AI-powered Structural variation detection

Kai Ye

Xi'an Jiaotong University, China

kaiye@xjtu.edu.cn

Education and Work experience

Education

■ 1995~2003	Wuhan University, CN	Biopharmaceutical Science	B.S.; M.S.
■ 2004~2008	Leiden University, NL	Computer Science	PhD

Work

■ 2008~2009	European Bioinformatics Institute, UK	Supervised by Rolf Apweiler	Postdoc
■ 2009~2012	Leiden University Medical Center, NL	Molecular Epidemiology	Assistant professor
■ 2012~2016	Washington University in St. Louis, US	Genome Institute	Assistant professor
■ 2016~ now	Xi’an Jiaotong University, CN	Automation	Professor

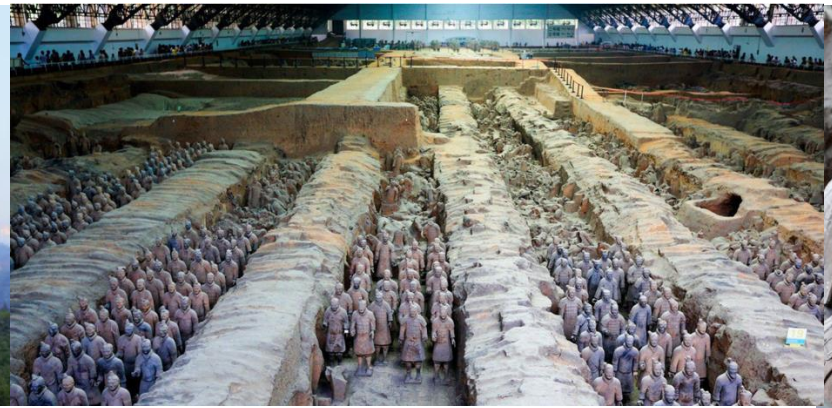
Education and Work experience

Education

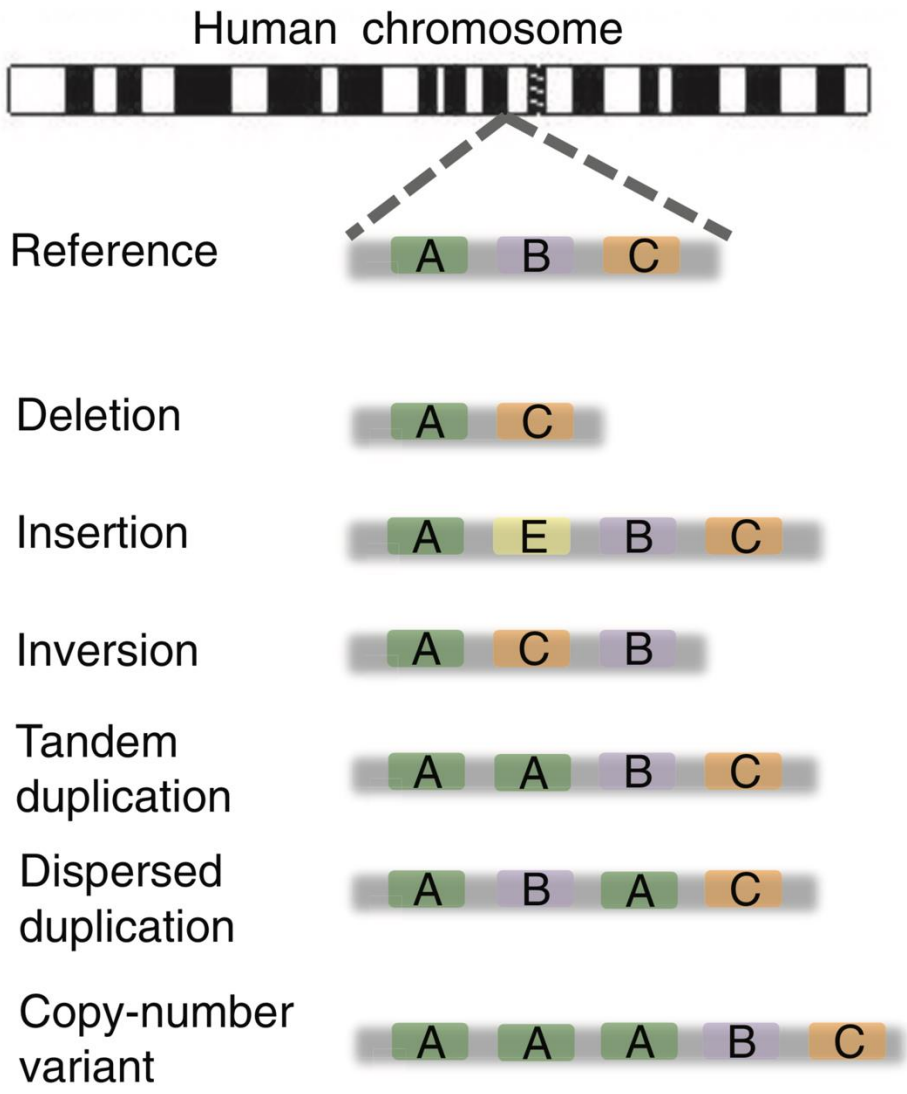
■ 1995~2003	Wuhan University, CN	Biopharmaceutical Science	B.S.; M.S.
■ 2004~2008	Leiden University, NL	Computer Science	PhD

Work

■ 2008~2009	European Bioinformatics Institute, UK	Supervised by Rolf Apweiler	Postdoc
■ 2009~2012	Leiden University Medical Center, NL	Molecular Epidemiology	Assistant professor
■ 2012~2016	Washington University in St. Louis, US	Genome Institute	Assistant professor
■ 2016~ now	Xi'an Jiaotong University, CN	Automation	Professor

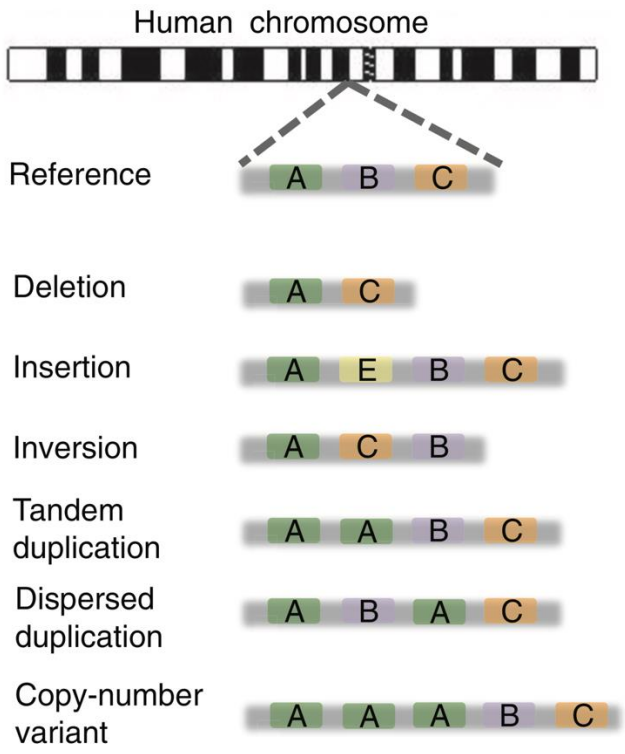


Structural variation (SV)

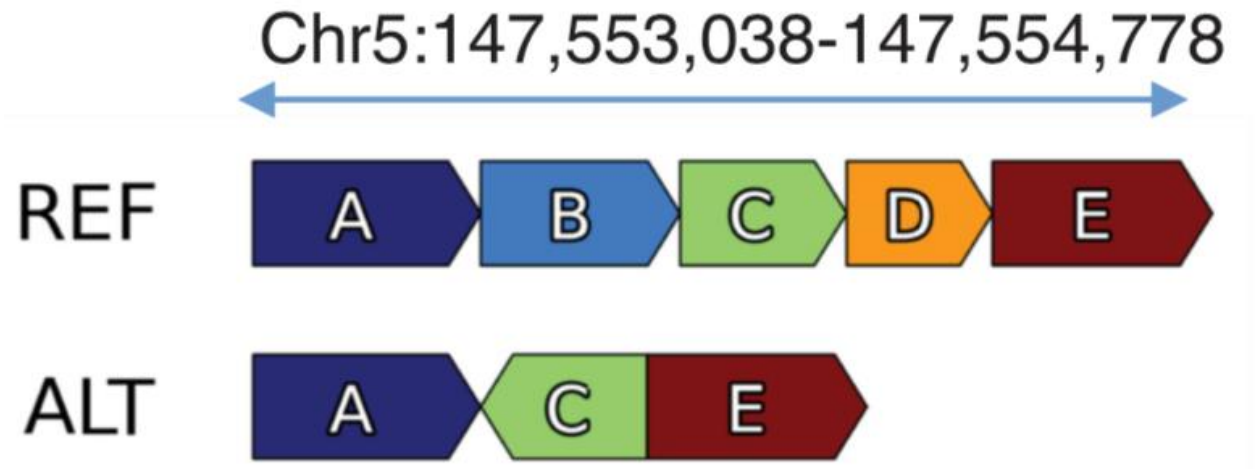


Structural variation (SV)

Simple SV

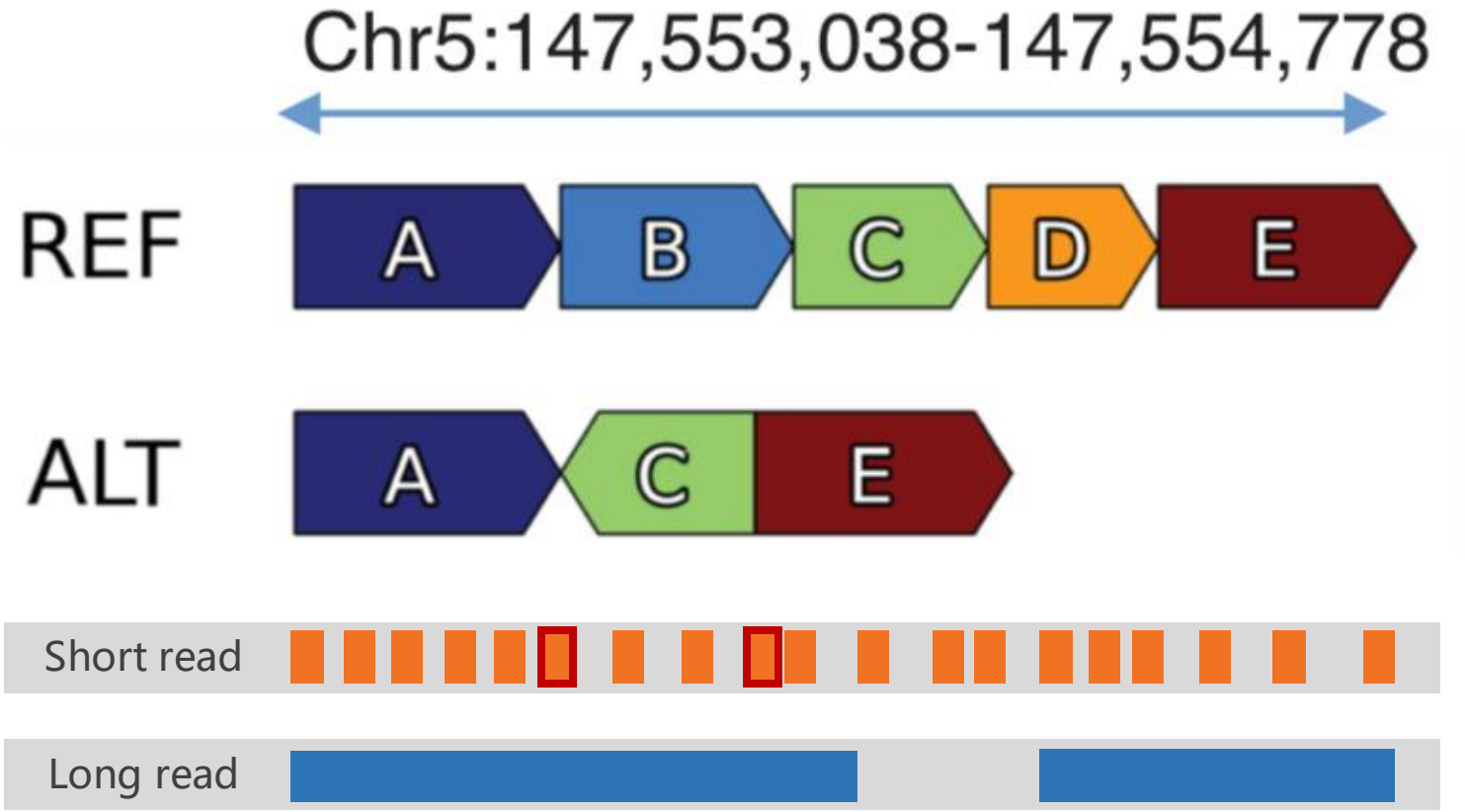


Complex SV



Combination of multiple simple alternations

Long reads are better for SV detection

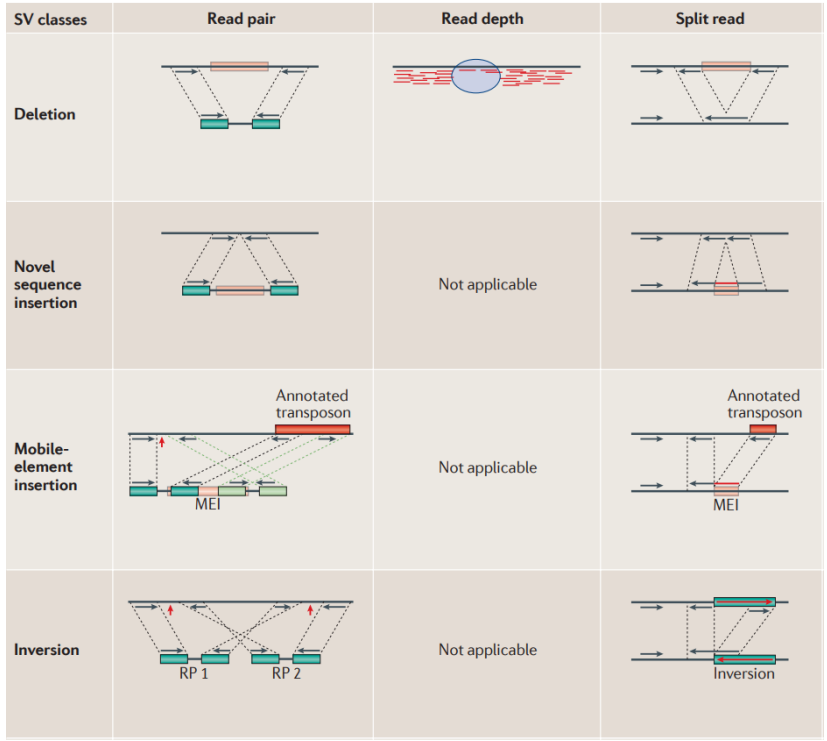


Short reads provide linking information between two adjacent segments

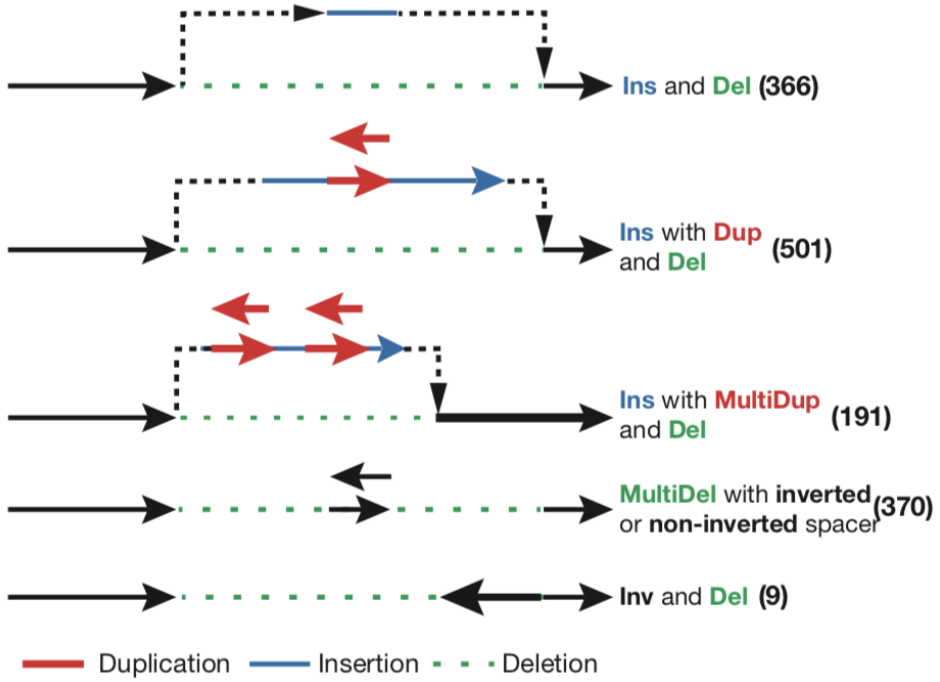
but too short to reveal connection among breakpoints

Challenges of detecting complex SV with long reads (1)

Model based approaches

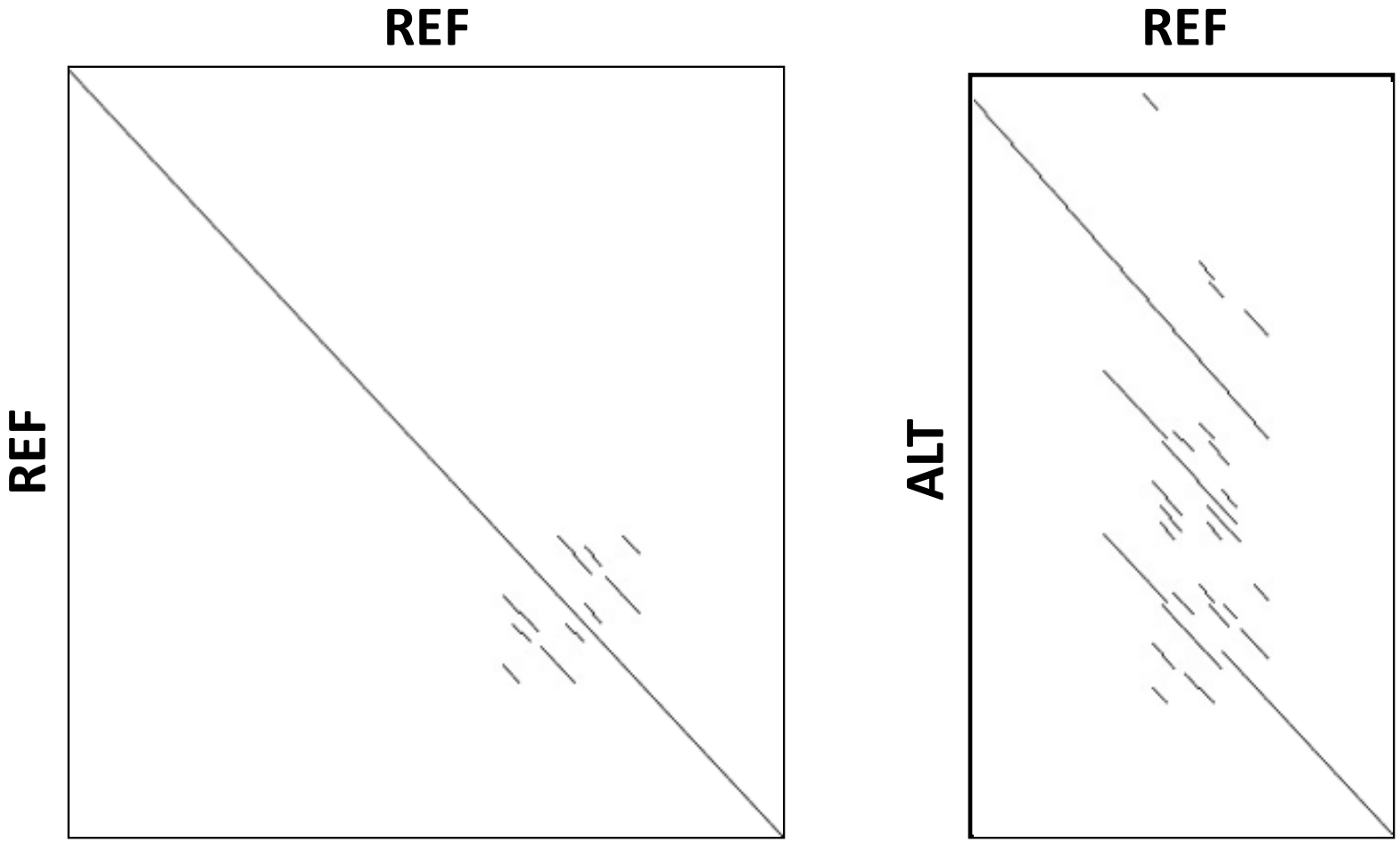


Hard to model and code



Model-based approach hinders detection of unknown types of complex SV

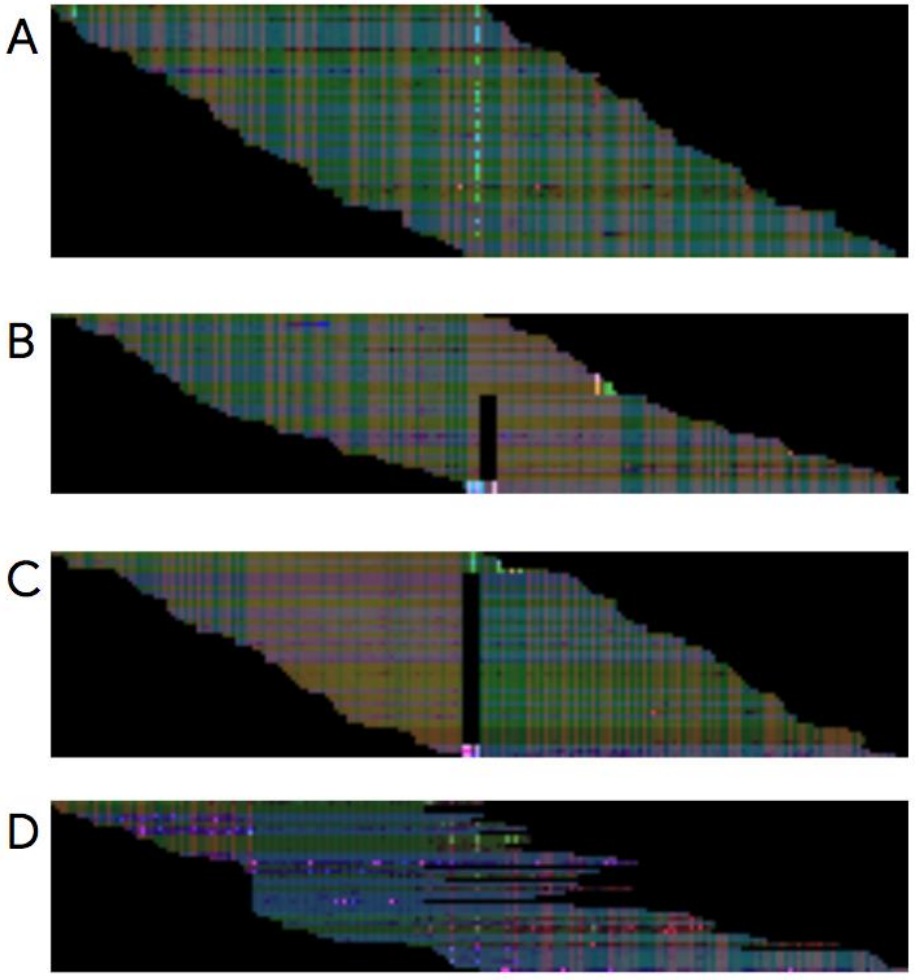
Challenges of detecting complex SV with long reads (2)



Repetitive regions complicate detection of SV

❖ Deep learning in variant detection

Google DeepVariant

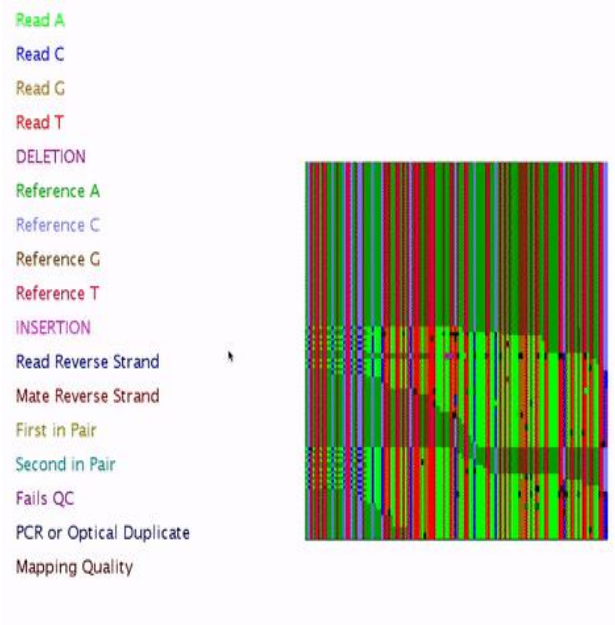


True SNP on one chromosome pair.

Deletion on one chromosome.

Deletion on both chromosomes.

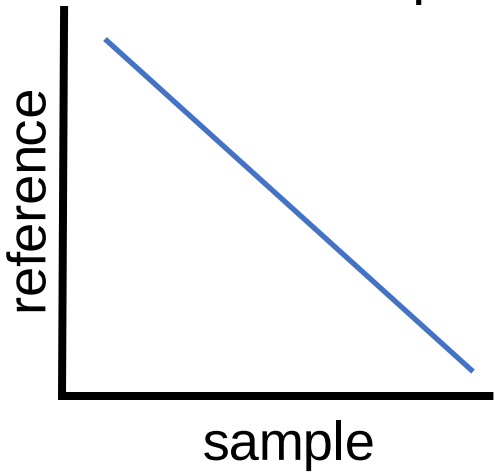
False variant caused by errors.



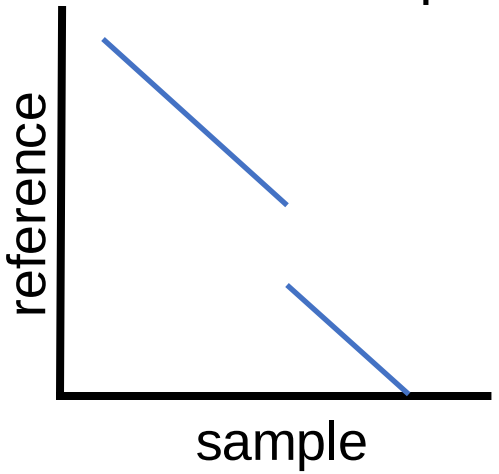
The sequence-to-image coding strategy of DeepVariant does not work on SV

Coding sequence similarity

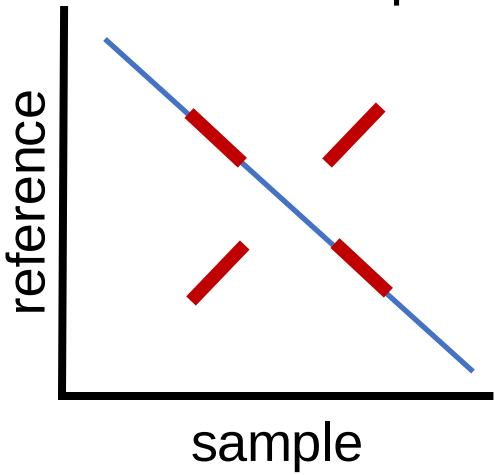
Identical & no repeat



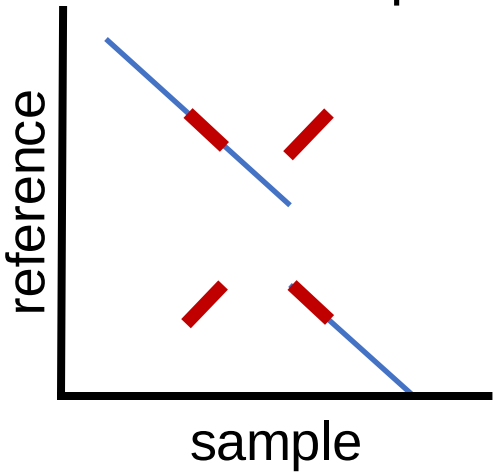
Different & no repeat



Identical & repeat



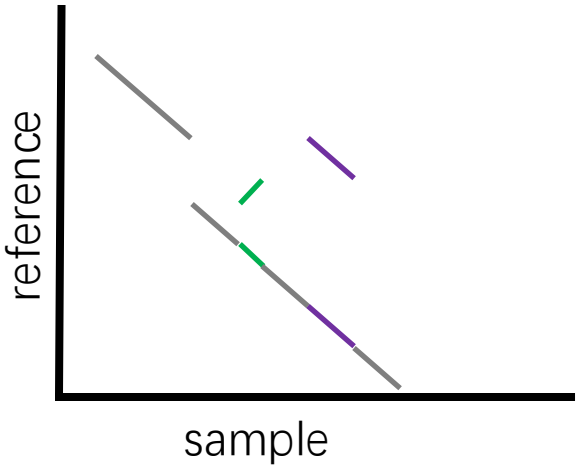
Different & repeat



Coding sequence similarity with denoised dotplot

Encode SV

ALT-to-REF

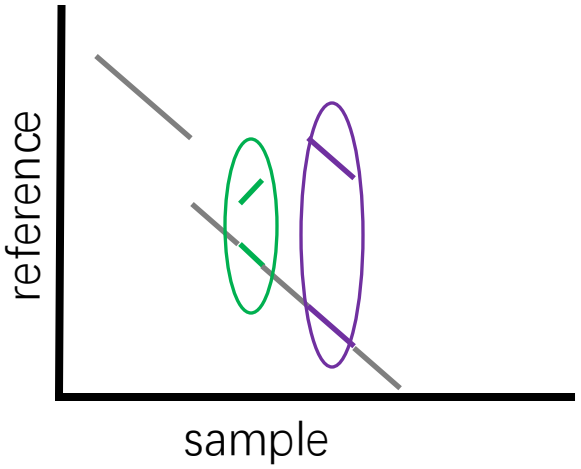


Lin et al., ***Nature Methods***, 2022

Coding sequence similarity with denoised dotplot

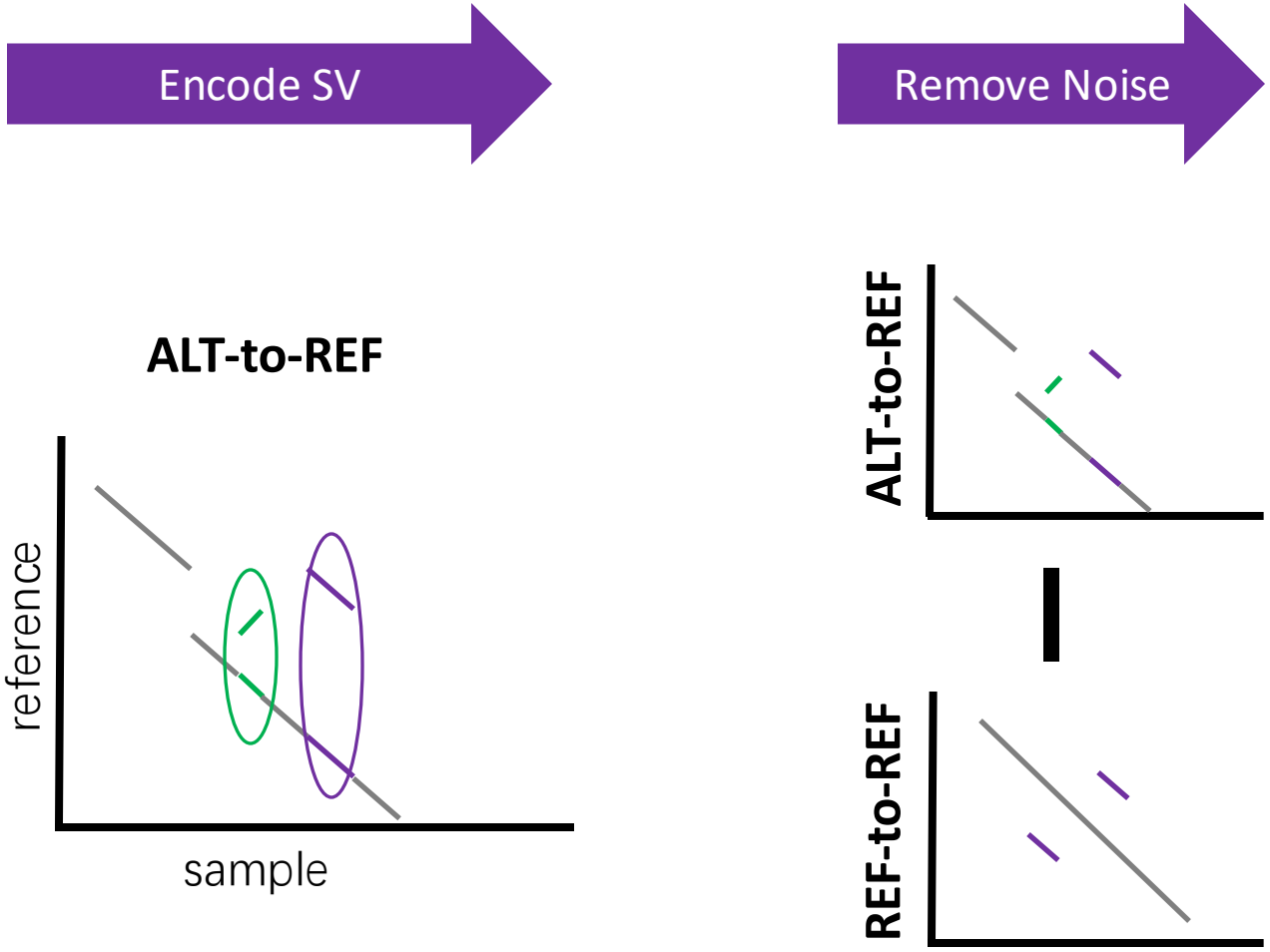
Encode SV

ALT-to-REF



Lin et al., ***Nature Methods***, 2022

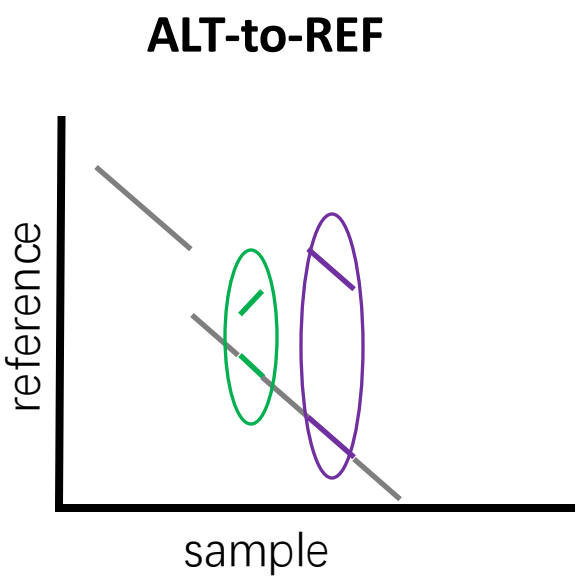
Coding sequence similarity with denoised dotplot

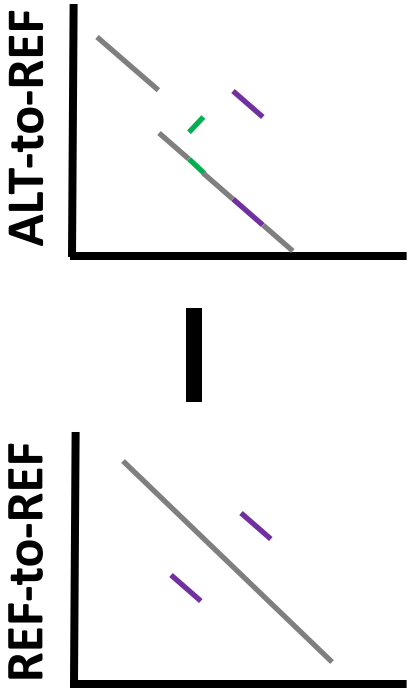


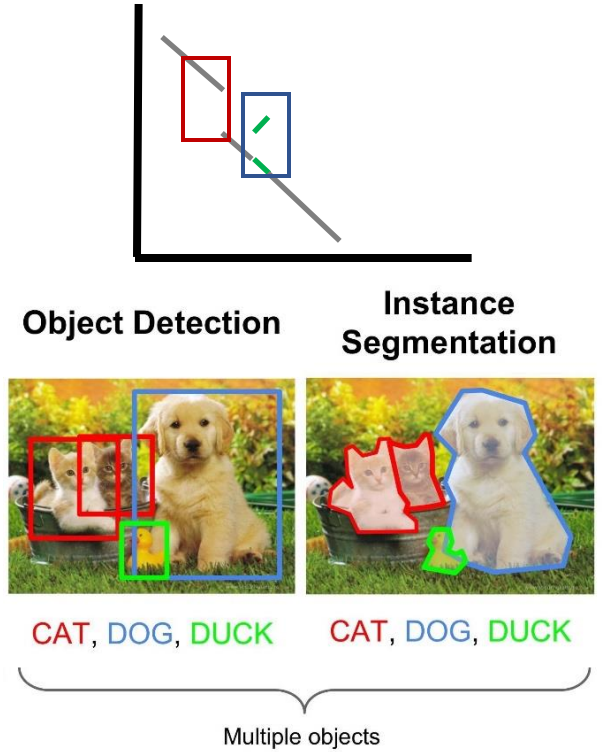
Lin et al., ***Nature Methods***, 2022

Coding sequence similarity with denoised dotplot

Encode SV → Remove Noise → Detection

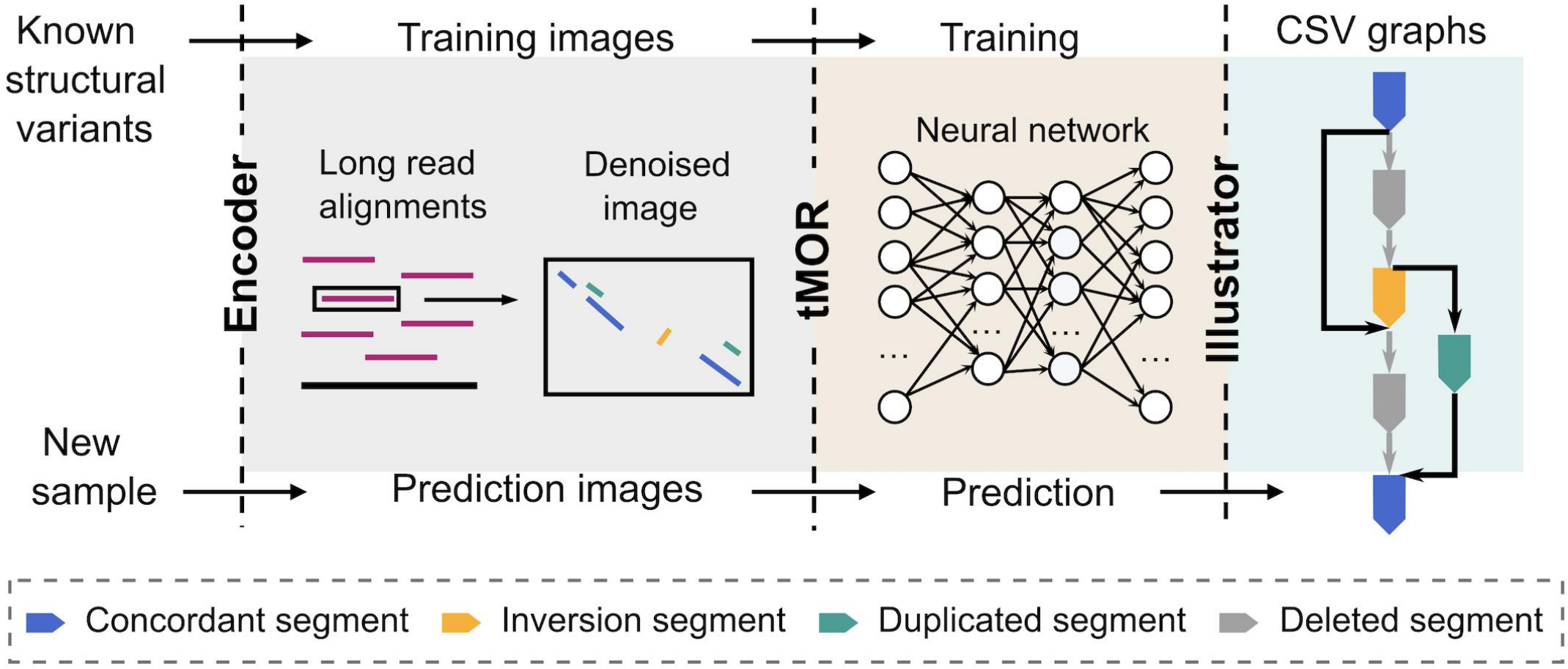






Lin et al., ***Nature Methods***, 2022

SVision workflow



Lin et al., ***Nature Methods***, 2022

Image representation and object detection

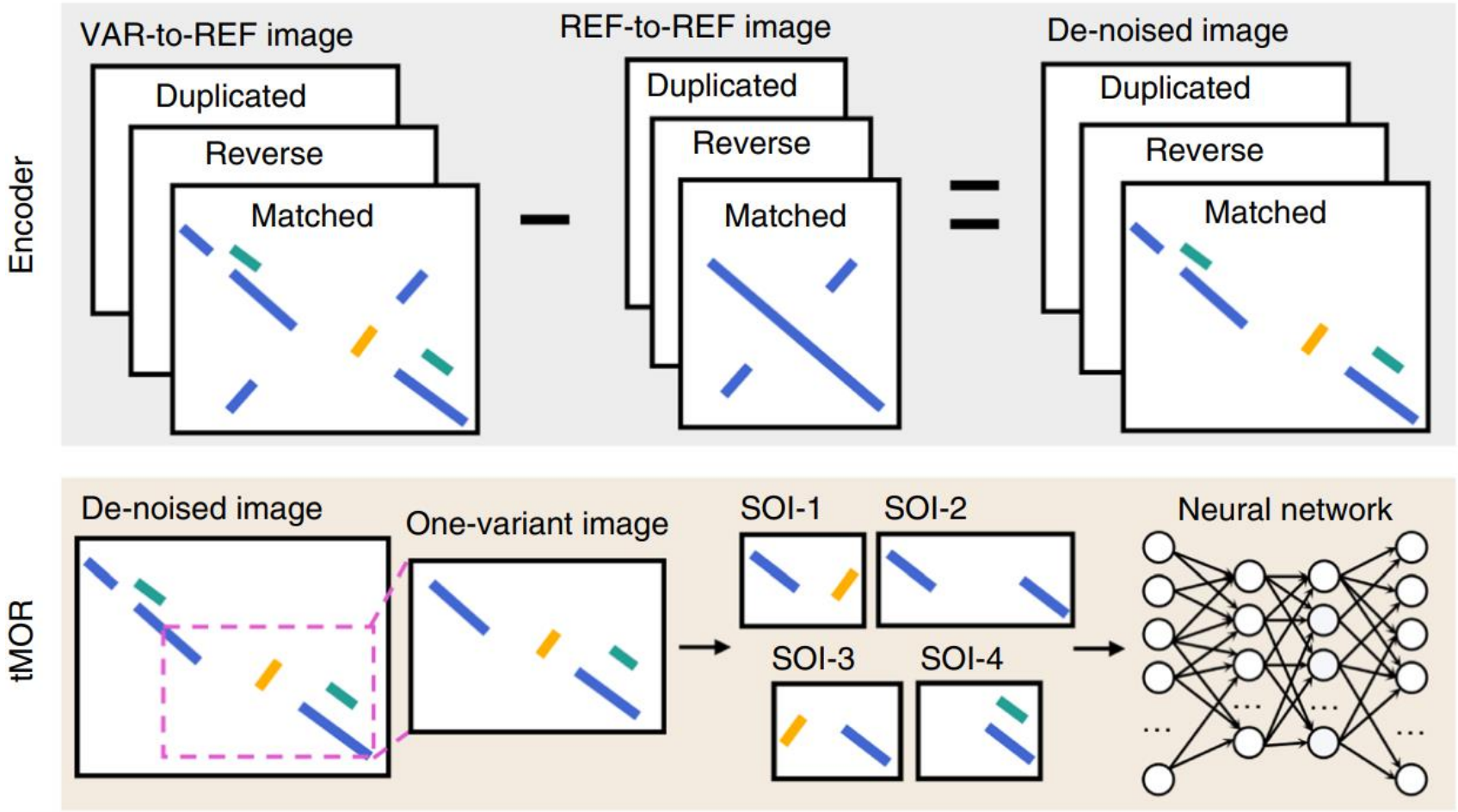


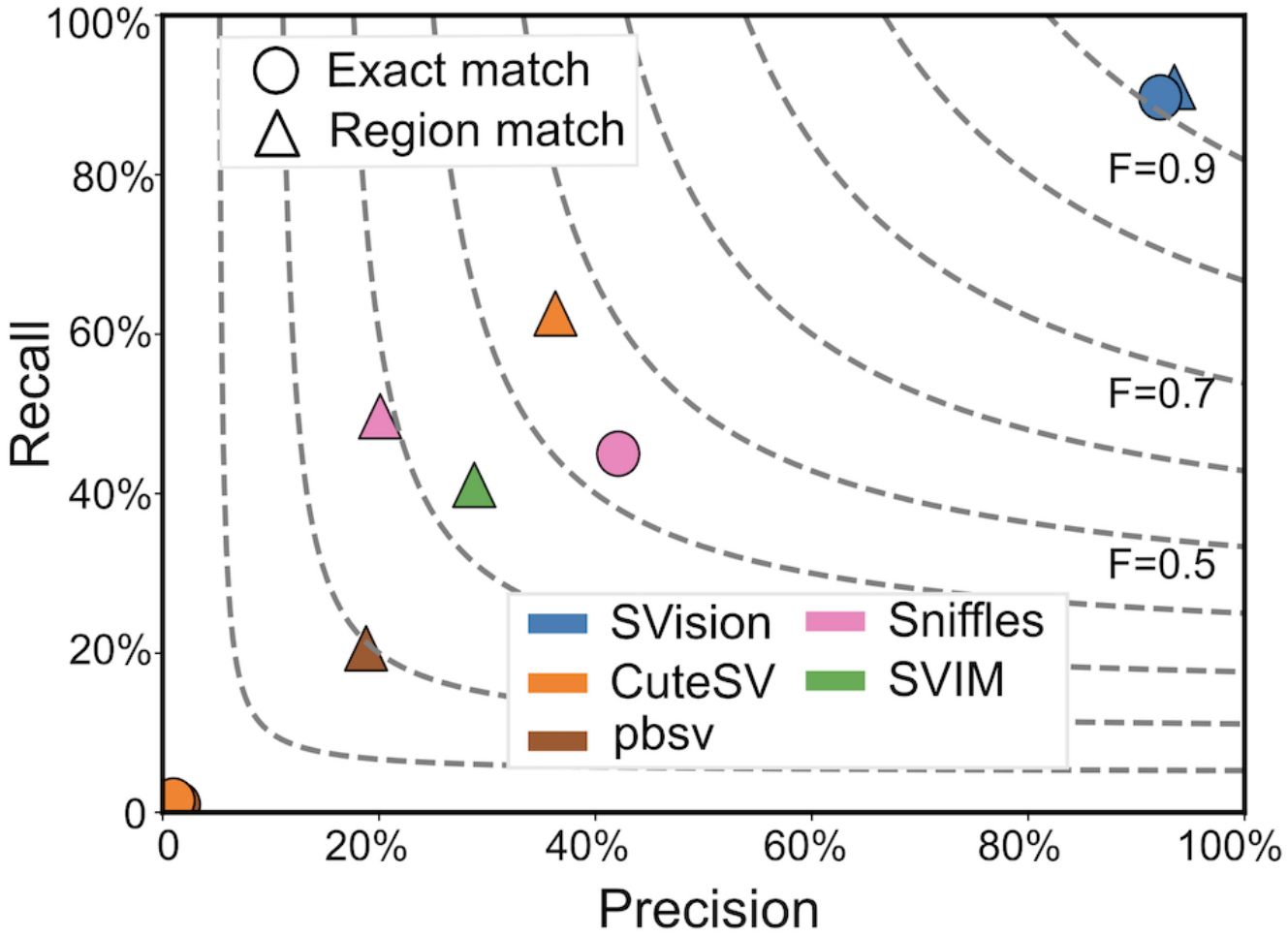
Image representation and de-noising

Object detection

Five basic structural variant types

Lin et al., ***Nature Methods***, 2022

Calling simulated complex SV

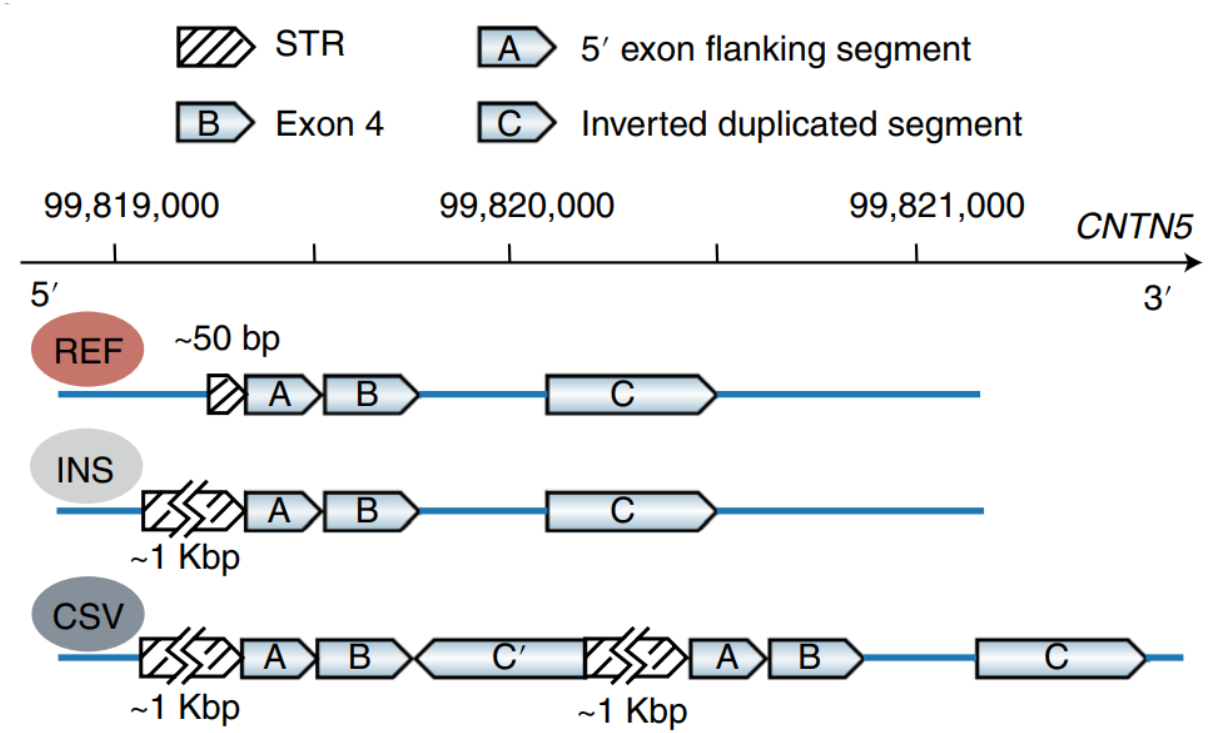


- Region match: correct discovery a CSV site;
- Exact match: correct discovery of CSV site and its subcomponents

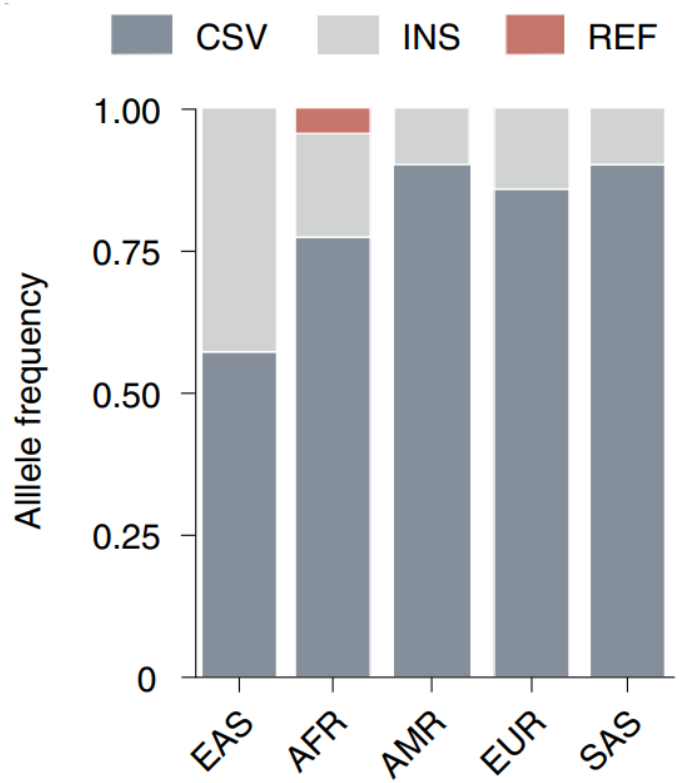
Lin et al., ***Nature Methods***, 2022

❖ Resolving gene-related complex SV

Resolved CSV in *CNTN5* coding region



Population diversity



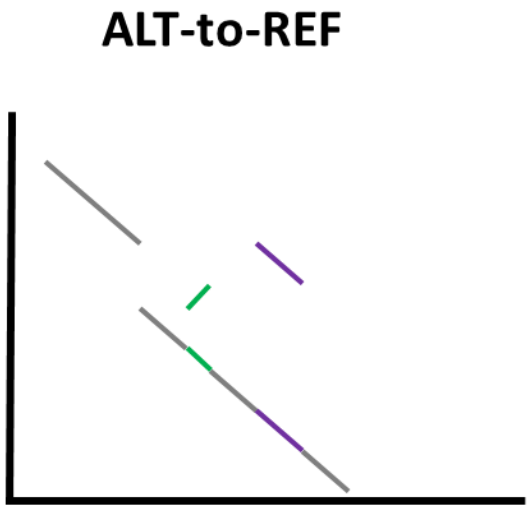
Lin et al., ***Nature Methods***, 2022

❖ Where we did well and how to improve?

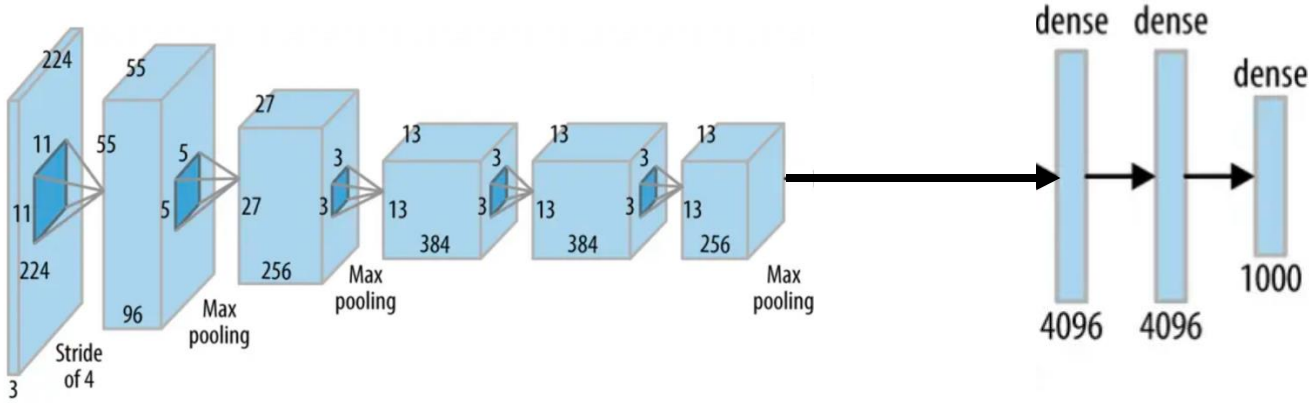
Sequence to image

Feature extraction

Classification



AlexNet

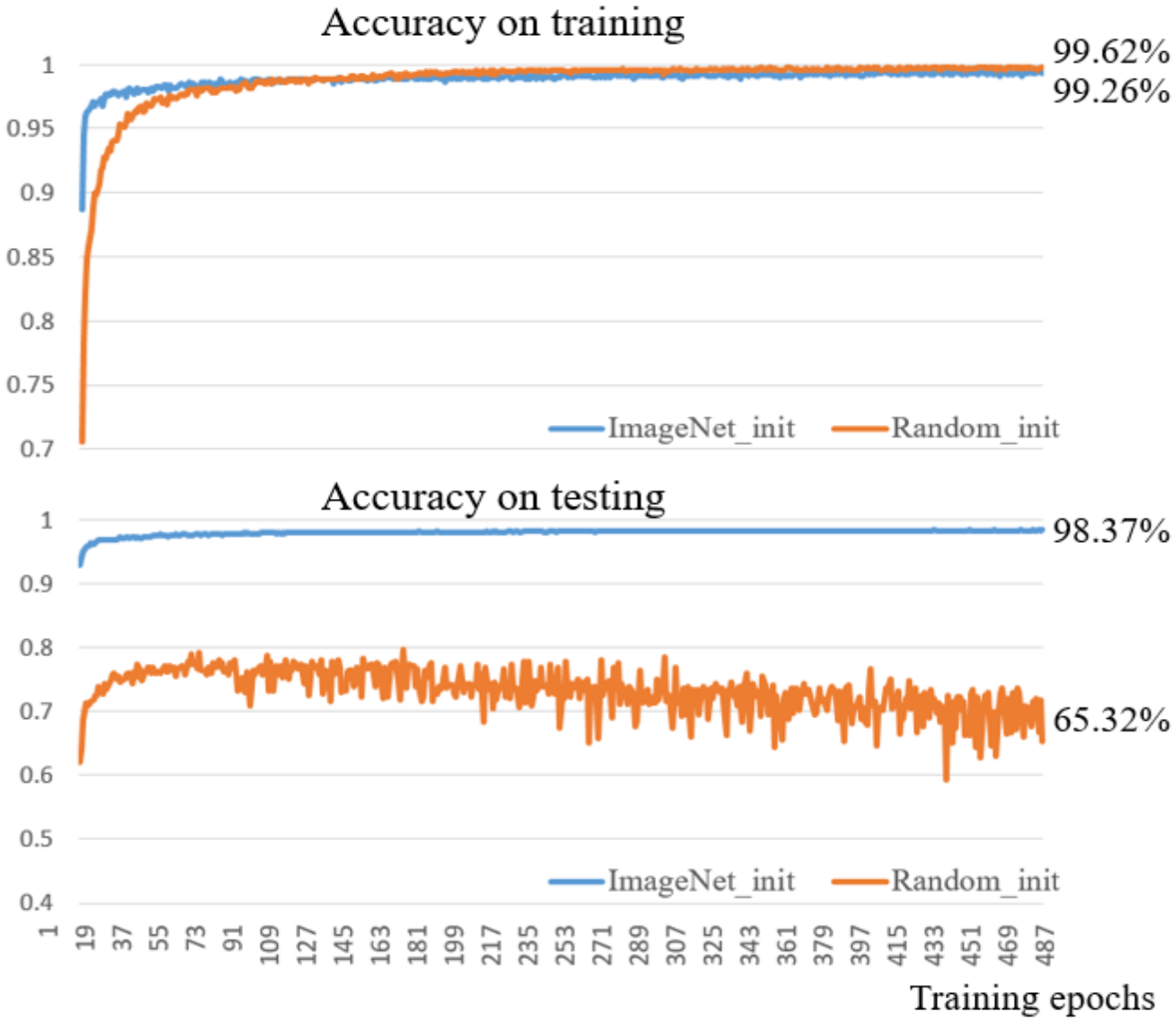


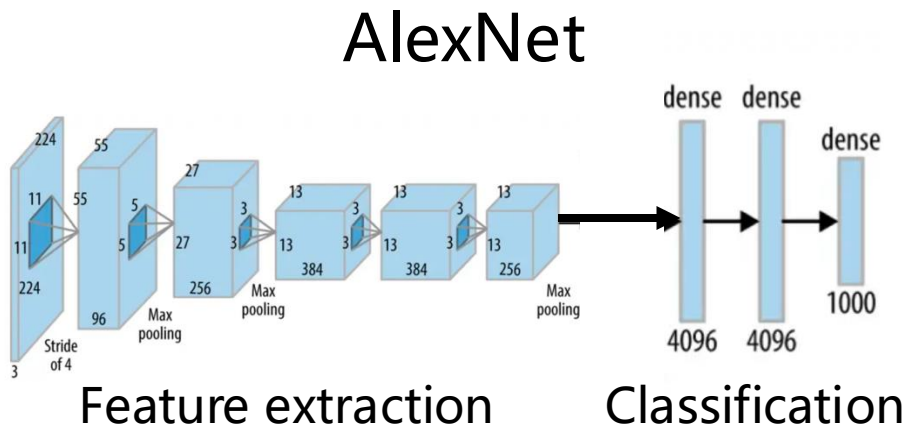
Denoised similarity image

ImageNet setting

trained

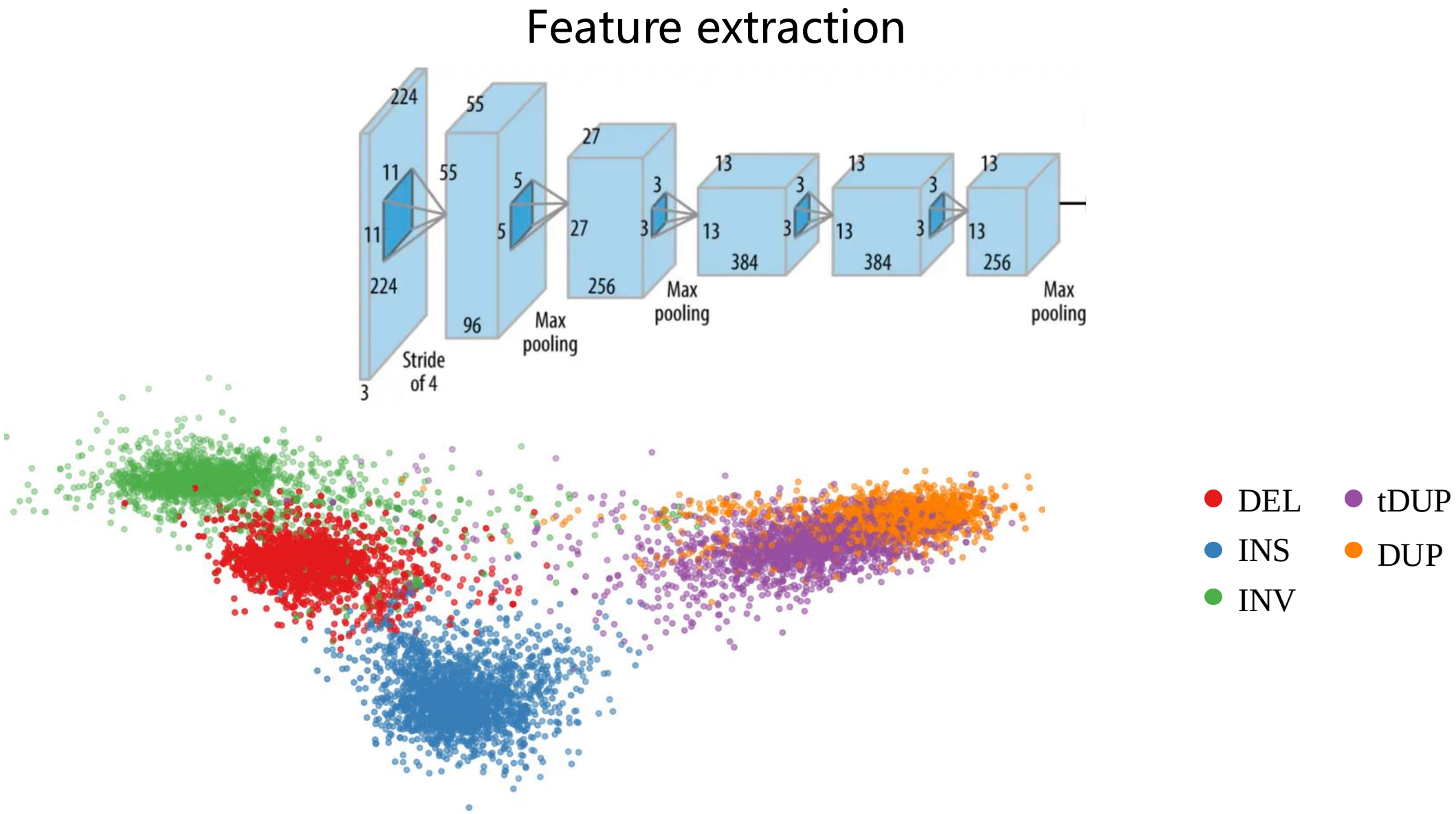
❖ Training AlexNet for SVision





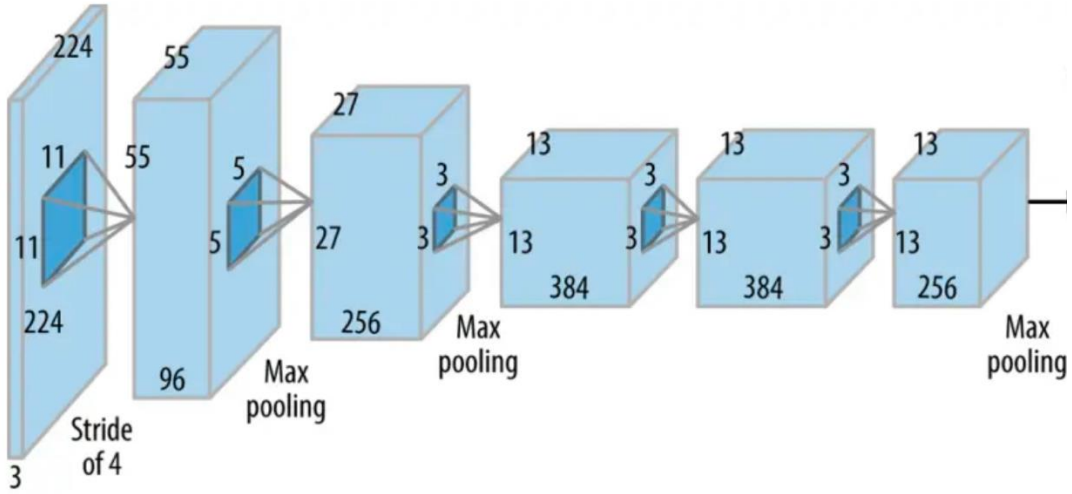
- ImageNet_init = set 5 feature extraction layers as ImageNet setting, only train 3 classification layers
- Random_init = train all 8 layers (5 feature extraction + 3 classification layers)

❖ AlexNet with ImageNet_init setting extracts fine features from training set



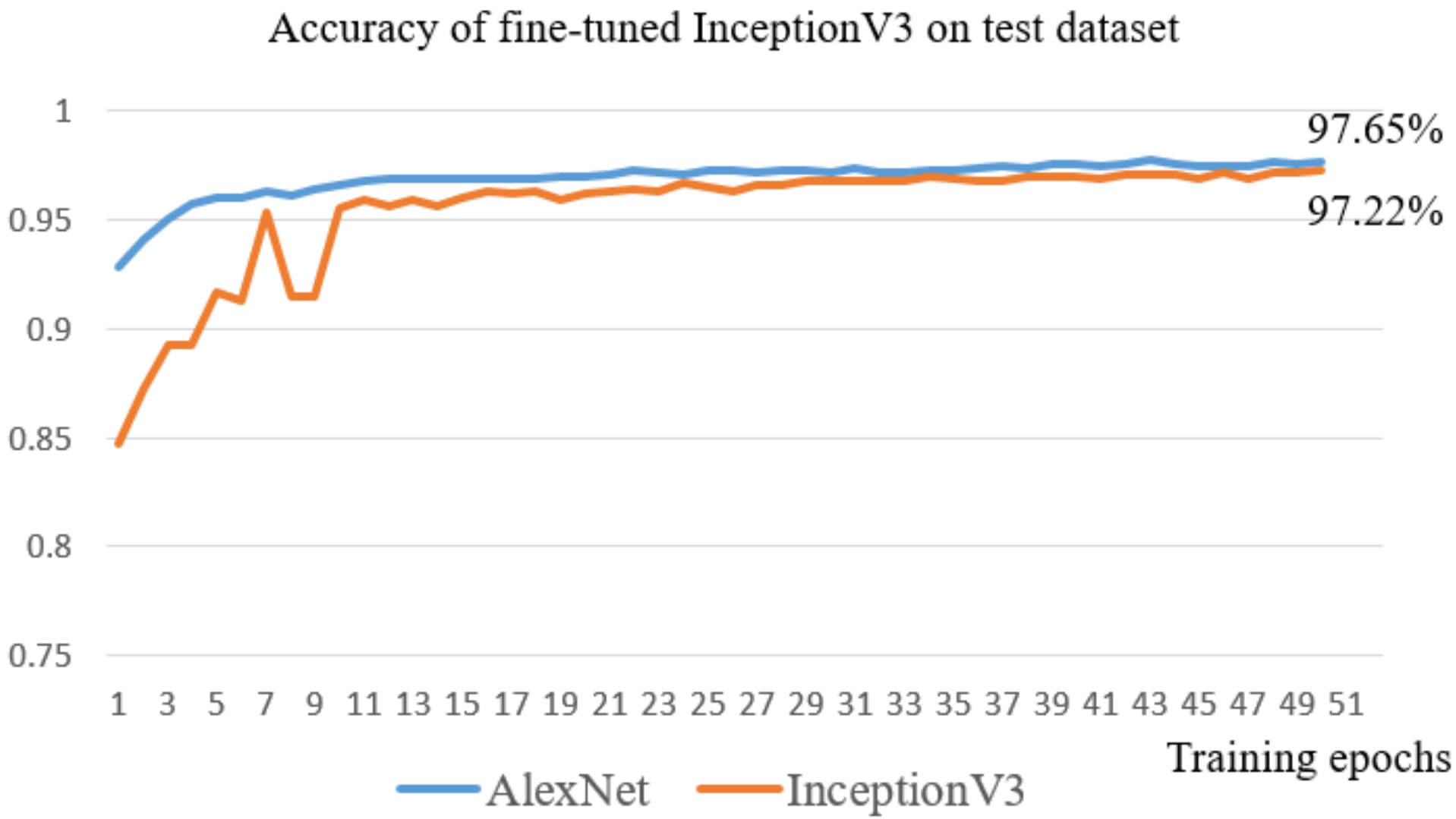
❖ Features extracted by AlexNet with ImageNet_init fit other ML approaches

Feature extraction



Methods	Recall	Precision
Logistic regression	88.49%	91.14%
SVM	97.12%	97.09%
Random forests	96.59%	96.77%

❖ Other CNN framework works too

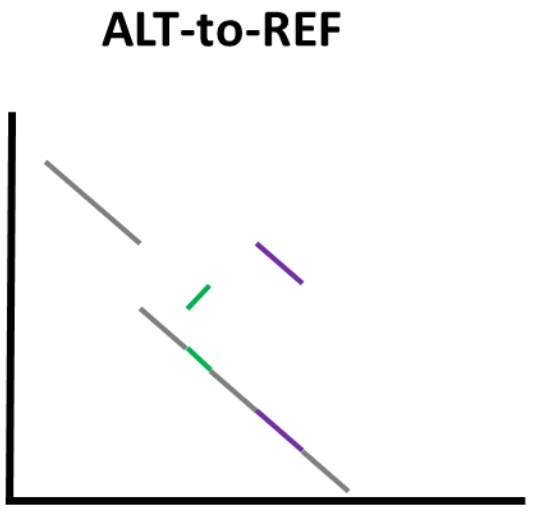


❖ Where we did well and how to improve?

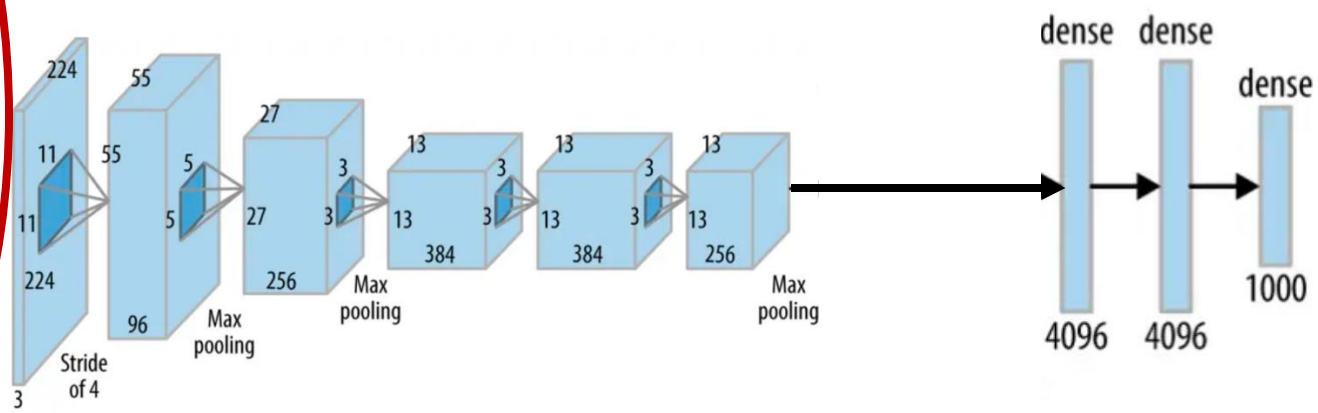
Sequence to image

Feature extraction

Classification



AlexNet



Denoised similarity image

ImageNet setting

trained

nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01609-w

Check for updates

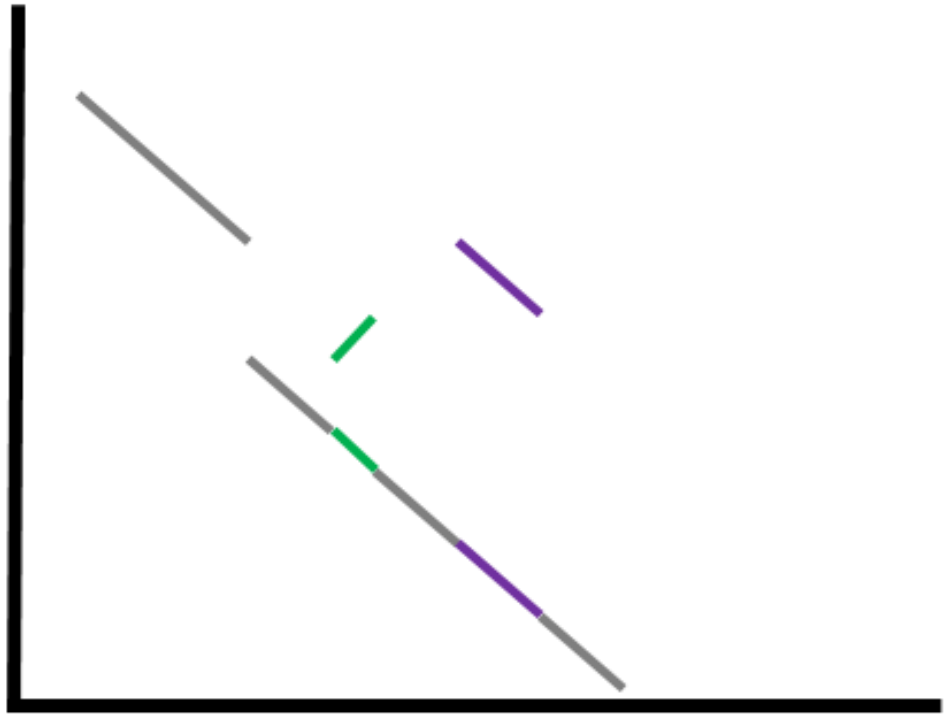
SVision: a deep learning approach to resolve complex structural variants

Jiadong Lin[1,2,3,4,13], Songbo Wang[1,2,3,13], Peter A. Audano[5], Deyu Meng[1,6,7], Jacob I. Flores[5], Walter Kosters[4], Xiaofei Yang[1,8], Peng Jia[1,2,3], Tobias Marschall[9], Christine R. Beck[5,10] and Kai Ye[1,2,3,11,12] ✉

Where we did well and how to improve?

SVision Seq2image

Compare ref and sample



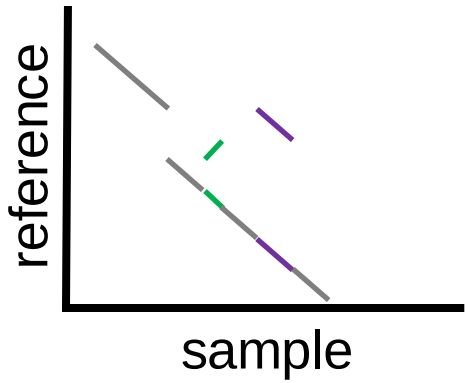
SVision Seq2image encodes

- Mapping information
- Sequence features
- Domain knowledge

New application scenarios: SVision-pro

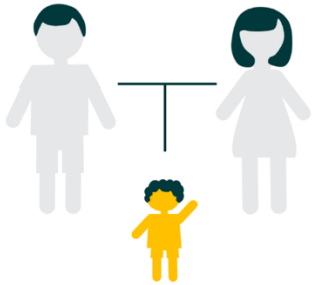
Single genome variant calling

Sample vs Reference
germline

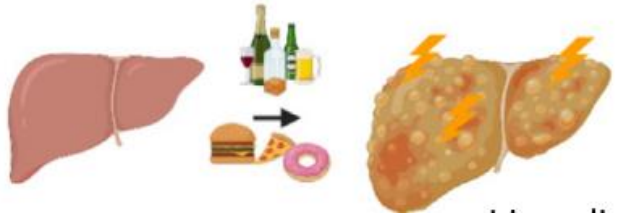


Multi-genome comparison

Sample vs Sample
De novo or somatic



De novo variants

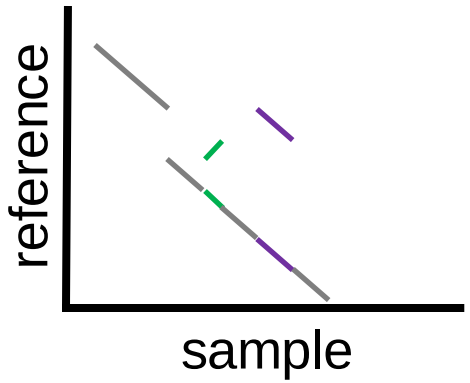


Somatic variants

New application scenarios: SVision-pro

Single genome variant calling

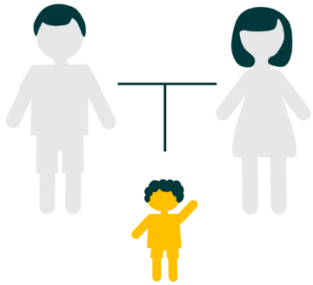
Sample vs Reference
germline



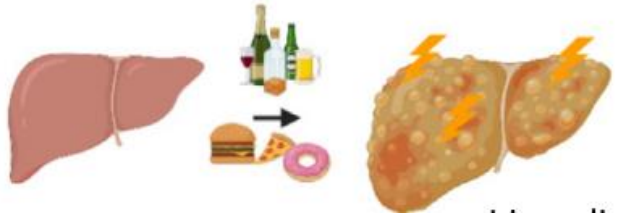
Diff(reference, sample)

Multi-genome comparison

Sample vs Sample
De novo or somatic



De novo variants



Somatic variants

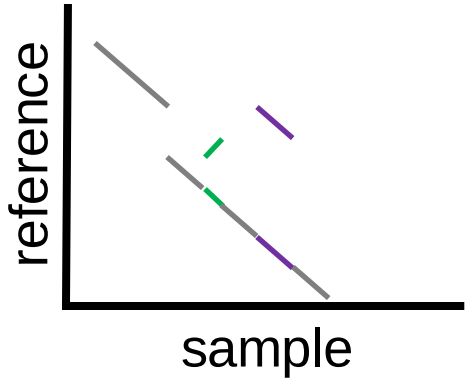
Diff(Diff(reference, normal), Diff(reference, tumor))

New application scenarios: SVision-pro

Single genome variant calling

Sample vs Reference

germline

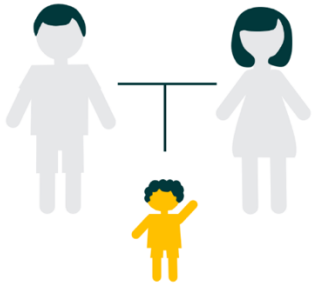


Diff(reference, sample)

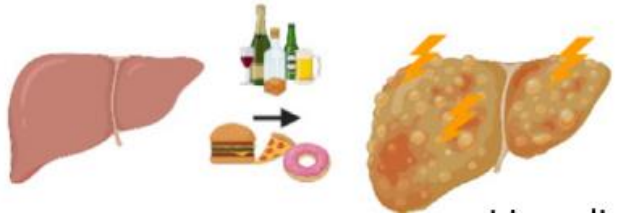
Multi-genome comparison

Sample vs Sample

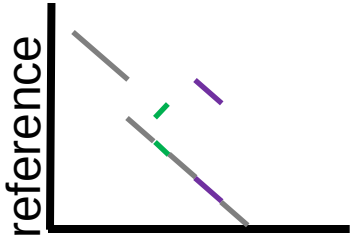
De novo or somatic

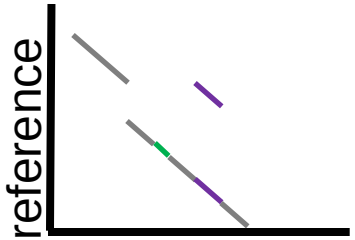


De novo variants



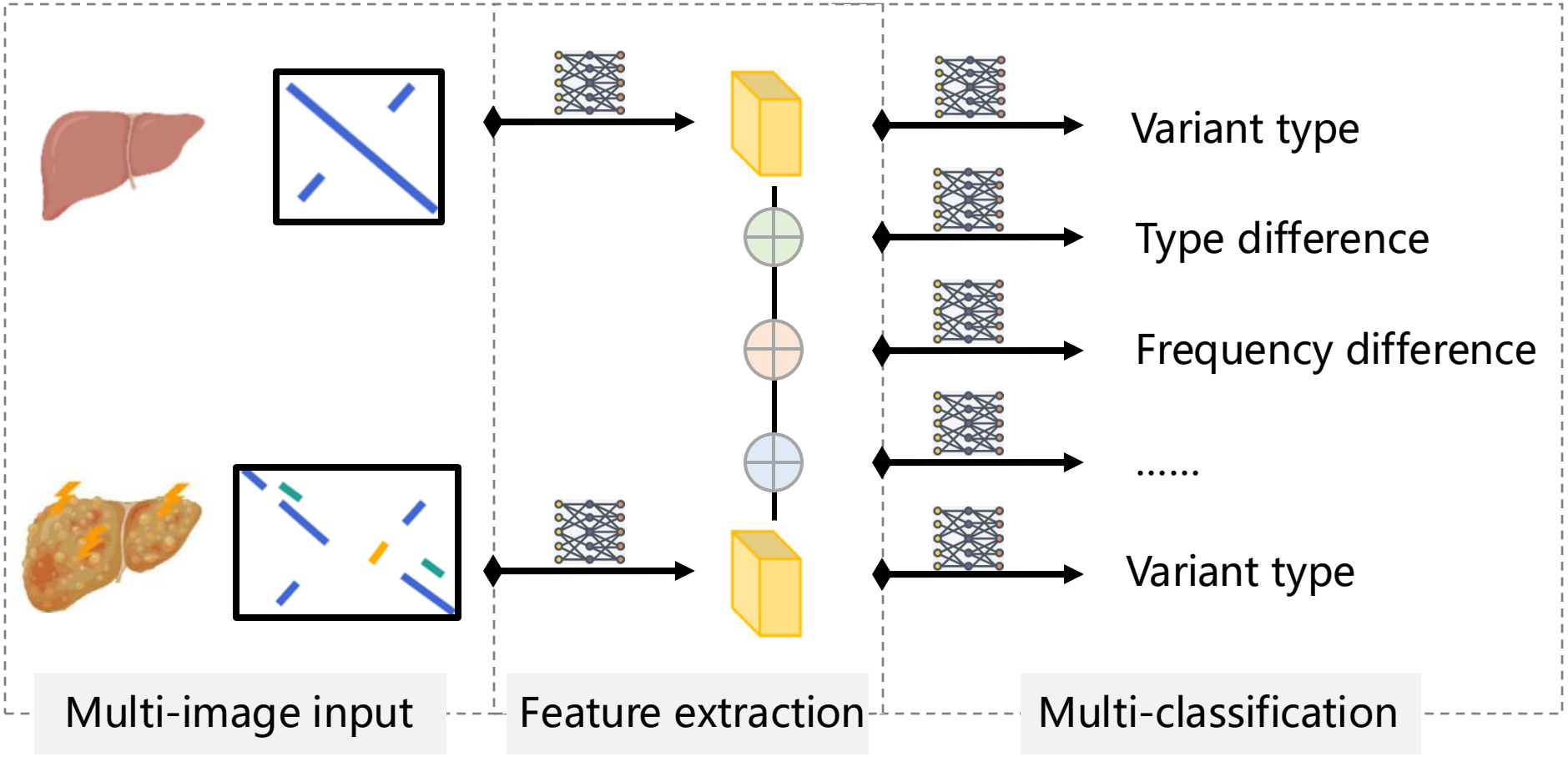
Somatic variants

Diff(

, 

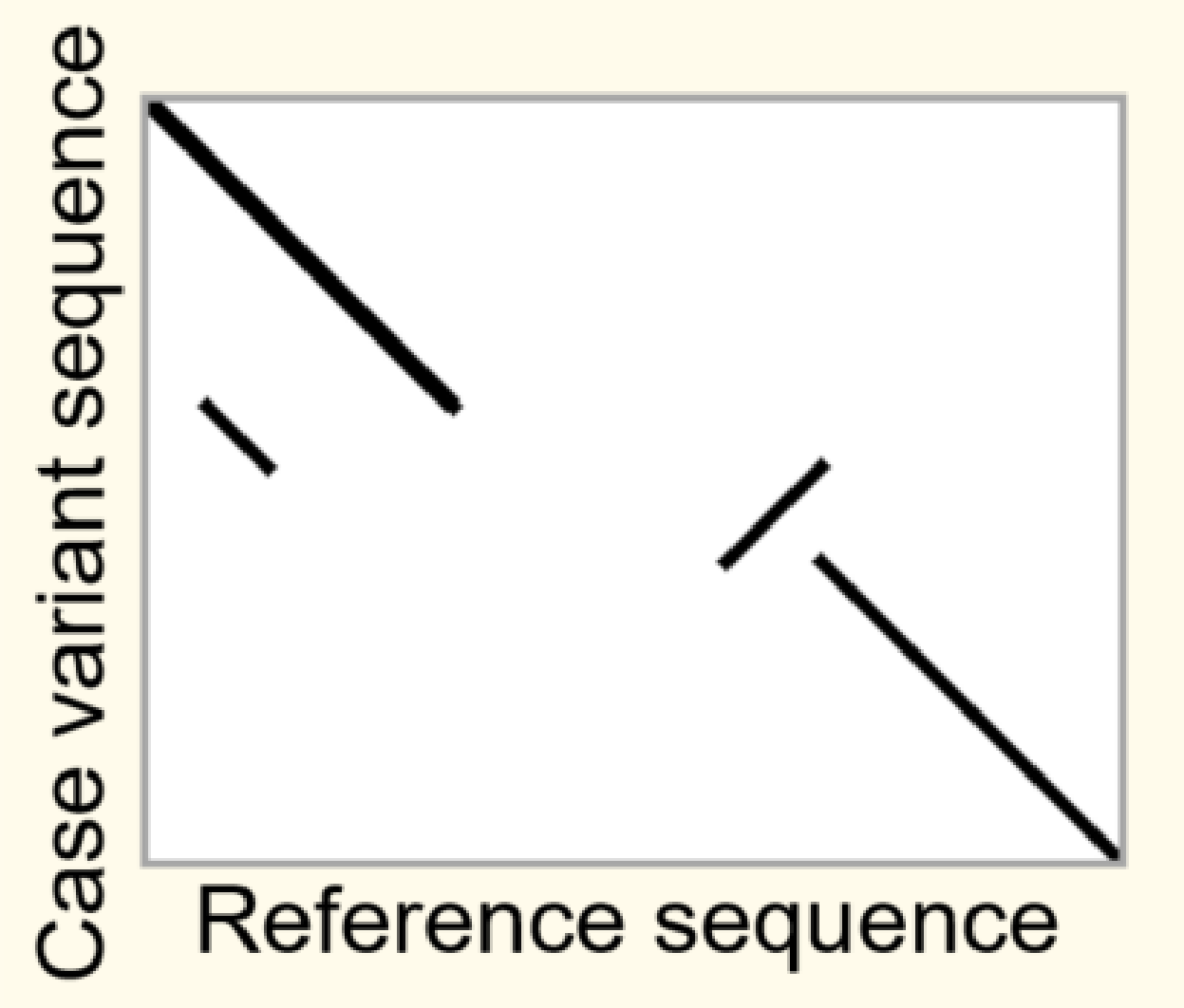
)

Do we directly compare two images? NO



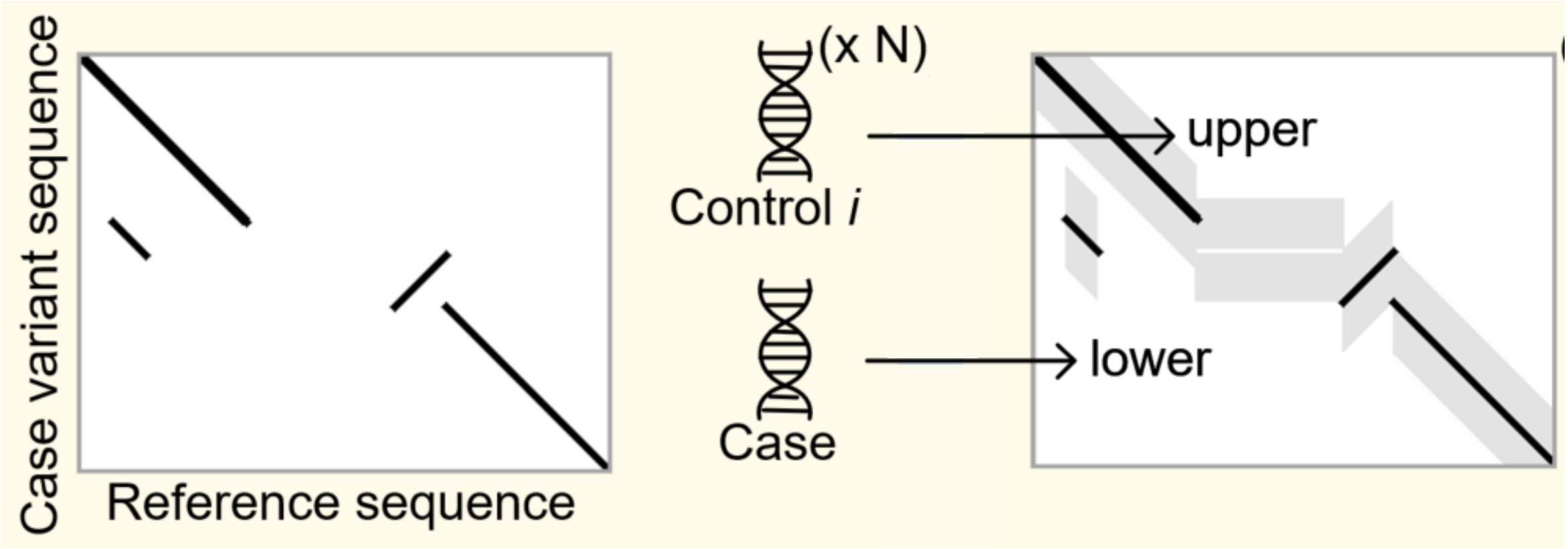
Abandon direct comparison due to computational burden
we need a lightweight and elegant approach

Our strategy: Genome-to-genome representation within one image



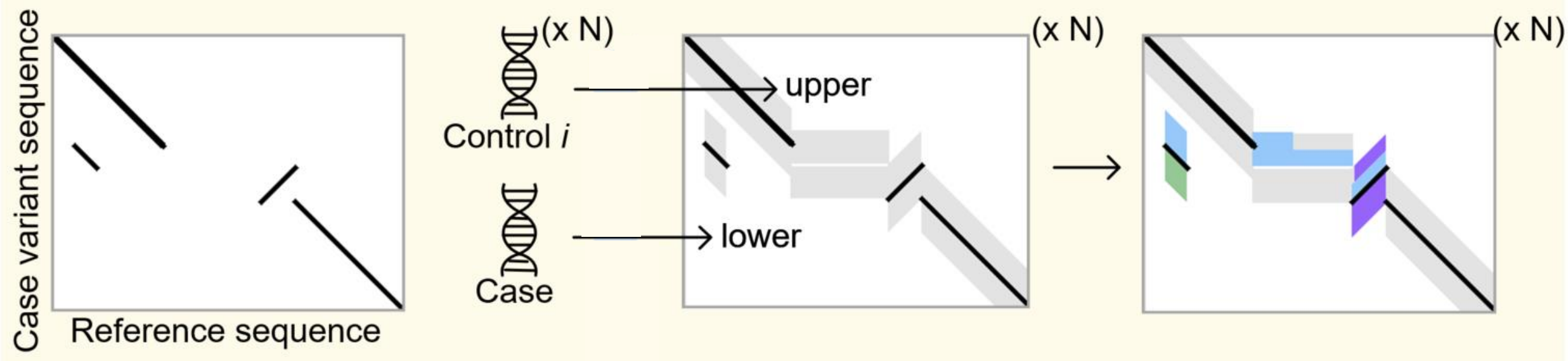
Wang et al., ***Nature Biotechnology***, 2024

Our strategy: Genome-to-genome representation within one image



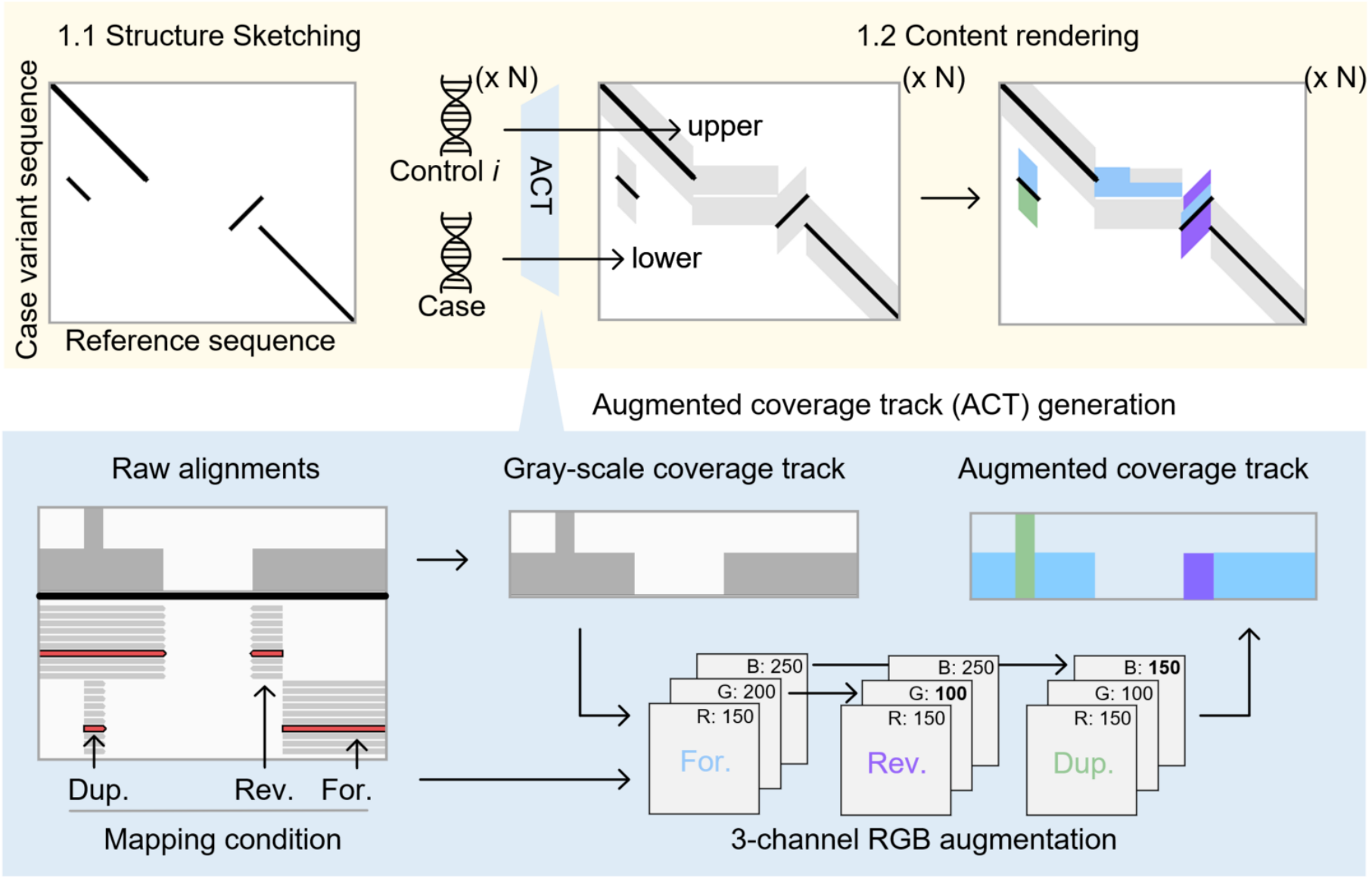
Wang et al., ***Nature Biotechnology***, 2024

Our strategy: Genome-to-genome representation within one image



Wang et al., ***Nature Biotechnology***, 2024

Our strategy: Genome-to-genome representation within one image



Wang et al., ***Nature Biotechnology***, 2024

Flexible image setting

Flexible image properties for different sensitivity requirements

Image size: 256 x 256
Track height: 25
Min-representable AF: 0.04

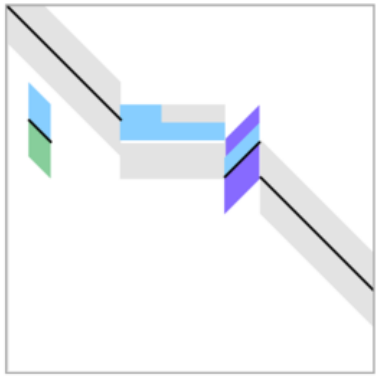


Image size: 512 x 512
Track height: 50
Min-representable AF: 0.02

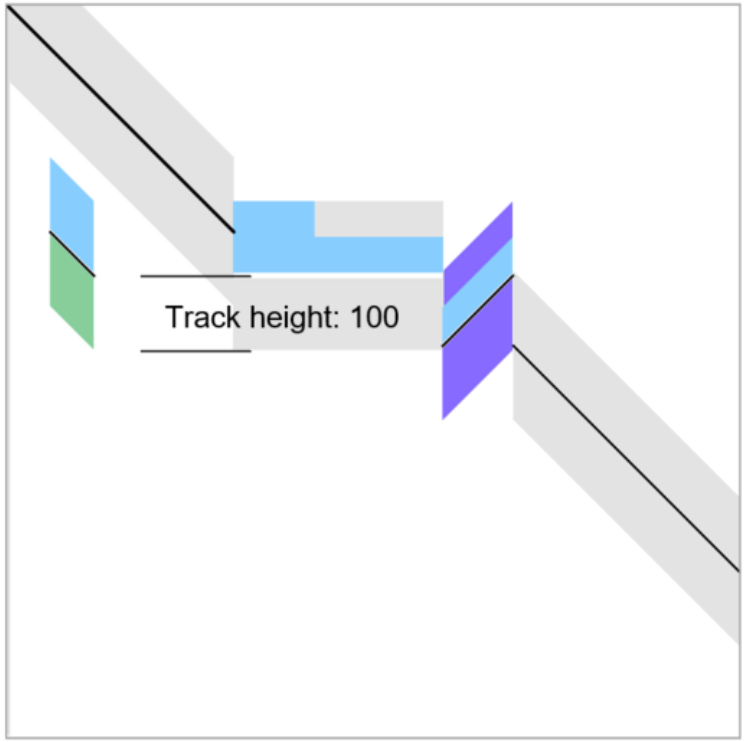


Image size: 1024 x 1024
Track height: 100
Min-representable AF: 0.01

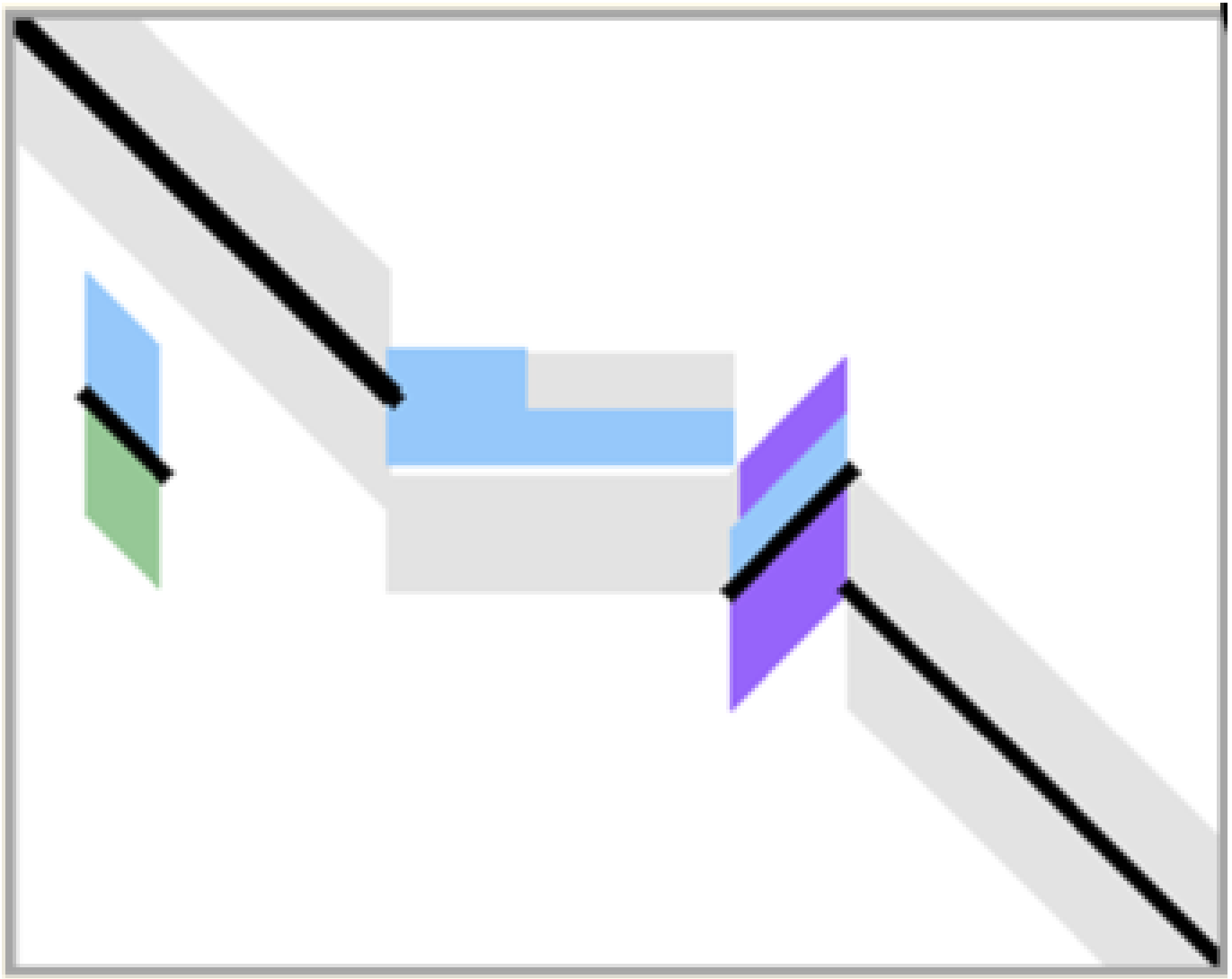
.....
(customization)

- Image sizes
- Track heights
- Colors

Larger image size leads to lower AF

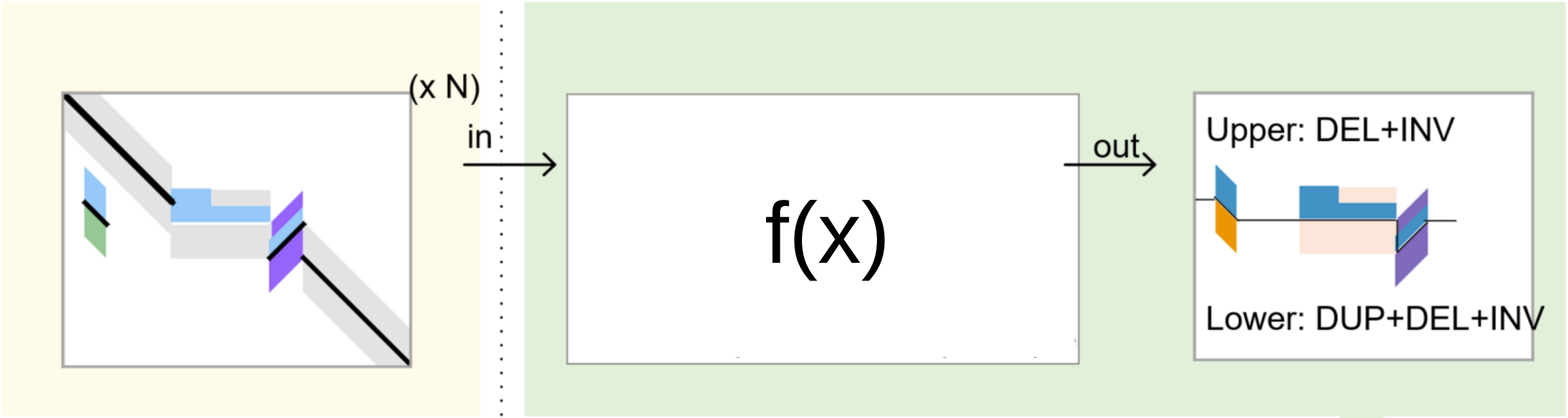
Wang et al., ***Nature Biotechnology***, 2024

Comparative recognition module



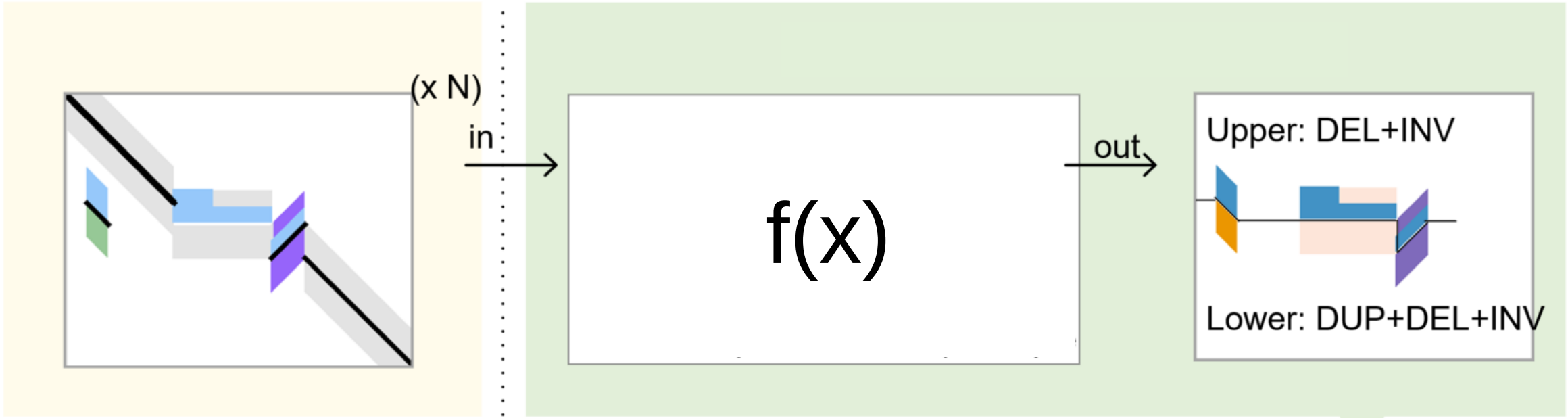
Wang et al., ***Nature Biotechnology***, 2024

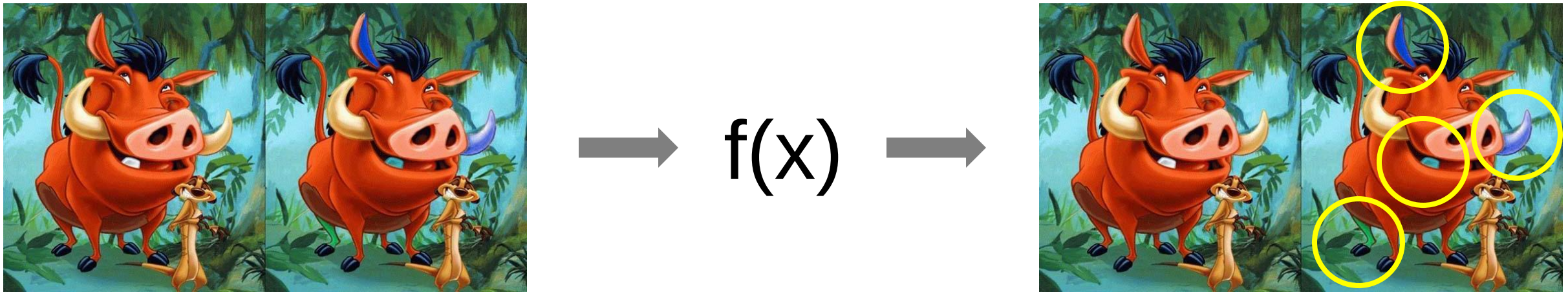
Comparative recognition module



Wang et al., ***Nature Biotechnology***, 2024

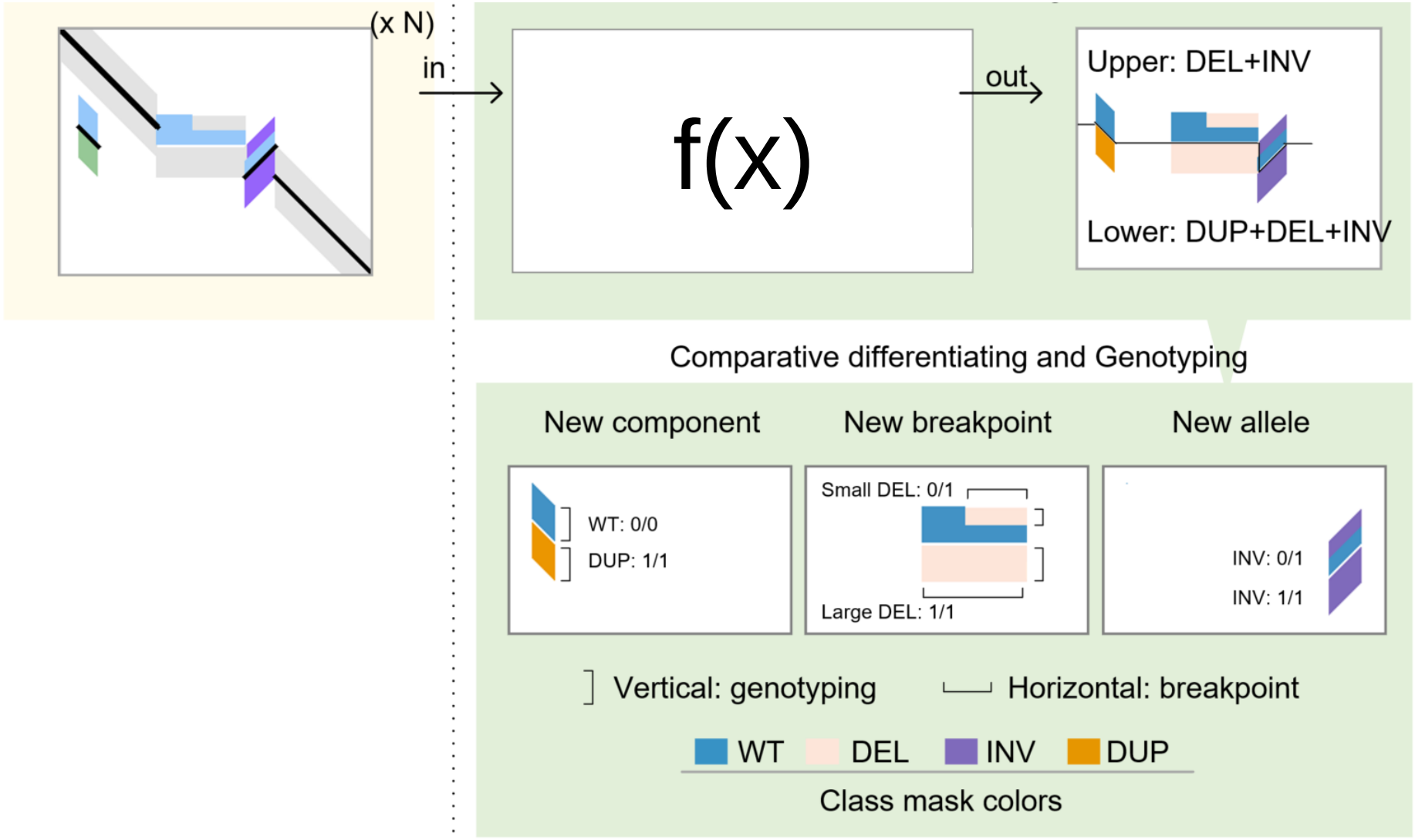
Comparative recognition module





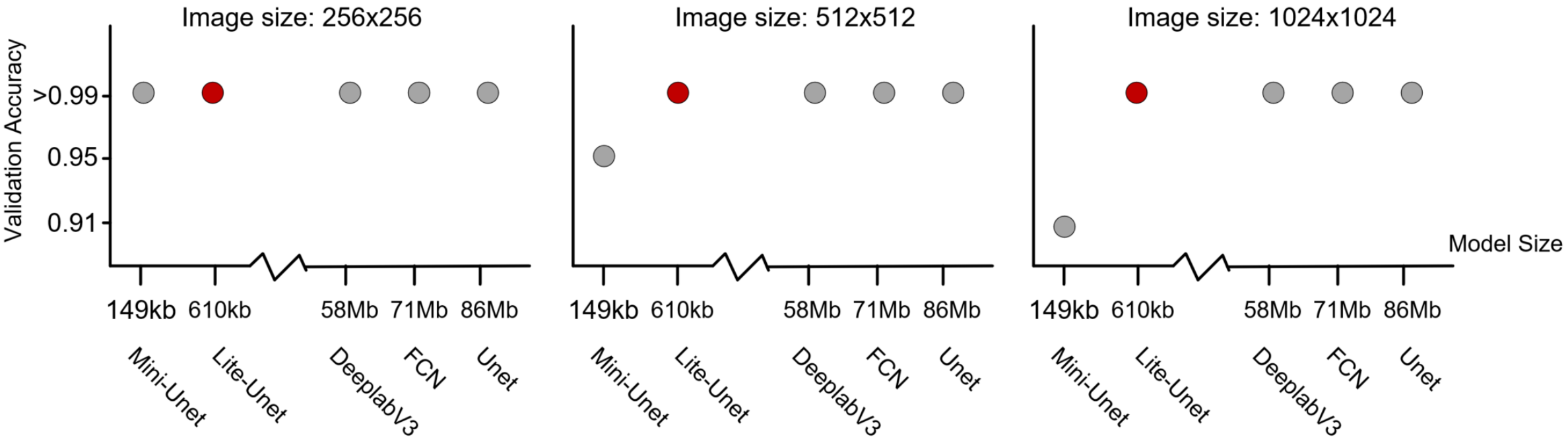
Wang et al., ***Nature Biotechnology***, 2024

Comparative recognition module



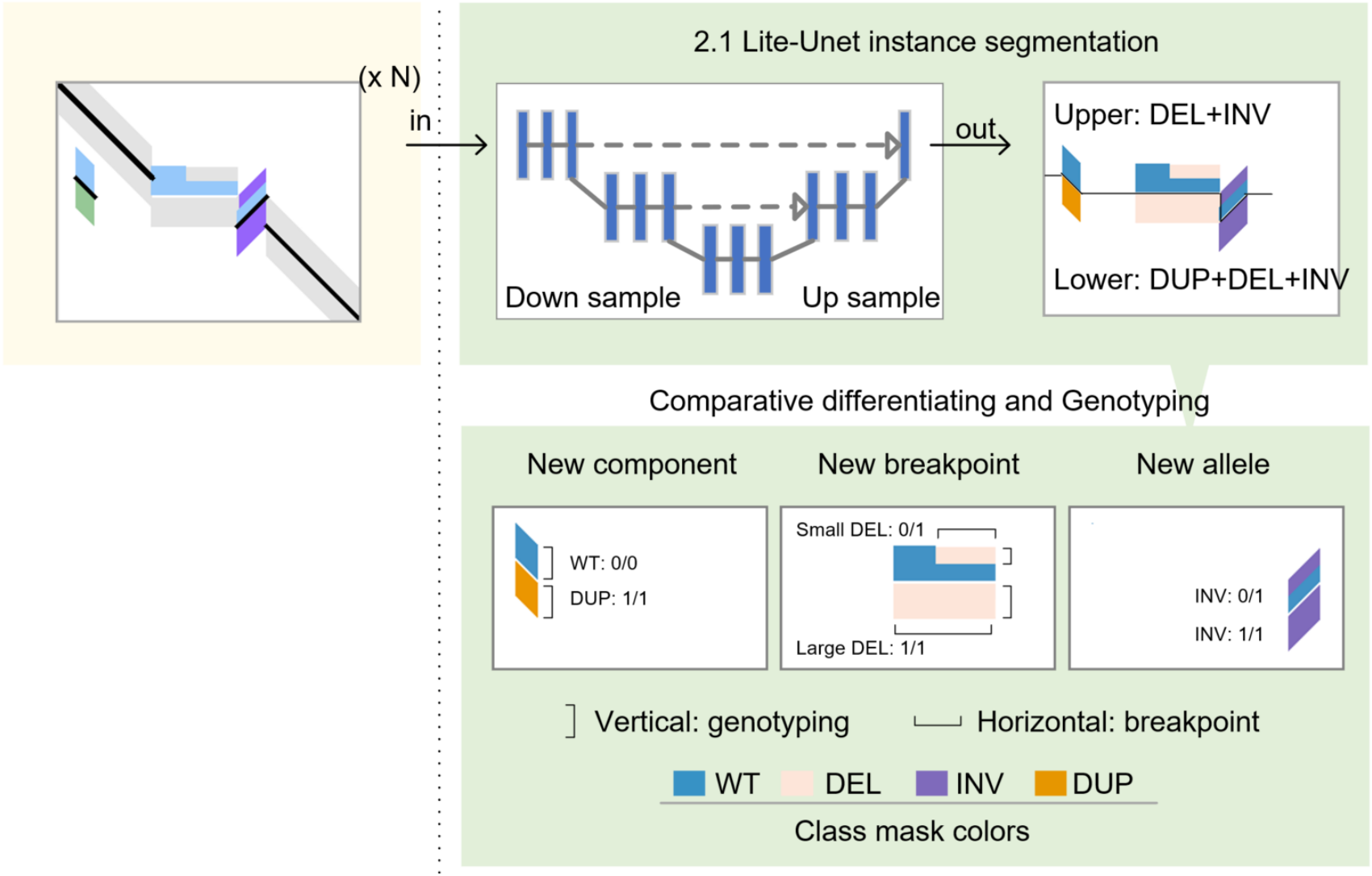
Wang et al., ***Nature Biotechnology***, 2024

Model selection



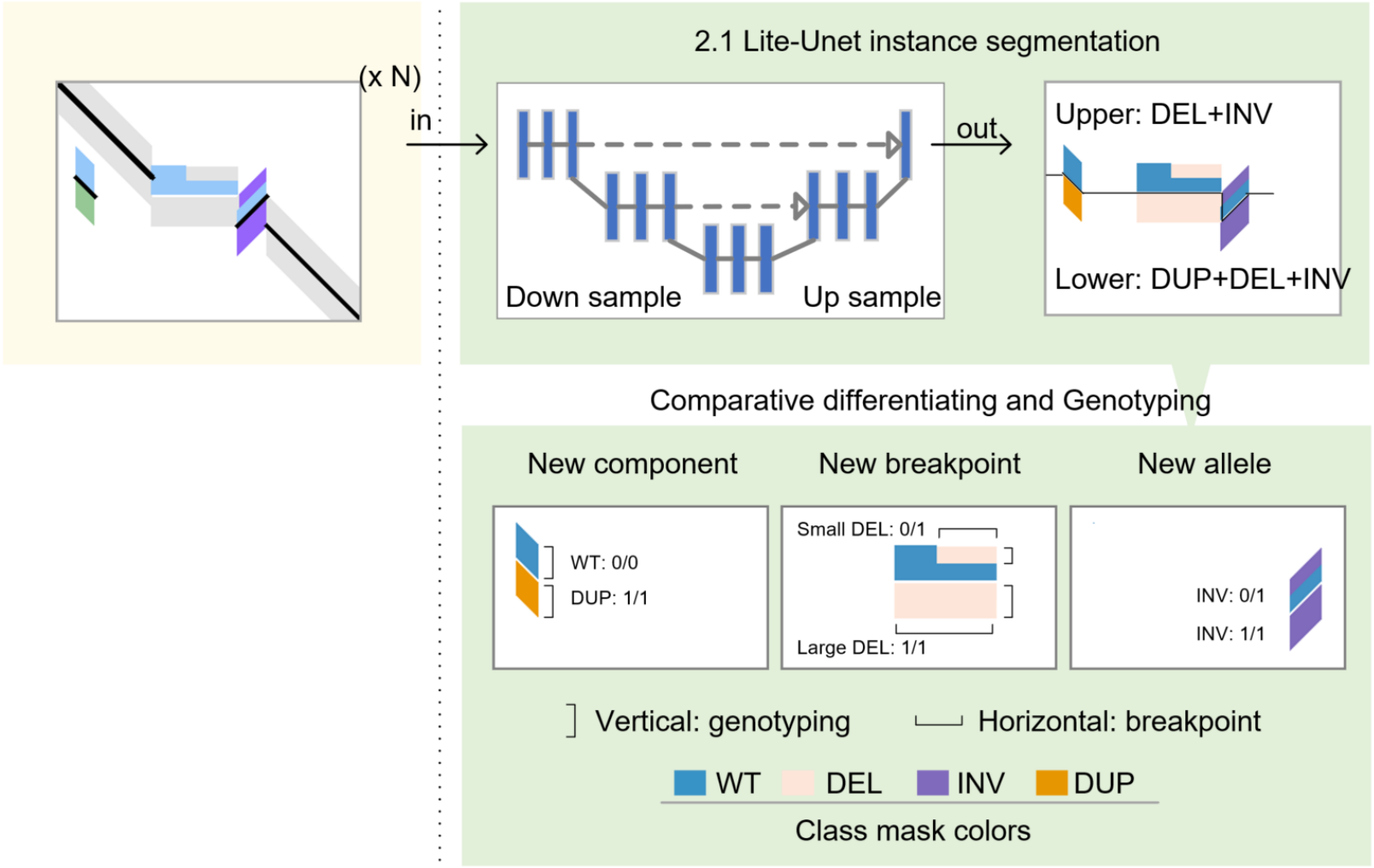
Wang et al., ***Nature Biotechnology***, 2024

Comparative recognition module



Wang et al., ***Nature Biotechnology***, 2024

Comparative recognition module



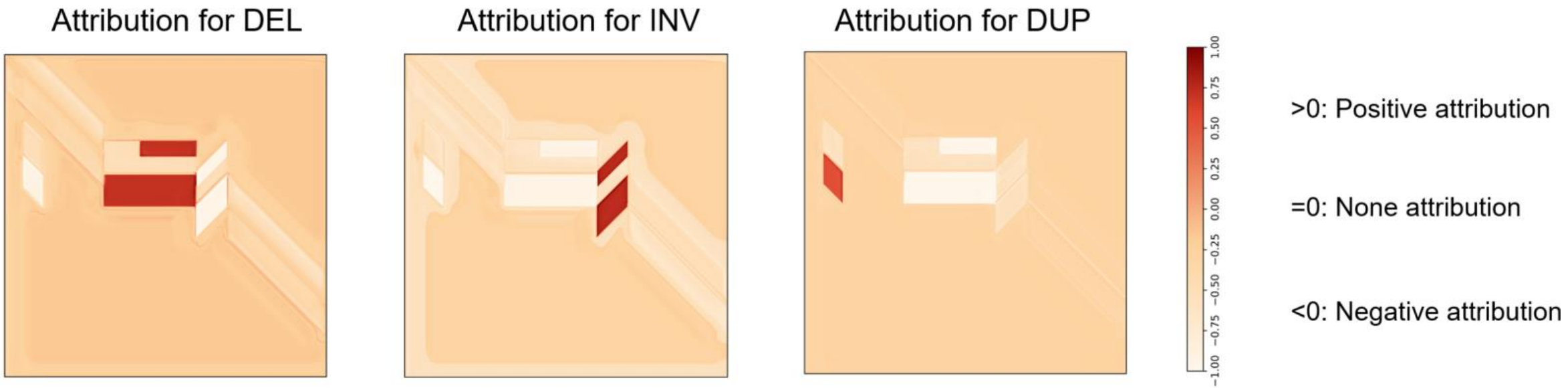
Segmentation mask assigns each pixel to a specific type

Horizontal and vertical comparison between upper panel and lower panel

Inter-genome SV differences

Wang et al., ***Nature Biotechnology***, 2024

Interpretability

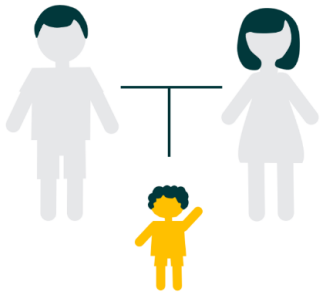


(Layer Grad-Cam)

Wang et al., ***Nature Biotechnology***, 2024

Performance 1. Multi-genome genotyping

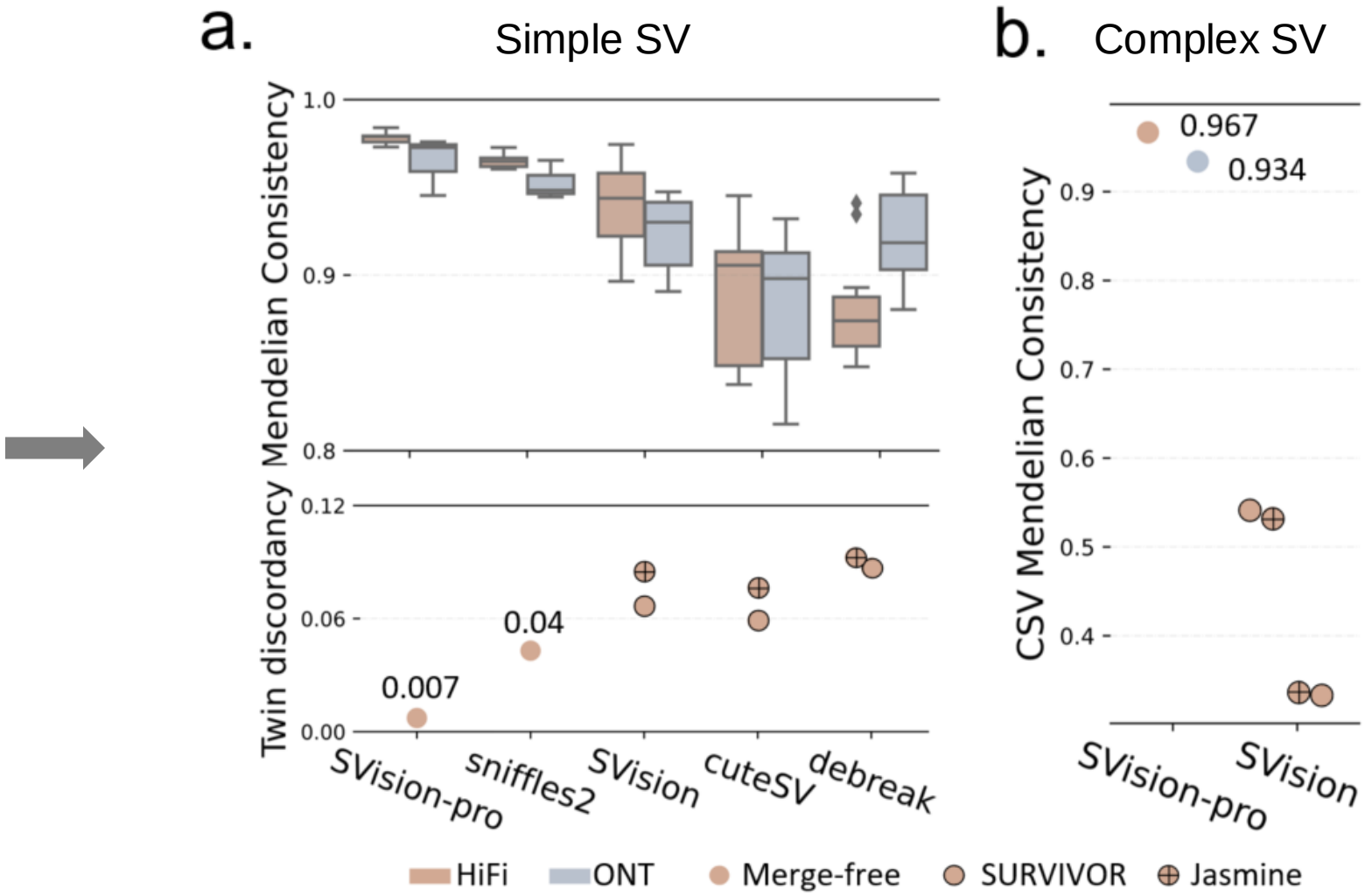
Trio datasets



X 5

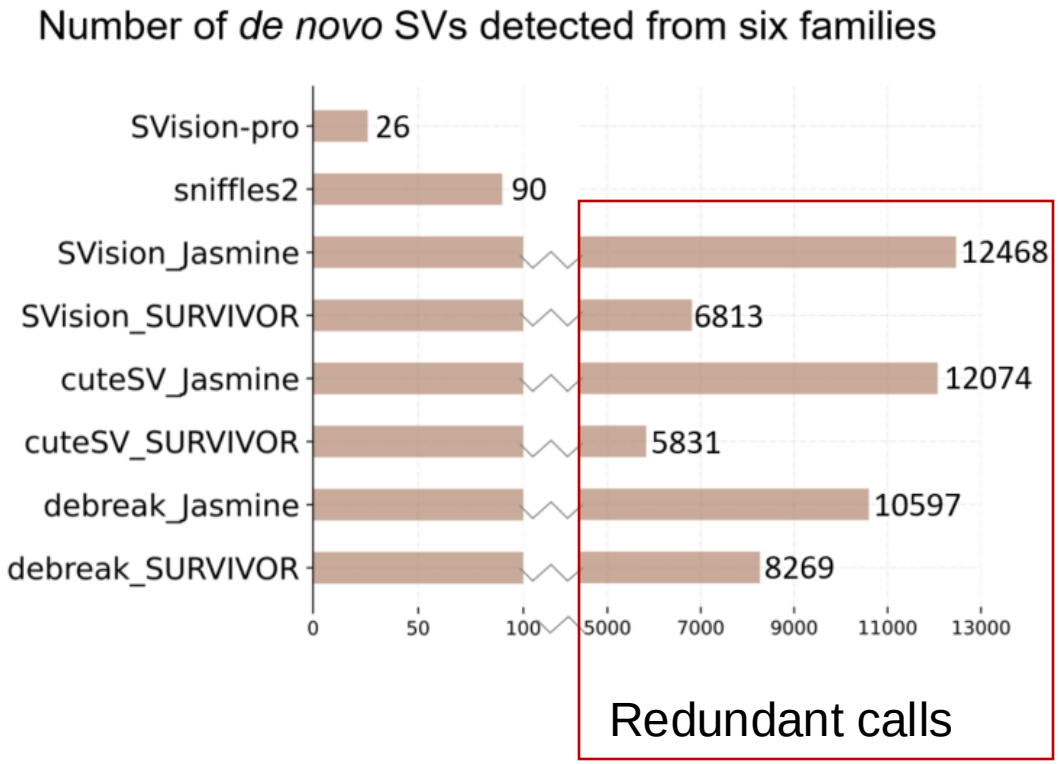
Quartet dataset

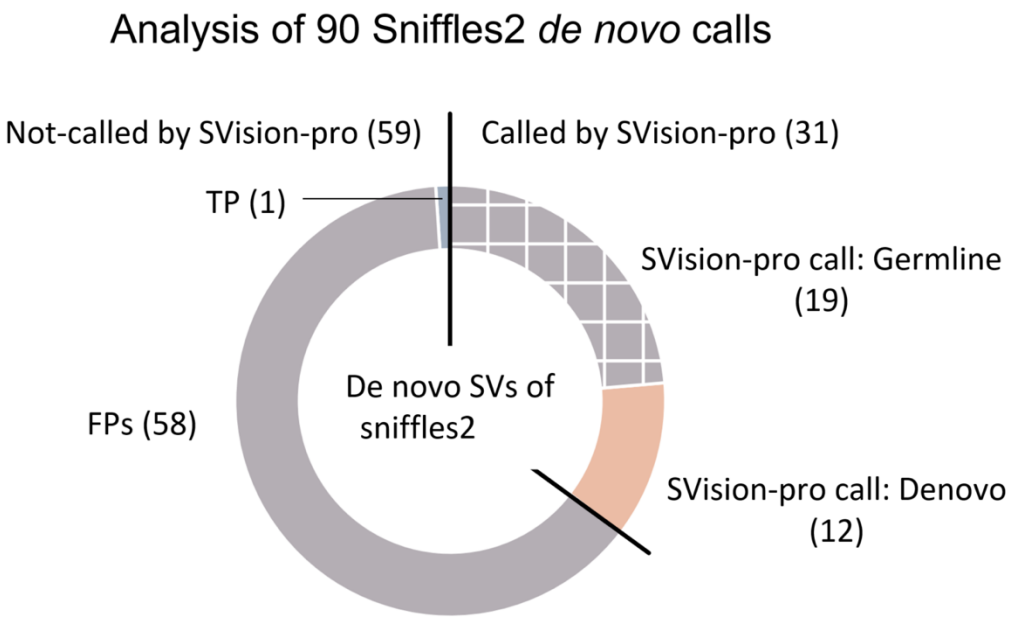
X 1



Wang et al., ***Nature Biotechnology***, 2024

Performance 2. *De novo* SV



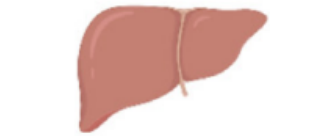


SVision-pro significantly reduces false-positive *de novo* calls

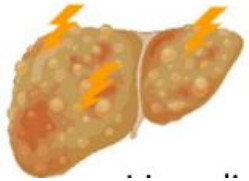
Wang et al., ***Nature Biotechnology***, 2024

Performance 3. Low-frequency somatic SV

Simulated normal-tumor pairs



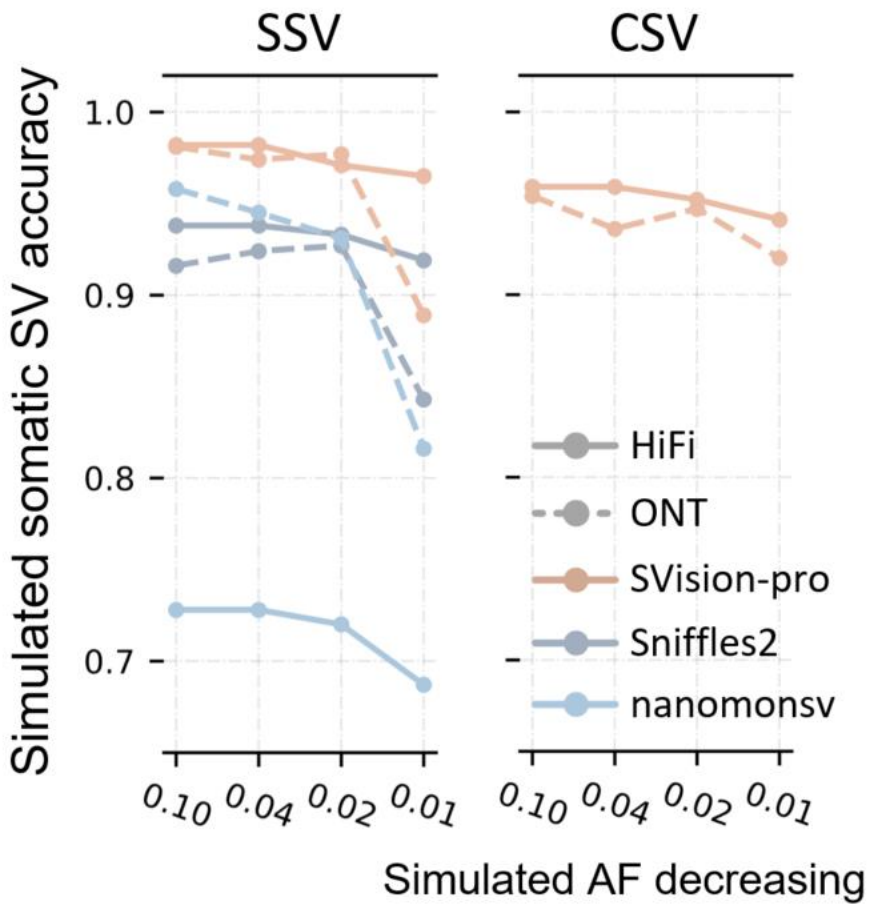
Normal genome



Tumor genome

implant

SSVs CSVs



Wang et al., ***Nature Biotechnology***, 2024

nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-024-02190-7

De novo and somatic structural variant discovery with SVision-pro

Received: 1 August 2023

Accepted: 27 February 2024

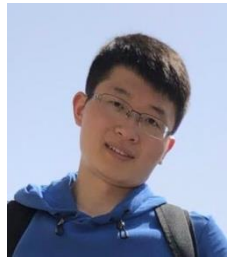
Published online: 22 March 2024

Songbo Wang[1,2,3], Jiadong Lin[2,3], Peng Jia[1,2,3], Tun Xu[2,3], Xiujuan Li[2,3], Yuezhuangnan Liu[4], Dan Xu[4], Stephen J. Bush[2,3], Deyu Meng[3,5,6,7] & Kai Ye[1,2,3,4,8,9]

Acknowledgements

国家自然科学基金委员会 NSFC
中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

西安交通大學 XI'AN JIAOTONG UNIVERSITY



Mr Songbo Wang
PhD student
Automation department

Dr Jiadong Lin
Now Postdoc @ Eichler lab

Contact: kaiye@xjtu.edu.cn