Interpretation

Assembly

```

...AGCAAAAGCAGGGGAAAACAAAAGCAACAAAAATGAAGGCAACTAGTAG
TTCTGCTATATACATTTGCAACCGCAAATGCAGACACATTATGTATAGGTTATCA
TGCGAACAAATTCACAGACACTGTAGACACAGTACTAGAAAAGAATGTAACAG
TAACACACTCTGTAACTTCTAGAAGACAAGCATAACGGGAAACTATGCA...
    
```

1. Timely sharing of sequences of new viruses is critical and fair sharing mechanisms exist via the GISAID platform



2. Theoretical 3D structure and phenotype effects of mutations and drug/vaccine candidates

WHO's Chief Scientist ([Swaminathan](#) *Nature* 2020) recognized GISAID as "game changer."

MIT Tech review breakthrough technology 2022



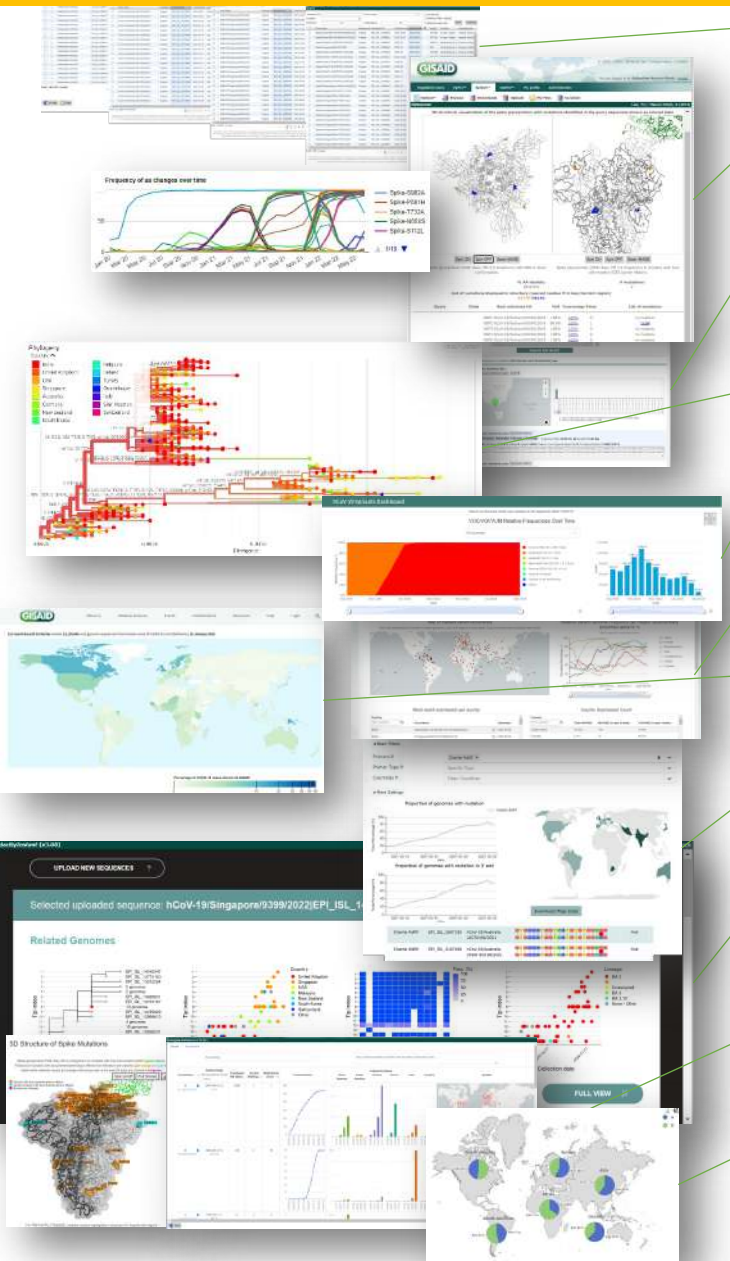
Mentioned by PM during National Day Rally, Minister in Parliament etc.

BII/GIS has global (GISAID, WHO, CEPI) and national role (NCID, MOH) in points 1-3





Current status of key tools being expanded to more pathogens



Tool	Purpose	EpiFlu	EpiCoV	EpiRSV	EpiPox
Browse, Search, Download	Data accessibility	Yes	Yes	Yes	Yes
Flu/CoV/RSV/Pox-Server	Interpret mutation effects	Yes	Yes	Yes	Yes
BLAST	Genomic Epidemiology - Detail	Yes	Yes	Yes	No
Phylodynamics	Genomic Epidemiology - overview	Yes	Yes	Yes	Yes
Variant Frequency	Major variant tracking	Yes	Yes	Yes	Yes
Variant Tracker	Major variant tracking	Yes	Yes	Yes	Yes
Submission Tracker	Surveillance capacity monitoring	Yes	Yes	Yes	Yes
PrimerChecker	Surveillance capacity monitoring	Yes	Yes	Yes	Near Future
AudacityInstant	Genomic Epidemiology - Detail	Yes	Yes	Near Future	Near Future
EmergingVariants	Emerging variant tracking	Yes	Yes	Yes	Yes
EpiCharts (beta)	Graphical overview of user selected data	Yes	Yes	Yes	Yes

September 2022

Yes ... new in Nov 22    Yes ... new in Feb 23    Near Future ... later in 2023

# A scale-free view of virus evolution

Subtype



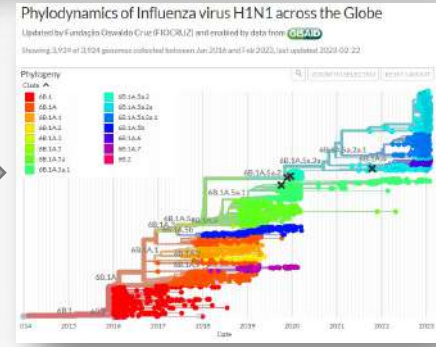
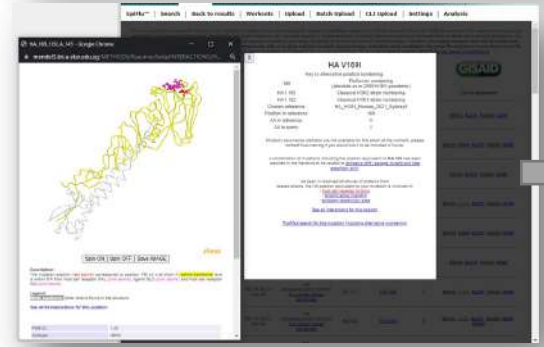
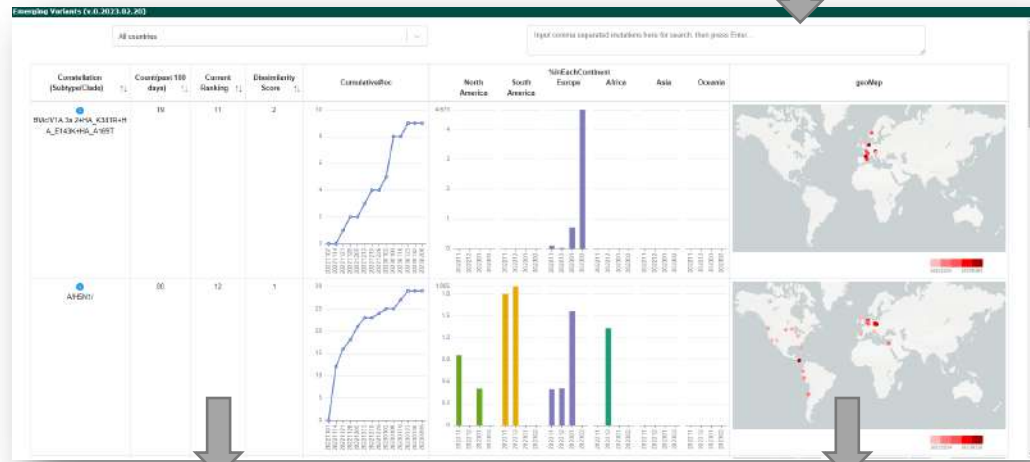
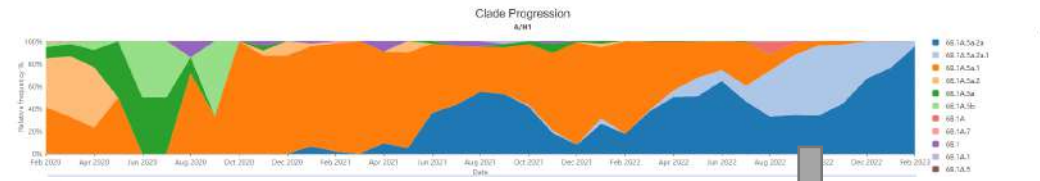
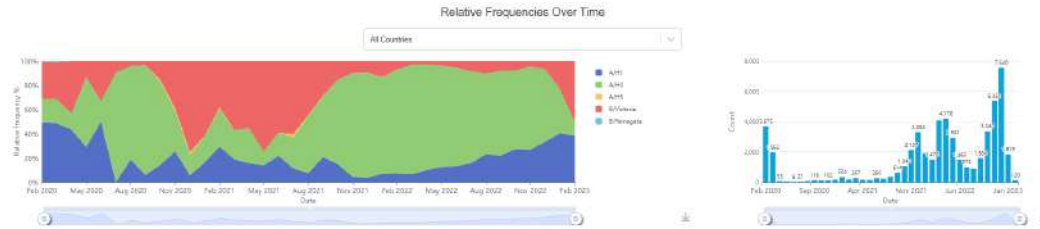
Clade



Emerging Variant  
(unique set of mutations)



Individual mutation



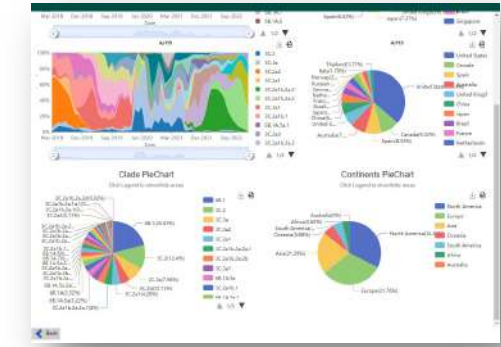
# Connected levels in GISAID's integrated tool ecosystem



Single entry detail

ID	Subtype	Country	Date	Accession	NCBI	EMBL	GenBank	ENA	NCBI	EMBL	GenBank	ENA
A/Spain/060503	H1N1	Spain	2009-06-25	EU021453	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021454	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021455	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021456	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021457	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021458	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021459	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021460	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021461	U001	U001	U001	U001	U001	U001	U001	U001
A/Spain/060503	H1N1	Spain	2009-06-25	EU021462	U001	U001	U001	U001	U001	U001	U001	U001

Set of entries in the database



Set of entries graphical summary

# FluClusterAI: Landing page

New, in preparation!



## FLU CLUSTER

Flu Cluster selection is vital for flu surveillance. A representative sample of the population is chosen based on factors like age, gender, and location to track the virus spread. This helps health authorities gain an accurate understanding of the virus and take steps to reduce its spread. Cluster selection plays a key role in flu prevention.

[Learn more](#)

### Upload input file

At least one sequence file is required

**Sequence file**  
Mandatory to generate the growth chart

sequences-sample-all.fa

**Metadata file** (optional)  
Upload Metadata, to view Enrichment Analysis

metadata-sample-all.tsv

**Phylogenetic Tree file** (optional)  
To generate your own phylogenetic tree.

Add file

Generate report

fasta file fasta fa

	A	C	D	E	F	G
1	sampleID	age	vaccine status	patient status	drug foldchange	HI titer
2	sample1	45		ICU	0.2	256
3	sample2	38		ICU	0.7	128
4	sample3	30			1.3	256
5	sample4	27		ICU	1.5	32
6	sample5	57		ICU		128
7	sample6	97			1.8	
8	sample7	57		ICU	0.9	256
9	sample8	81		ICU	1.7	256
10	sample9	81		ICU	1.5	128
11	sample10	22				64
12	sample11	17		ICU	1.3	32
13	sample12	30		ICU	1.9	
14	sample13	75	vaccine breakthrough		1.9	256
15	sample14				0.9	256
16	sample15	26	vaccine breakthrough			64
17	sample16	87	vaccine breakthrough		1.7	128
18	sample17	83	vaccine breakthrough		0.9	64
19	sample18	1	vaccine breakthrough		1.4	
20	sample19	90			0.1	32
21	sample20	68	vaccine breakthrough			128
22	sample21	46	vaccine breakthrough		0.9	32
23	sample22	17	vaccine breakthrough		1.8	128
24	sample23	46	vaccine breakthrough		1	32
25	sample24	0			0.5	
26	sample25	24	vaccine breakthrough			128
27	sample26	29	vaccine breakthrough		0.9	128
28	sample27	40	vaccine breakthrough		0	32
29	sample28				0.1	64
30	sample29	38			0.2	128
31	sample30	13				
32	sample31	29			0.6	128
33	sample32	48			1.1	32
34	sample33					256
35	sample34	37			0.1	128
36	sample35	46				32
37	sample36	46			0.6	
38	sample37	41			0.6	32
39	sample38	35			1	128
40	sample39	34			1.9	64
41	sample40	65			1.8	128
42	sample41	33			0.6	64

metadata file (tsv/csv/json)

(Sample dataset: NPHL Singapore dataset H1N1pdm09i 2016-2017)



# FluClusterAI: Meta-Enrichment Analysis

Descriptive paragraphs are automatically generated and under representation of metadata.



❖ Odds Ratio of FluCluster against all matched constellations is used to identify enrichment of specific phenotype with certain FluCluster based on user supplied metadata.

## Question to confirm:

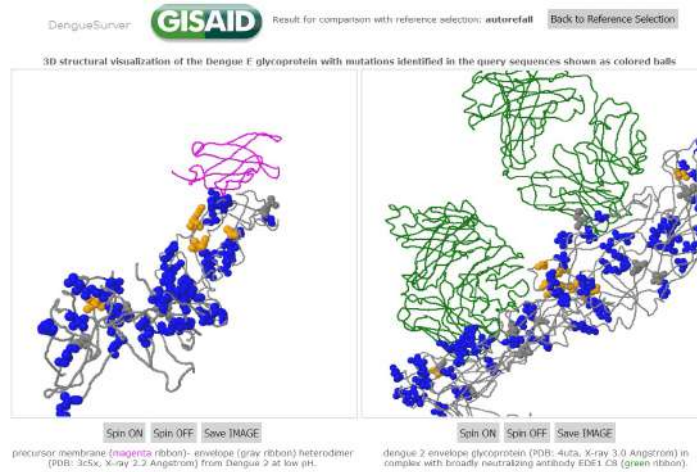
- Currently, missing metadata are left uncolored (white). Do let us know if you prefer us to color them as green (insignificant).

New, in preparation!

# DengueSurver & ChikSurver: Tools to judge relevance of mutations that can affect viral fitness

DengueSurver  
ChikSurver

New, in preparation!



<https://mendel3.bii.a-star.edu.sg/METHODS/denguesurver/current/>  
<https://mendel3.bii.a-star.edu.sg/METHODS/chiksurver/current/>

## Reference Genomes

	DengueSurver	ChikSurver
# ref genomes	4	1
# genotype ref genomes	45	3

**E V365I**

Key to alternative position numbering:  
E 365 V DENV1 numbering  
E 365 V DENV2 numbering  
E 363 V DENV3 numbering  
E 365 T DENV4 numbering

Chosen reference: E DENV2/Thailand/16681/1984  
Position in reference: 365  
AA in reference: V  
AA in query: I

Known effect(s) of series of mutations including position equivalent to your mutation:  
Protein: E  
Coronavirus type: human DENV1 (1993)  
Mutation (as in paper): V365I from series M196V, V365I, T405I  
neutral AA: M, V  
neg. eff. AA: V,I  
Effect: Virulence

**Comment:**  
amino acid changes at positions E196 and E405 could facilitate the oligomeric assembly of DEN-1 envelope glycoproteins in mouse neuroblastoma cells.  
[Literature reference](#)  
(Mutation V365I from series M196V, V365I, T405I in the paper is at an equivalent position of the mutation in your query)

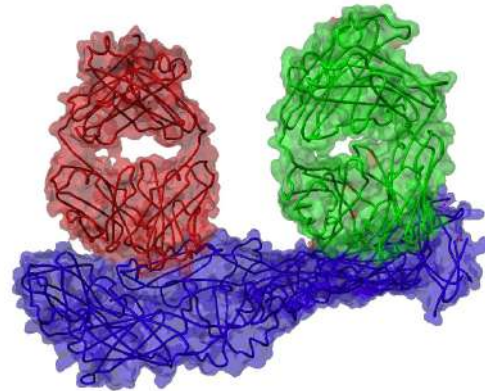
Mutation E V365I already occurred 19 times (0.41% of all samples with E sequence) in 10 countries. The first strain with this mutation, collected in August 1965, was hDenV2/Nigeria/CBEID-112345/1966. The mutation most recently occurred in strain hDenV2/Cambodia/NIH-109-0294/2020, collected in June 2020. [\(see map\)](#)  
[See detailed global statistics for this position](#)

A combination of mutations including the position equivalent to E 365 has been reported in the literature to be related to [Virulence](#).  
[PubMed search for this mutation](#)

## Literature

	DengueSurver	ChikSurver
Drug resistance	2	0
Virulence	14	9
Antigenic drift / escape mutant	7	0
Host specificity change/shift	0	79
Other (enzyme activity, affects protein accumulation/stability/function)	38	18
<b>total literature entries</b>	<b>61</b>	<b>96</b>

## 3D Structure Interactions



	DengueSurver	ChikSurver
# PDB structures	197	?



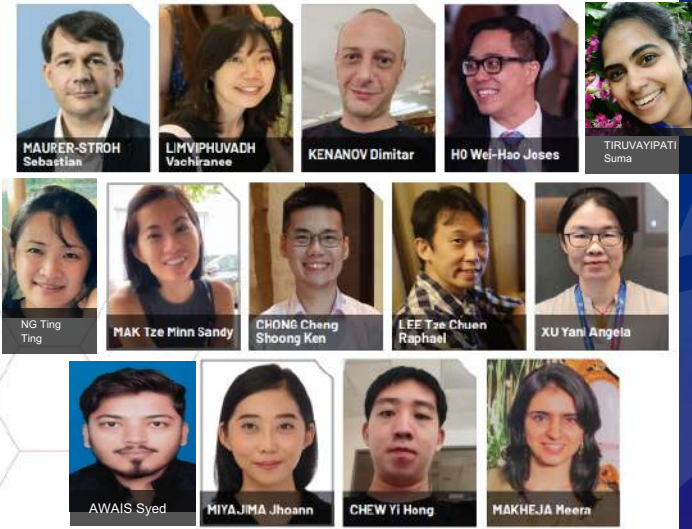
The image displays multiple overlapping screenshots of the GISAID EpiX™ web application interface. The main components visible are:

- Search and Navigation:** A top navigation bar with the GISAID logo and tabs for different pathogen types (EpiFlu™, EpiCoV™, EpiRSV™, EpiPox™, EpiArbo™, EpiX™). A search bar is prominently featured.
- Released Files Table:** A table listing various viral sequences with columns for Name, Accession ID, Location, and Substitutions. The table shows a large number of entries, with a total count of 368,052 isolates.
- Single Upload Form:** A detailed form for uploading genetic sequence data. It includes sections for:
  - Pathogen detail:** Submission name, Accession ID, Pathogen Kingdom, Pathogen Family, and Passage details/history.
  - Sample information:** Collection date, Location, Additional location information, Host, Additional host information, Sampling strategy, Gender, Patient age, Patient status, and Additional clinical information.
  - Outbreak Detail:** Specimen source and Last vaccinated.
- User Interface Elements:** The interface includes a user profile dropdown (e.g., "You are logged in as Sebastian Maurer-Stroh"), a "Submit for Review" button, and various help and navigation icons.

We gratefully acknowledge the Authors from Originating and Submitting laboratories of sequence data on which the analysis is based.



## Bioinformatics Institute (BII) Protein Sequence Analysis



# THANK YOU!