BIOINFODIAG

Réseau Français de Bioinformatique pour le Diagnostic

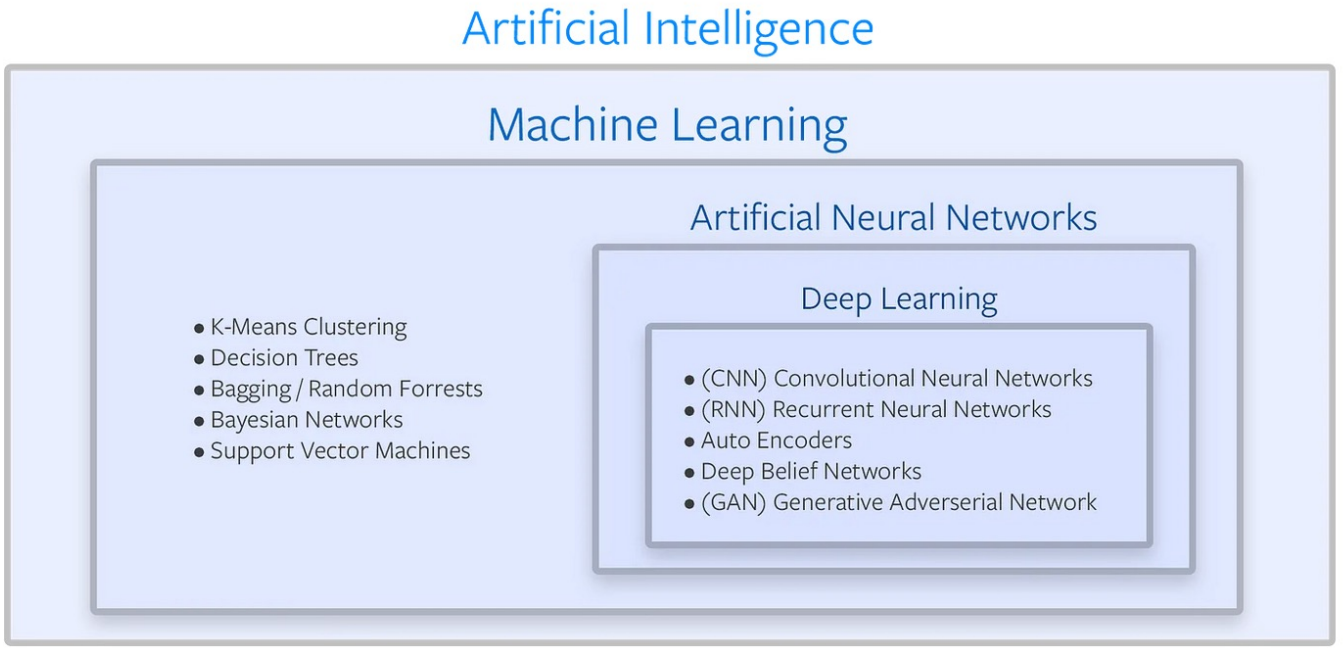
API HOUR

Le deep learning en génétique moléculaire: 3 outils qui changent la donne

David Baux, CHU de Montpellier

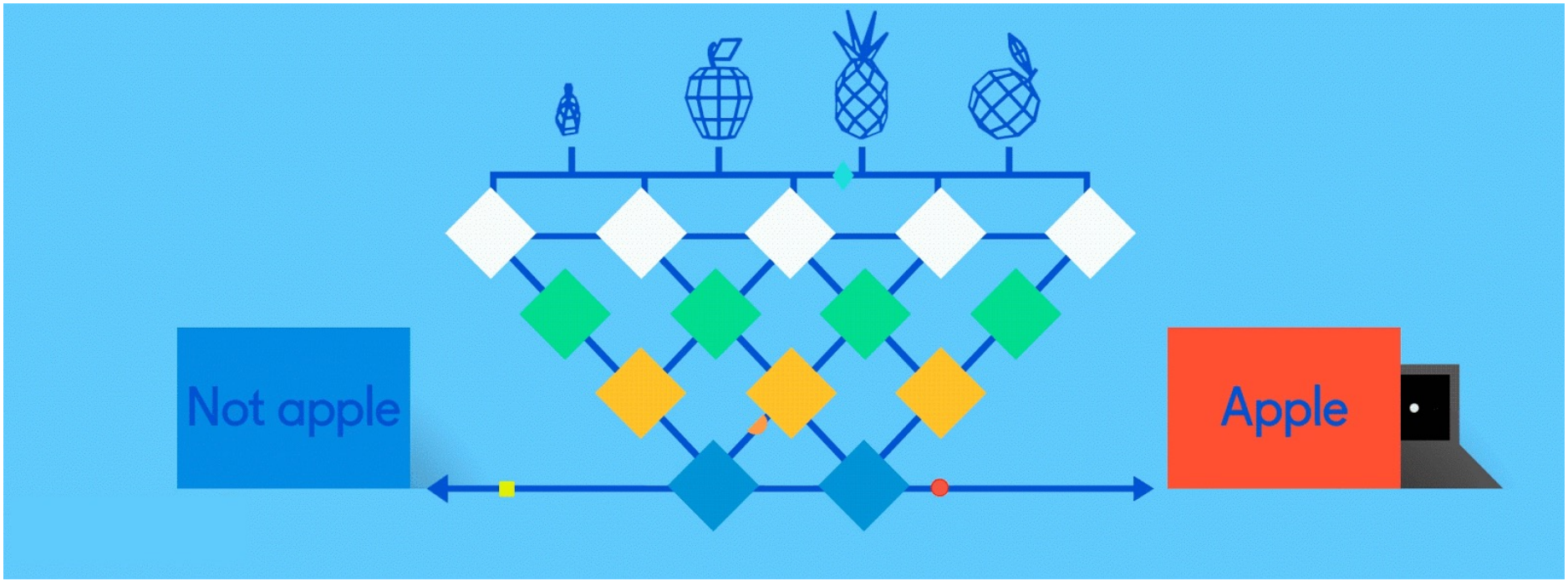
31 mai 2023

Disclaimer: je ne suis pas du tout expert en IA



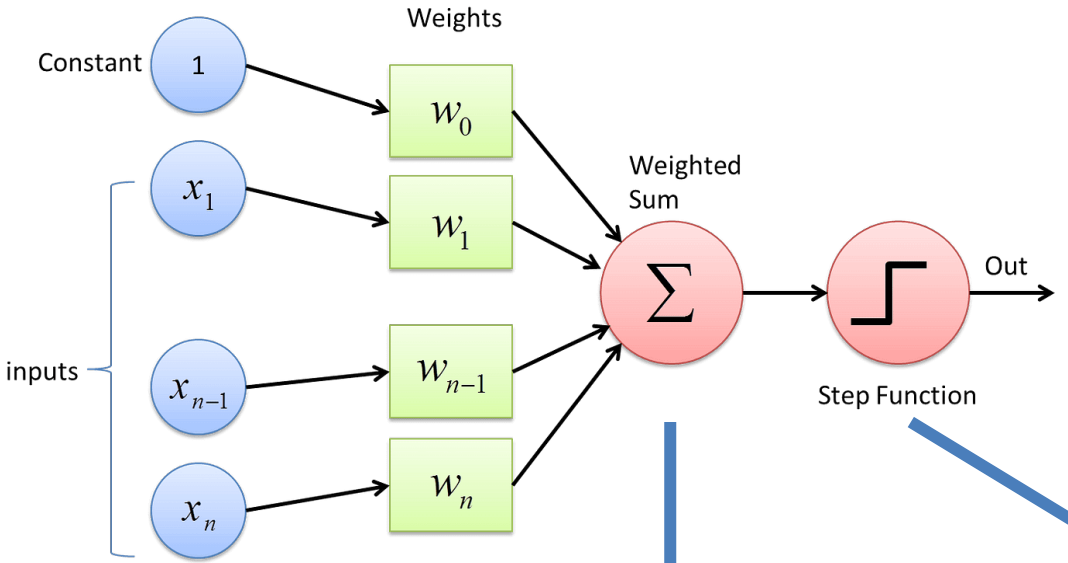
from Facebook Developer Circles Resources

https://yannicksergeobam.medium.com/comprendre-les-r%C3%A9seaux-de-neurones-convolutifs-cnn-d5f14d963714



https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53



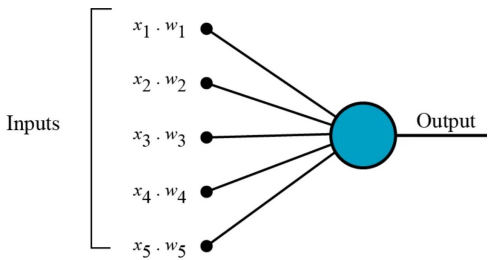


Fig: Multiplying inputs with weights for 5 inputs

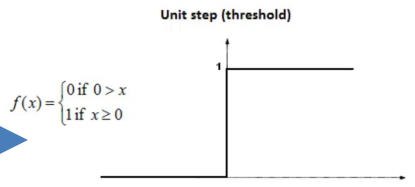
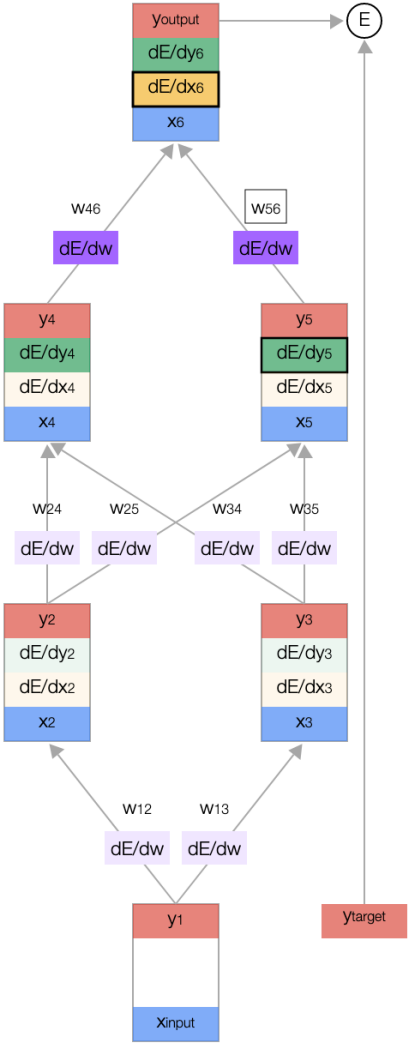
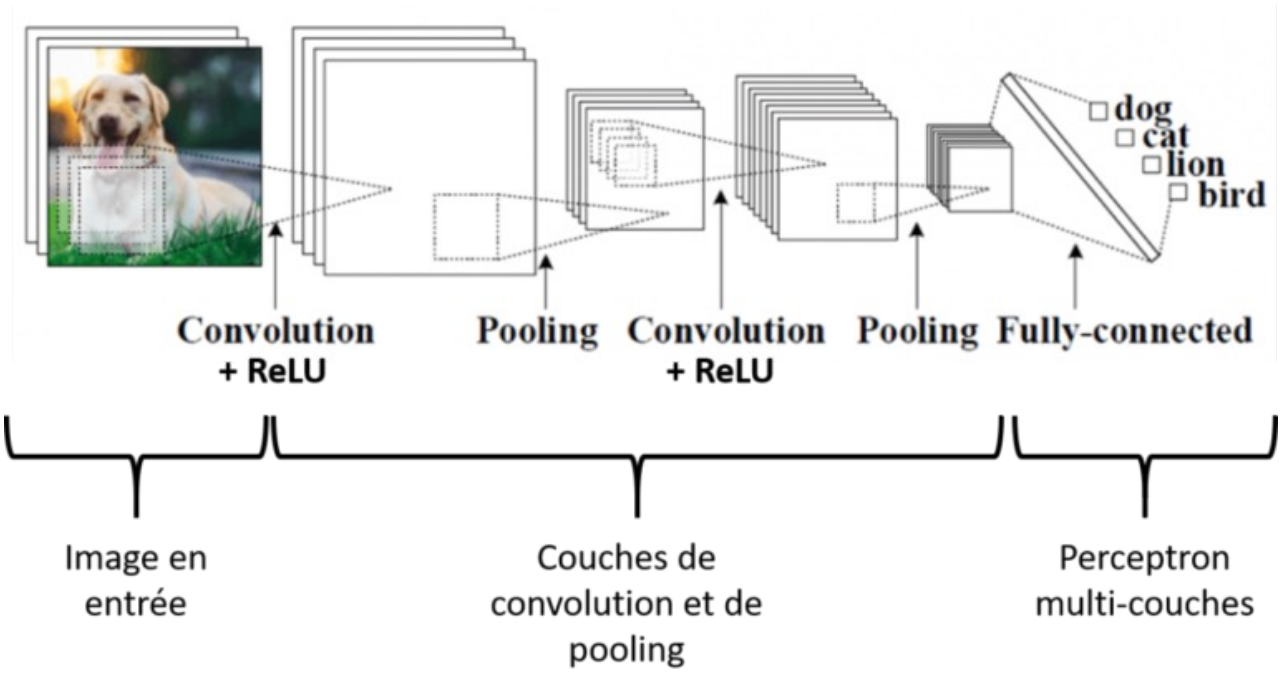


Fig: Unit Step Activation Function

Entrainement et rétropropagation de l'erreur: Mise à jour des pondérations



https://developers-dot-devsite-v2-prod.appspot.com/machine-learning/crash-course/b



https://yannicksergeobam.medium.com/comprendre-les-r%C3%A9seaux-de-neurones-convolutifs-cnn-d5f14d963714

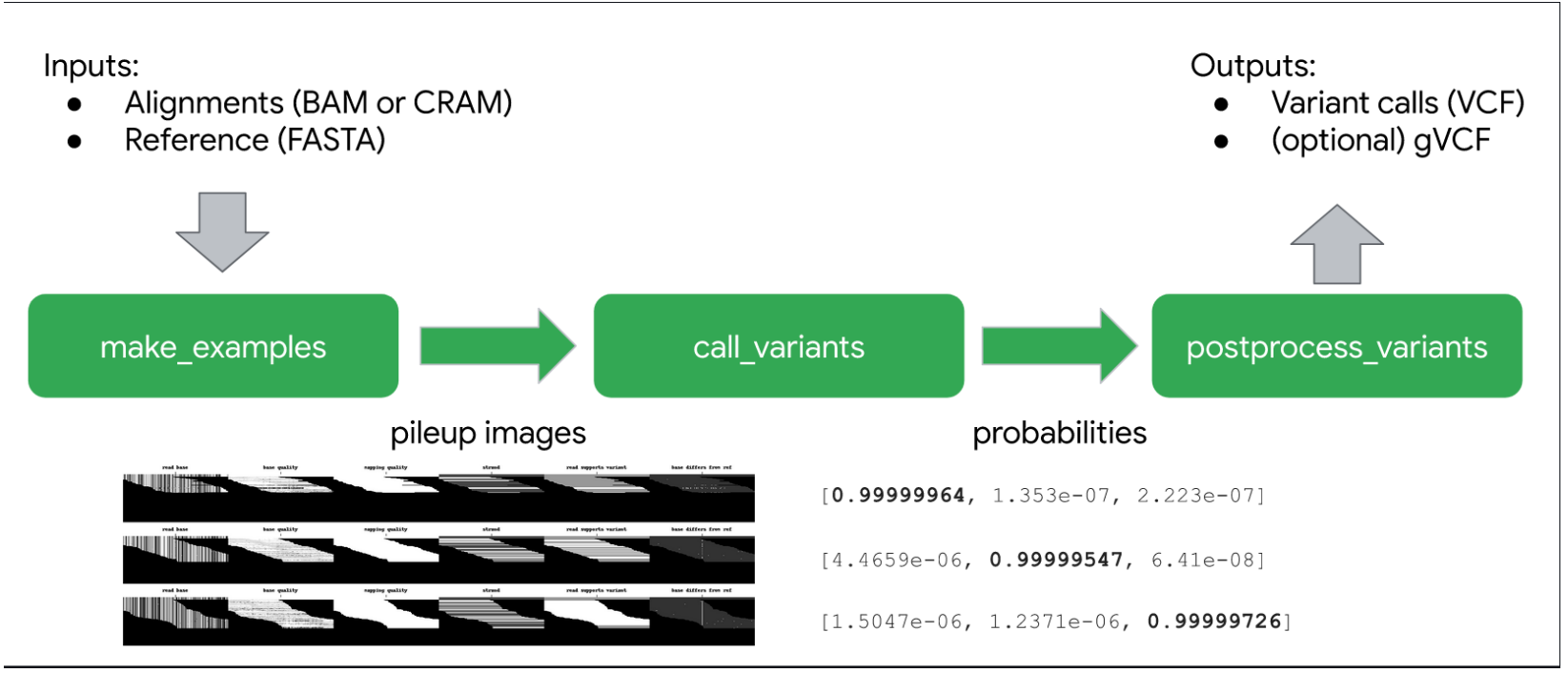
LETTERS

nature biotechnology

A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin[1,2], Pi-Chuan Chang[2], David Alexander[2], Scott Schwartz[2], Thomas Colthurst[2], Alexander Ku[2], Dan Newburger[1], Jojo Dijamco[1], Nam Nguyen[1], Pegah T Afshar[1], Sam S Gross[1], Lizzie Dorfman[1,2], Cory Y McLean[1,2] & Mark A DePristo[1,2]

Poplin et al., 2018
https://doi.org/10.1038/nbt.4235



https://github.com/google/deepvariant

Channels are shown in greyscale below in the following order:

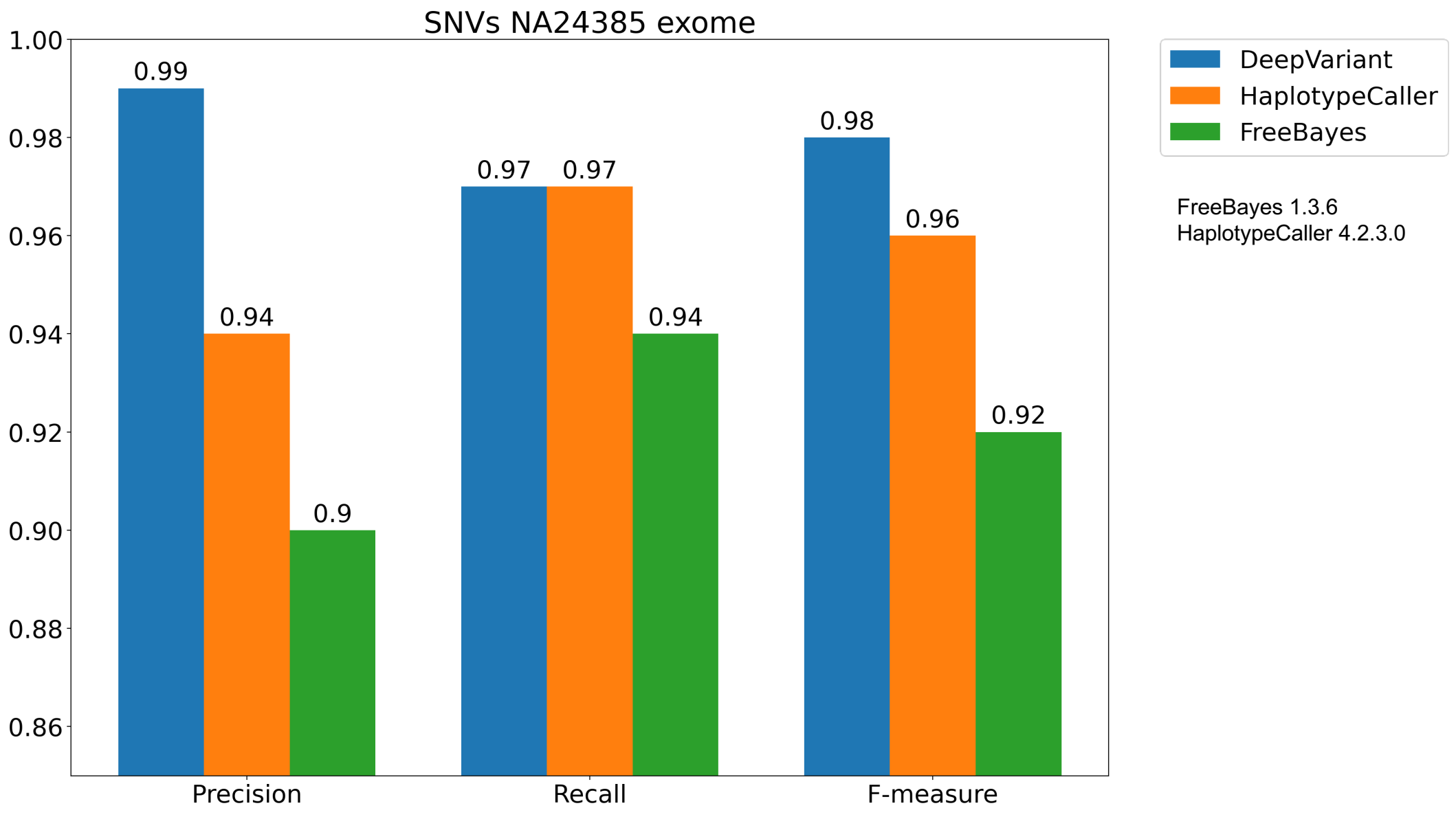
1. Read base: different intensities represent A, C, G, and T.
2. Base quality: set by the sequencing machine. White is higher quality.
3. Mapping quality: set by the aligner. White is higher quality.
4. Strand of alignment: Black is forward; white is reverse.
5. Read supports variant: White means the read supports the given alternate allele, grey means it does not.
6. Base differs from ref: White means the base is different from the reference, dark grey means the base matches the reference.

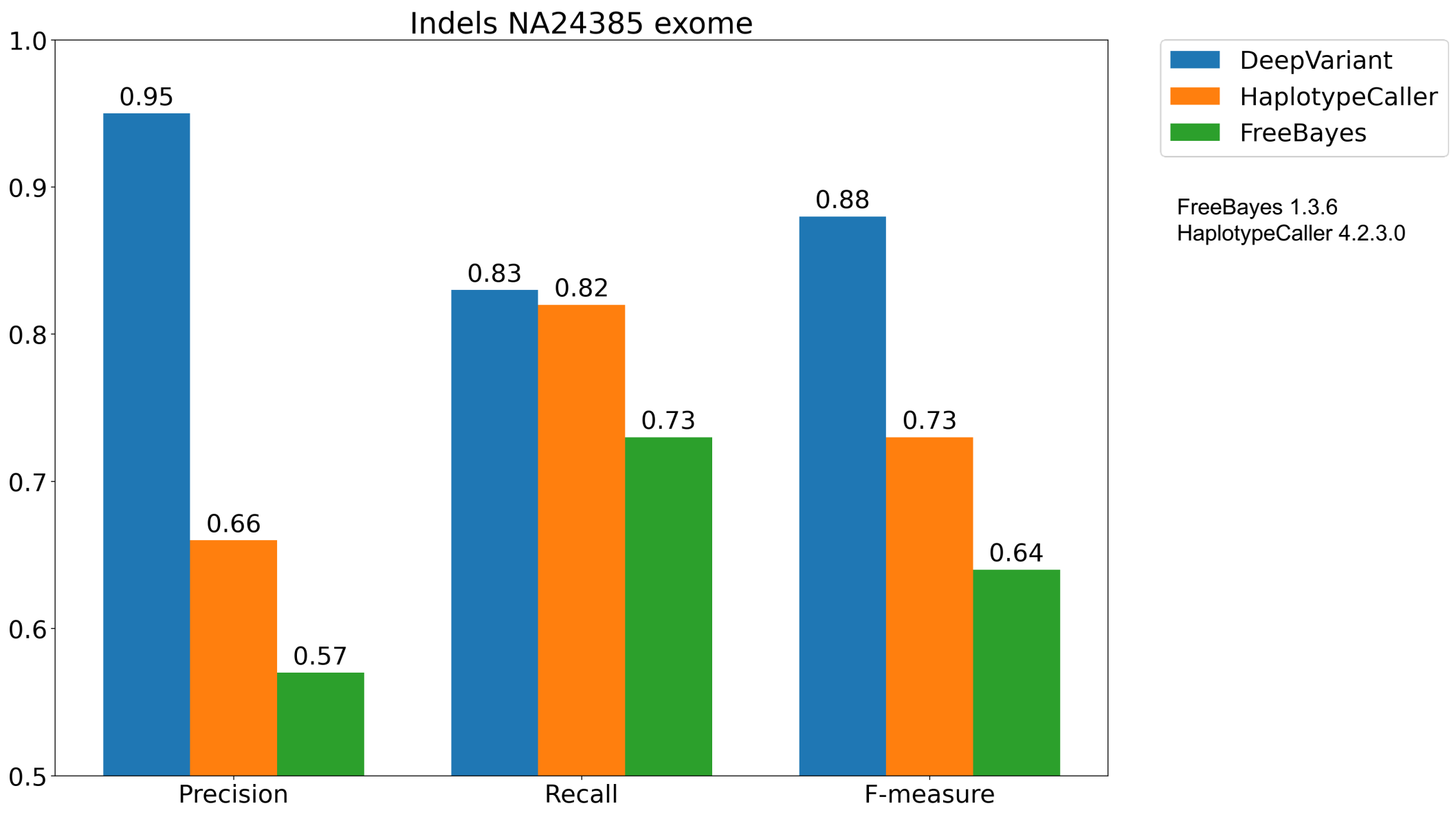
https://google.github.io/deepvariant/posts/2020-02-20-looking-through-deepvariants-eyes/

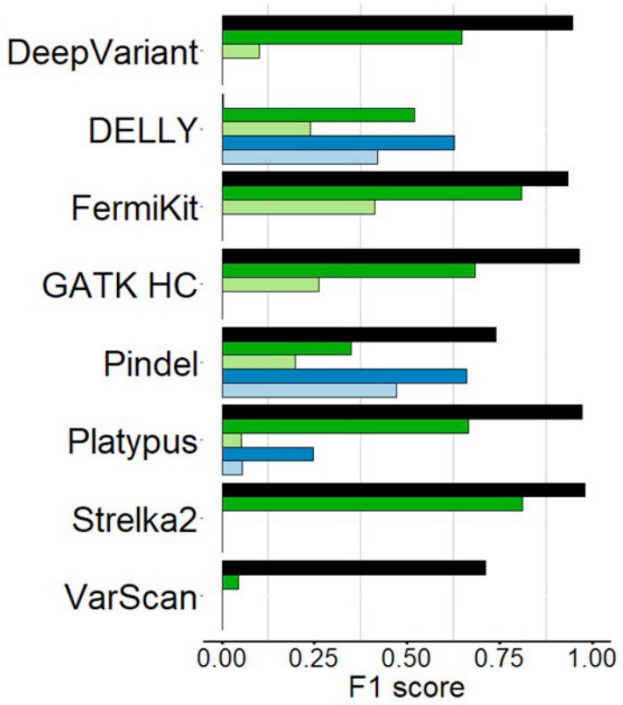
Deepvariant 1.5 VS NA24385

Type	Filter	TRUTH.TOTAL	TRUTH.TP	TRUTH.FN	QUERY.FP	METRIC.Recall	METRIC.Precision	METRIC.F1_Score
INDEL	ALL	1671	1405	266	76	0.840814	0.948959	0.891619
INDEL	PASS	1671	1405	266	76	0.840814	0.948959	0.891619
SNP	ALL	27362	26606	756	249	0.972370	0.990729	0.981464
SNP	PASS	27362	26606	756	249	0.972370	0.990729	0.981464

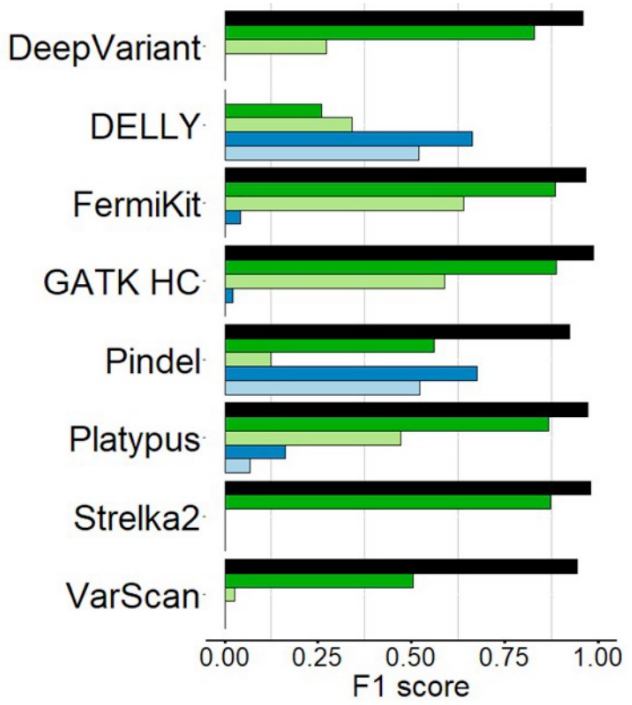
~186000 régions
38,8Mb
WES 60X


SNVs NA24385 exome
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.99
0.94
0.9
0.97
0.97
0.94
0.98
0.96
0.92
Precision
Recall
F-measure
DeepVariant
HaplotypeCaller
FreeBayes
FreeBayes 1.3.6
HaplotypeCaller 4.2.3.0


Indels NA24385 exome
1.0
0.9
0.8
0.7
0.6
0.5
0.95
0.66
0.57
0.83
0.82
0.73
0.88
0.73
0.64
Precision
Recall
F-measure
DeepVariant
HaplotypeCaller
FreeBayes
FreeBayes 1.3.6
HaplotypeCaller 4.2.3.0



30X; 100bp



30X; 250bp

Deletions size 1 - 20 bp 20 - 50 bp 50 - 200 bp 200 - 500 bp ≥ 500 bp

Semi-simulated dataset: 43,066 insertions and 45,223 deletions

Wang et al., 2022
https://doi.org/10.1371/journal.pcbi.1009269

Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

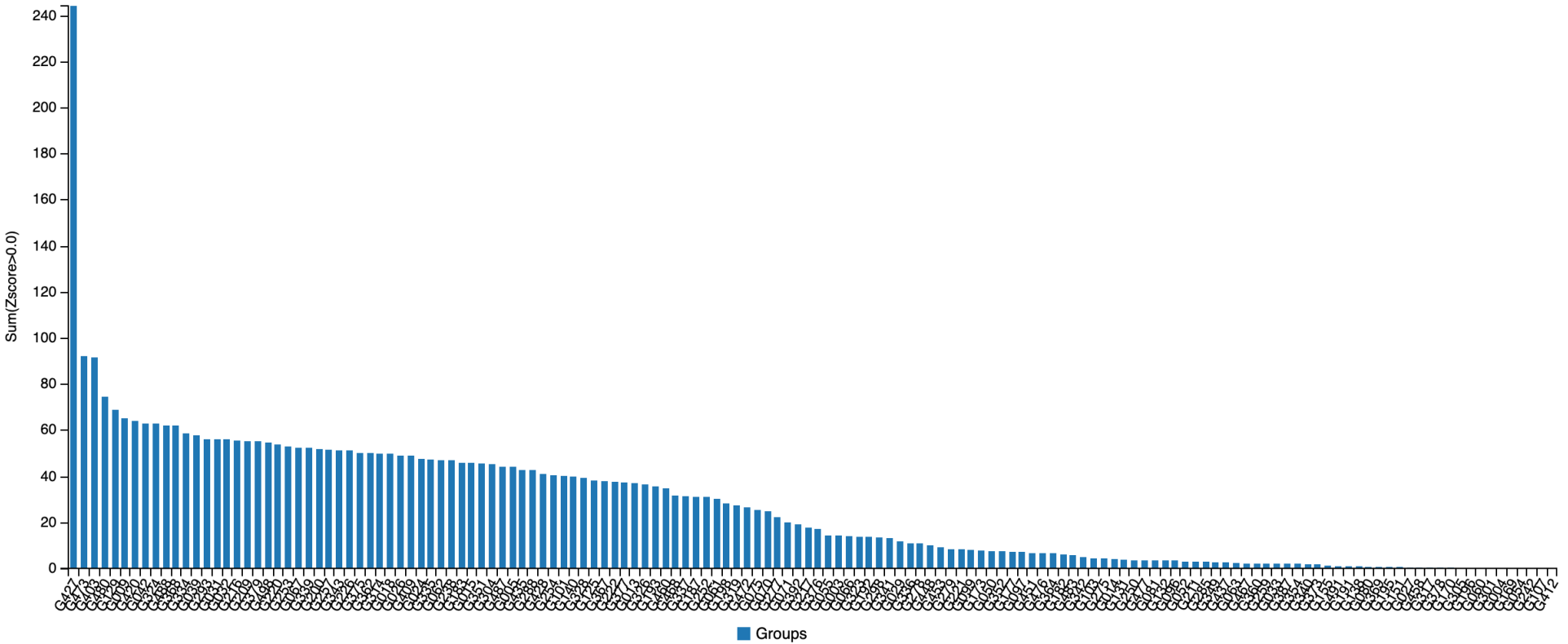
Published online: 15 July 2021

Open access

Check for updates

John Jumper[1,4]✉, 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 Žídek[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], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

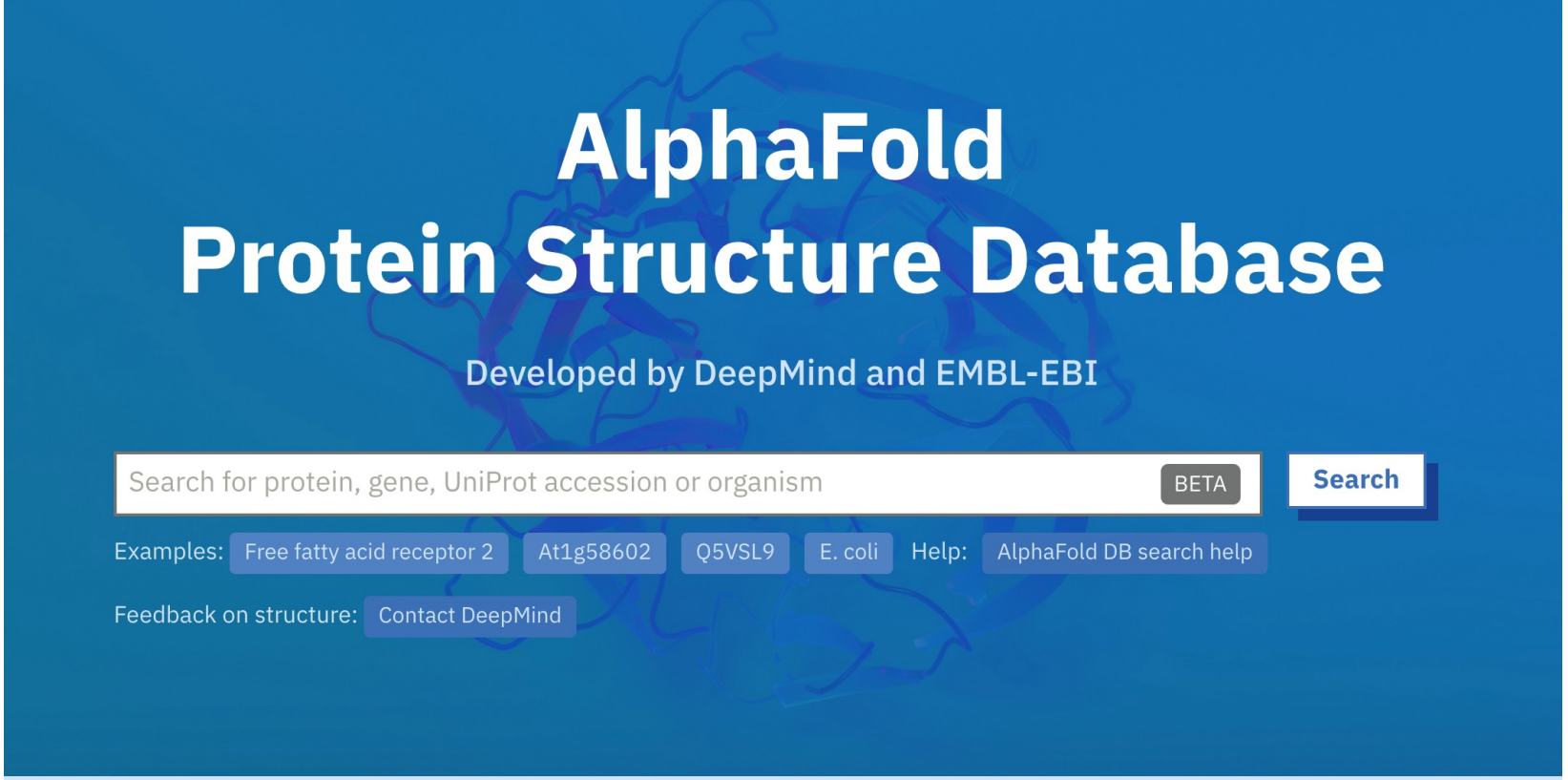
Jumper et al., 2021
https://doi.org/10.1038/s41586-021-03819-2



14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

TS Analysis : Group performance based on combined z-scores

https://predictioncenter.org/casp14/zscores_final.cgi



AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

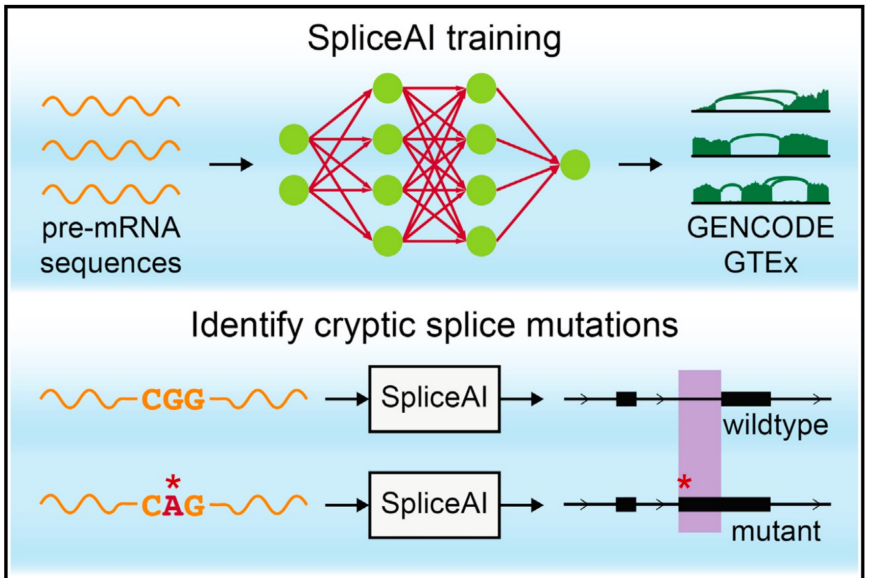
https://alphafold.com/

Article

Cell

Predicting Splicing from Primary Sequence with Deep Learning

Graphical Abstract



Authors

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, ..., Serafim Batzoglou, Stephan J. Sanders, Kyle Kai-How Farh

Correspondence

kfarh@illumina.com

In Brief

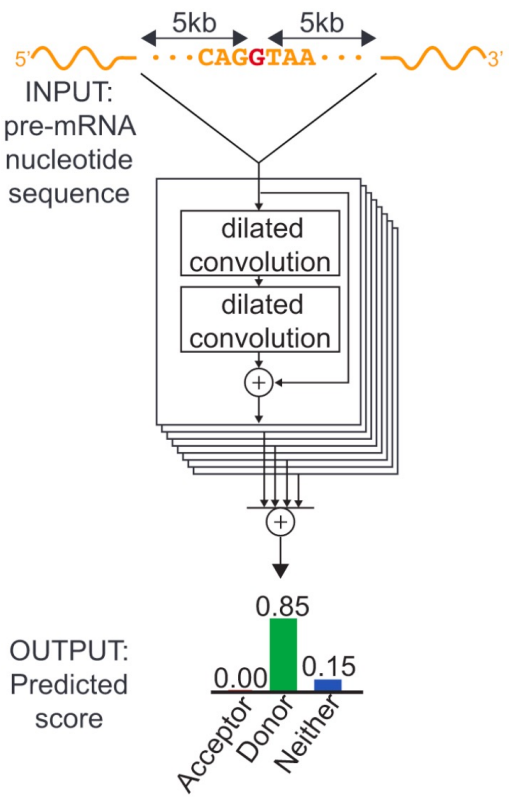
A deep neural network precisely models mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases.

Jaganathan et al., 2019
https://doi.org/10.1016/j.cell.2018.12.015

Operation mode - performance

Jaganathan et al., 2019
https://doi.org/10.1016/j.cell.2018.12.015

A



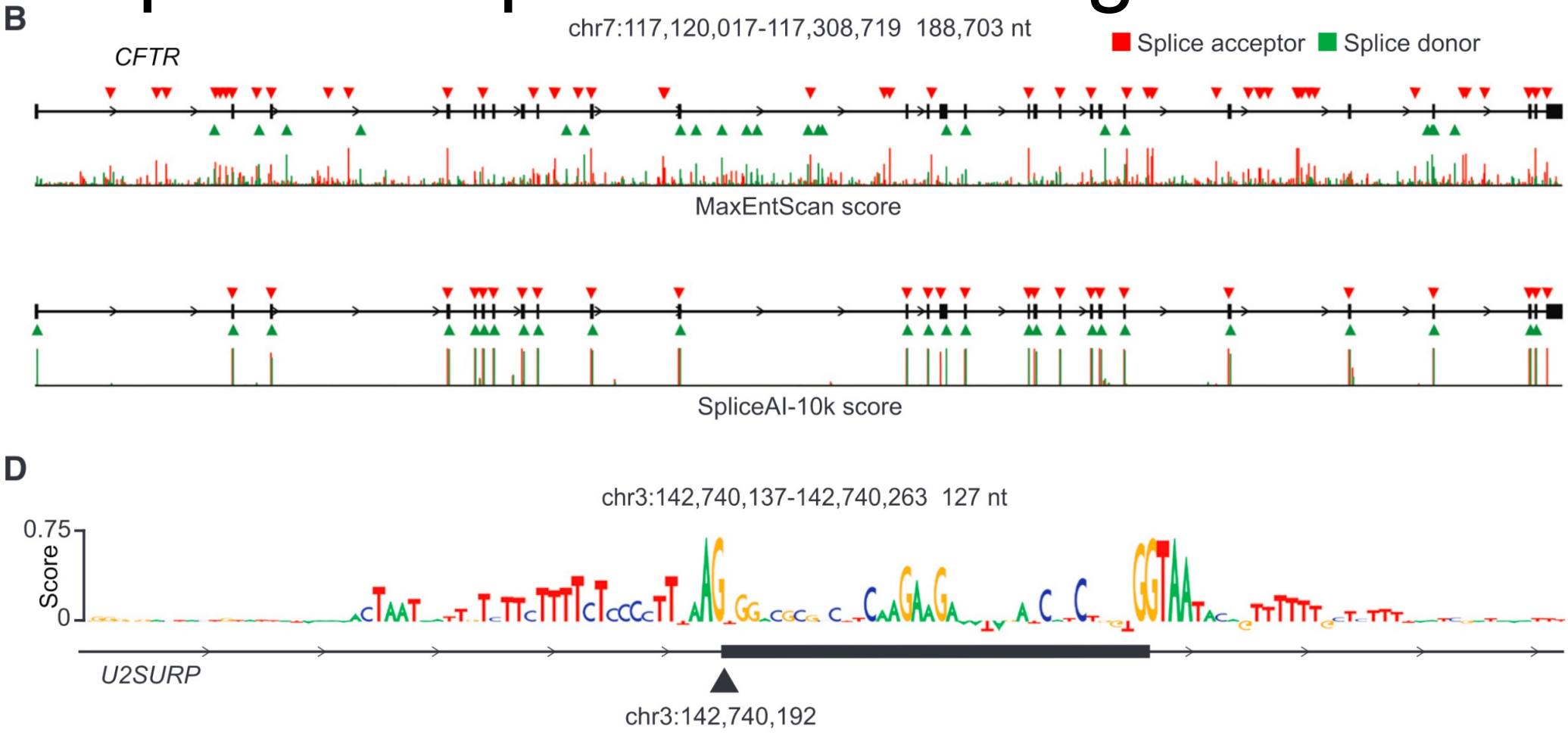
E

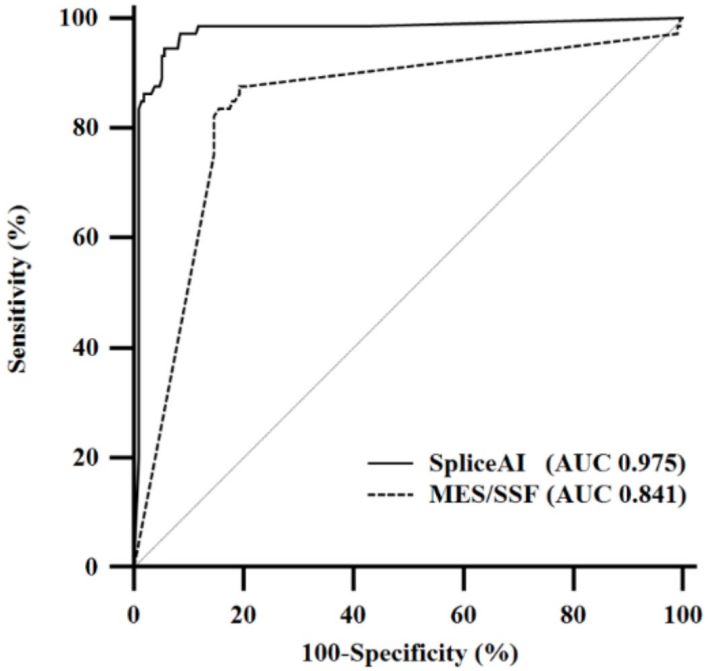
	Top-*k* accuracy	PR-AUC
SpliceAI-80nt	0.57	0.60
SpliceAI-400nt	0.90	0.95
SpliceAI-2k	0.93	0.97
SpliceAI-10k	**0.95**	**0.98**
GeneSplicer	0.30	0.23
MaxEntScan	0.22	0.15
NNSplice	0.22	0.15

Top-k accuracy is the fraction of correctly predicted splice sites at the threshold where the number of predicted sites is equal to the actual number of sites present.

Jaganathan et al., 2019
https://doi.org/10.1016/j.cell.2018.12.015

SpliceAI: splice sites recognition

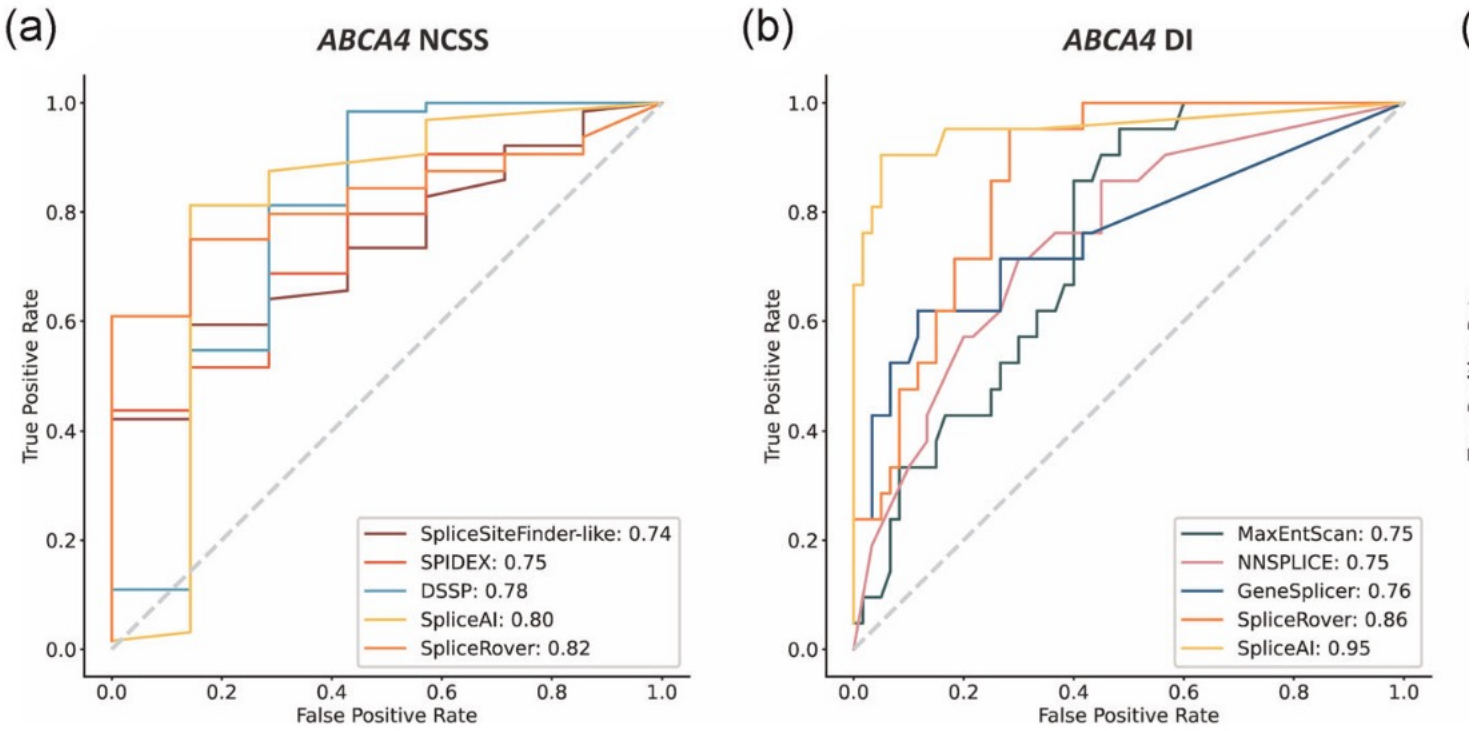




Method	Sensitivity	Specificity
	N/Total N% (95% CI)	N/Total N% (95% CI)
SpliceAI	69/73	200/212
	94.5% (86.6–98.5%)	94.3% (90.3–97.0%)

Performance Evaluation of SpliceAI for the Prediction of Splicing of *NF1* Variants (cut-off: 0.22, 285 variants, 73 avec effet, 212 sans effet, séquençage du cDNA)

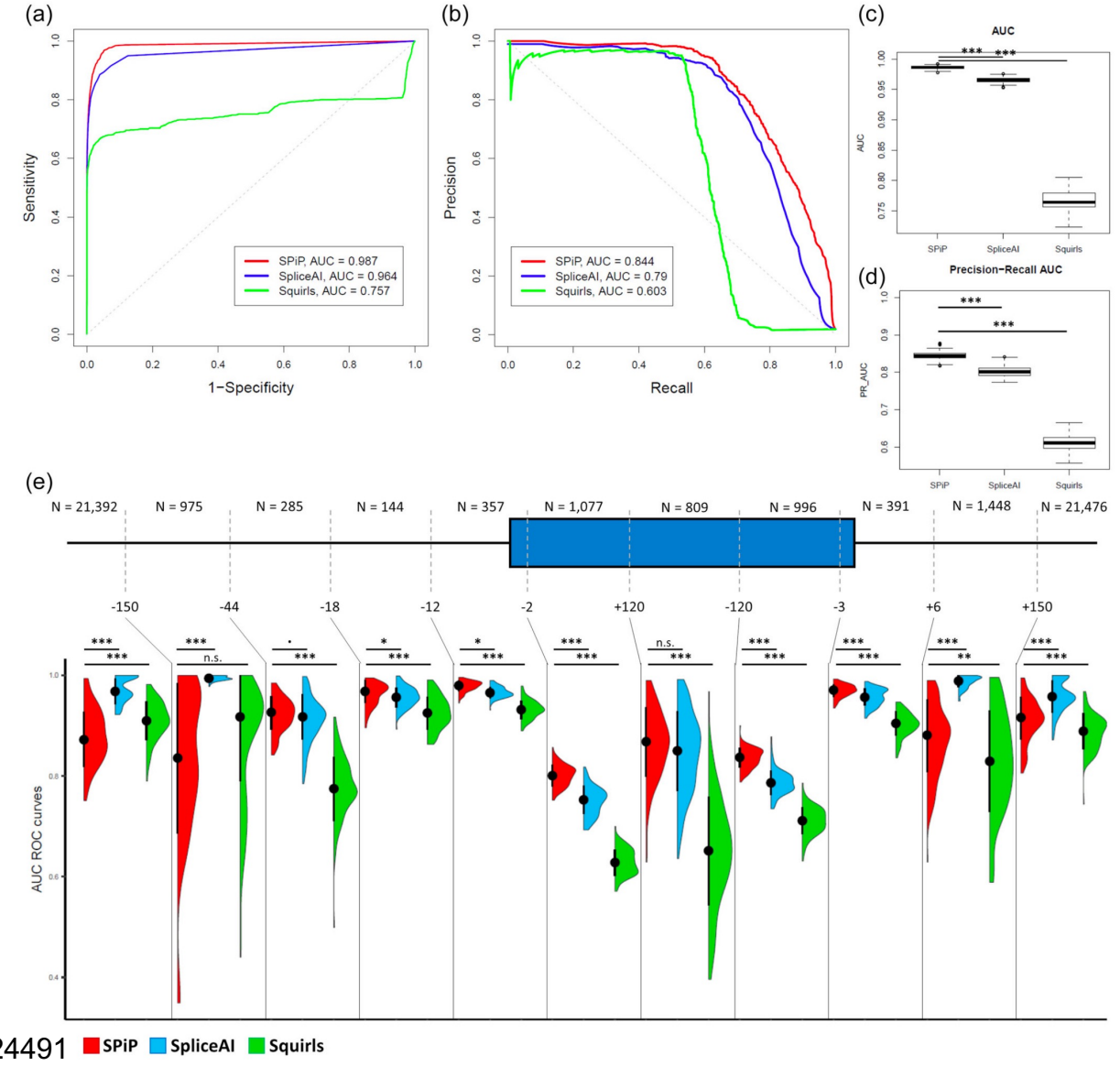
Ha et al., 2021
https://doi.org/10.3390/genes12091308



Riepe et al., 2021
https://doi.org/10.1002/humu.24212

Tool	*ABCA4* NCSS	*ABCA4* DI	*MYBPC3* NCSS	Suggested threshold
CADD	2.66	0.24	2.09	5′ extended: 7.39, 3′ intronic: 0.0964, exonic: 0.39
DSSP	0.01	0.13	0.01	0.30
GeneSplicer	0.18	0.05	0.21	–
MaxEntScan	0.26	0.31	0.24	0.10
MMSplice	1.42	–	1.37	2
NNSPLICE	0.13	0.40	0.30	0.05
Spidex	0.86	–	1.72	5
SpliceAI	0.19	0.18	0.11	0.20
SpliceRover	0.18	0.26	0.10	–
SpliceSiteFinder-like	0.01	0.12	0.09	0.05

Riepe et al., 2021
https://doi.org/10.1002/humu.24212



4616 variants
(1924 impacting)
RNA in vitro studies

Leman et al., 2022
https://doi.org/10.1002/humu.24491

Take home message

MONTPELLIER
CHU
CENTRE HOSPITALIER UNIVERSITAIRE

MoBiDiC

People

- Olivier Ardouin
- Corinne Bareil
- David Baux
- Thomas Guignard
- Simon Cabello
- Souphatta Sasorith
- Charles Van Goethem

Laboratoire de génétique moléculaire: groupe neurosensoriel

- Anne-Françoise Roux
- Luke Mansard
- Christel Vaché
- Valérie faugère
- Julie Bianchi
- Corinne Baudoin