UNIVERSITÉ CÔTE D'AZUR
FRANCE 2030 Initiative d'Excellence
IRCAN
chu de Nice
INRAE
anr

ABEILLE & VIOLA: novel tools to improve the diagnosis of mitochondrial diseases using omics and multi-omics data

15/05/2024

Justine LABORY, PhD student

Medical context

Mitochondrial diseases

Medical context

Mitochondrial diseases

Rare diseases

Medical context

Less than 1
person out of
2000

Mitochondrial diseases

Rare diseases

Medical context

Less than 1
person out of
2000

30 millions of people

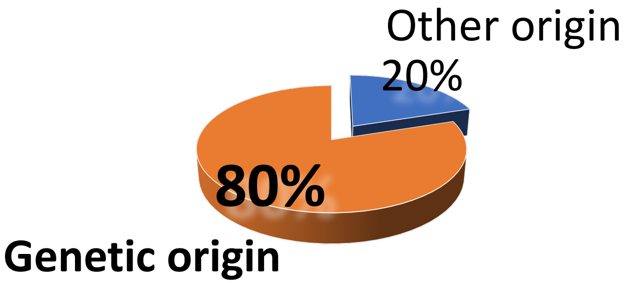
Mitochondrial diseases

Rare diseases

Medical context

Less than 1 person out of **2000**

30 millions of people



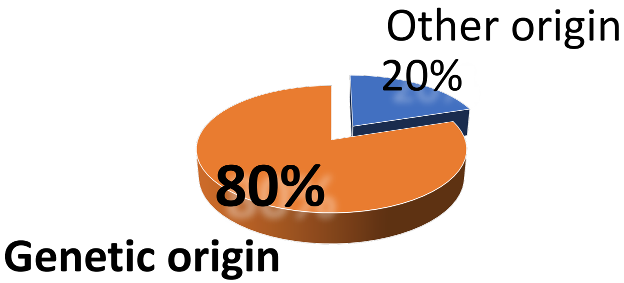
Mitochondrial diseases

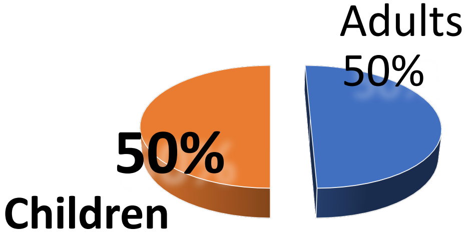
Rare diseases

Medical context

Less than 1 person out of **2000**

30 millions of people





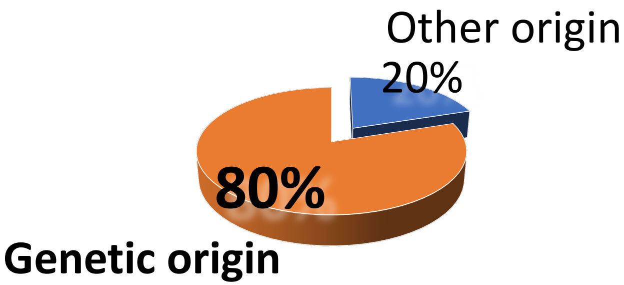
Rare diseases

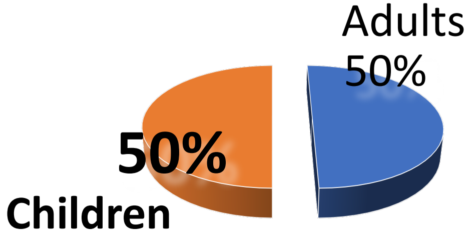
Mitochondrial diseases

Medical context

Less than 1 person out of **2000**

30 millions of people





Rare diseases

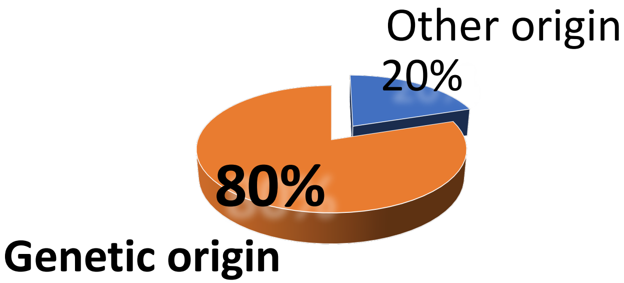
Mitochondria

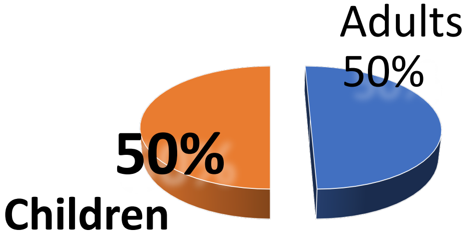
Mitochondrial diseases

Medical context

Less than 1 person out of **2000**

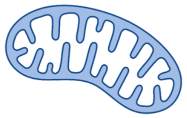
30 millions of people





Rare diseases

responsible for a wide variety of biochemical processes



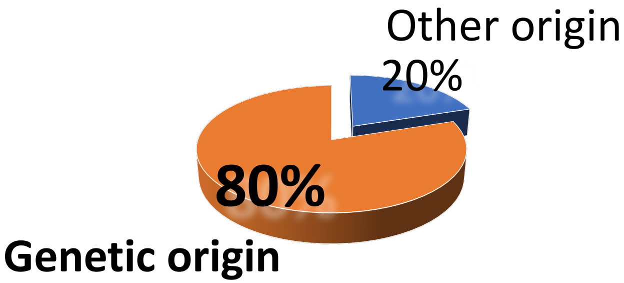
Mitochondria

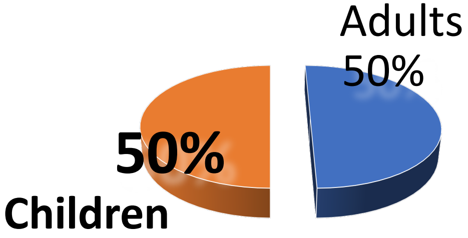
Mitochondrial diseases

Medical context

Less than 1 person out of **2000**

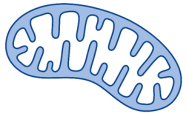
30 millions of people





Rare diseases

responsible for a wide variety of biochemical processes



Mitochondria

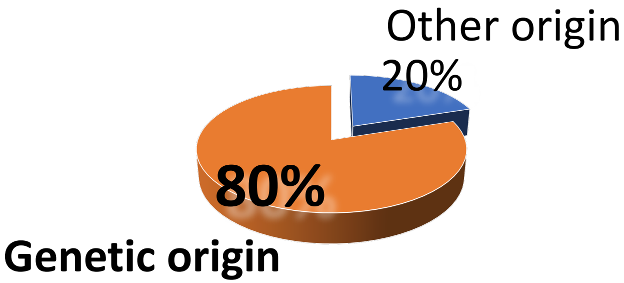
under the double control of mtDNA and nDNA

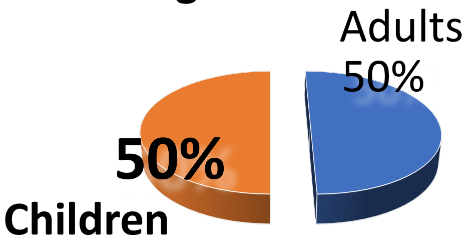
Mitochondrial diseases

Medical context

Less than 1 person out of **2000**

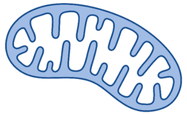
30 millions of people





Rare diseases

responsible for a wide variety of biochemical processes



Mitochondria

under the double control of mtDNA and nDNA

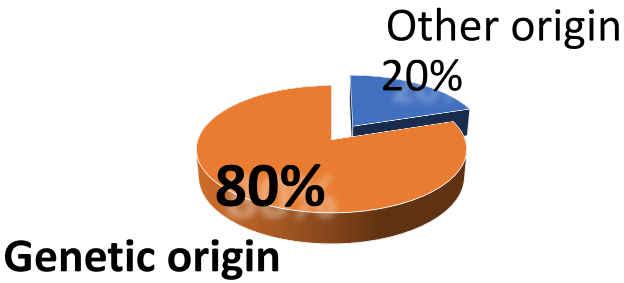
Mitochondrial diseases

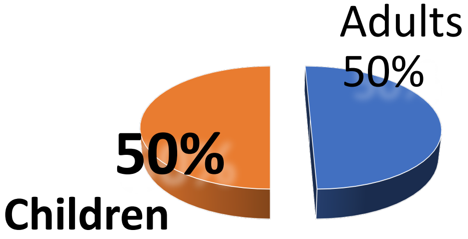
Disease

Medical context

Less than 1 person out of **2000**

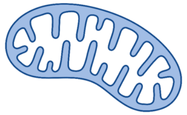
30 millions of people





Rare diseases

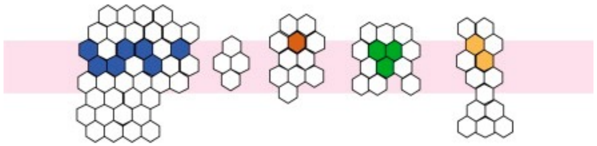
responsible for a wide variety of biochemical processes



Mitochondria

under the double control of mtDNA and nDNA

Mitochondrial diseases



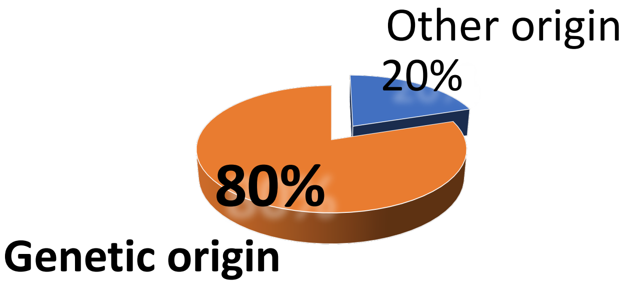
deficiency of the mitochondrial respiratory chain

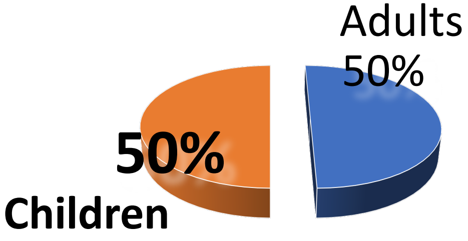
Disease

Medical context

Less than 1 person out of **2000**

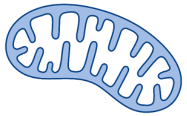
30 millions of people





Rare diseases

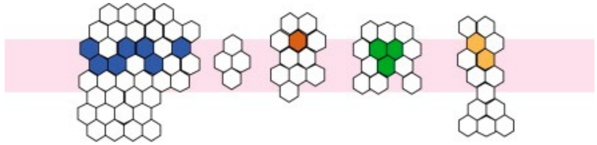
responsible for a wide variety of biochemical processes



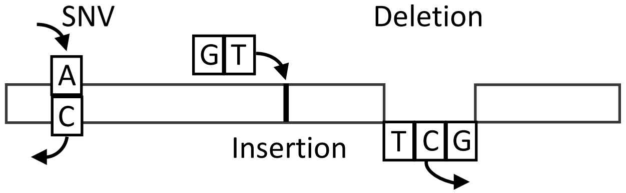
Mitochondria

under the double control of mtDNA and nDNA

Mitochondrial diseases



deficiency of the mitochondrial respiratory chain



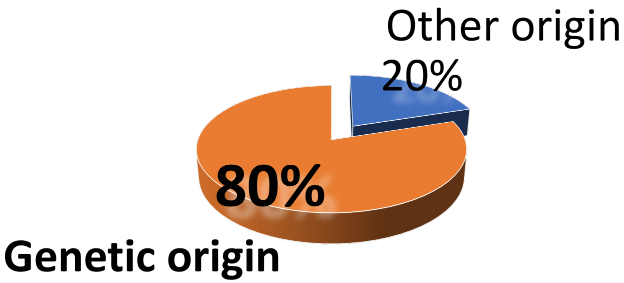
due to rare hereditary or spontaneous variants of mtDNA or nDNA

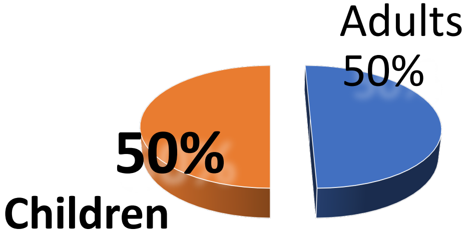
Disease

Medical context

Less than 1 person out of **2000**

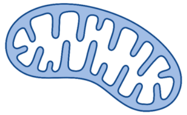
30 millions of people





Rare diseases

responsible for a wide variety of biochemical processes

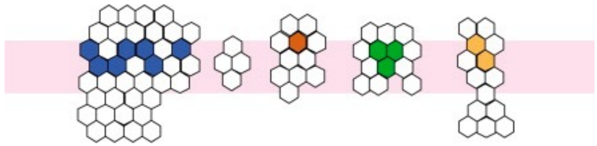


Mitochondria

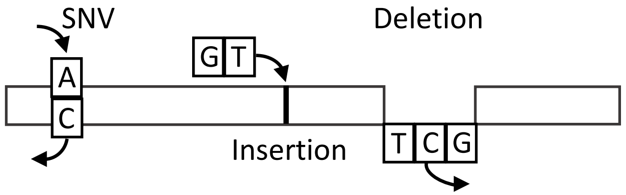
under the double control of mtDNA and nDNA

Mitochondrial diseases

Diagnosis



deficiency of the mitochondrial respiratory chain



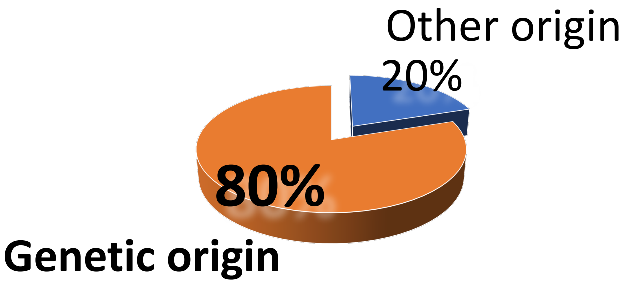
due to rare hereditary or spontaneous variants of mtDNA or nDNA

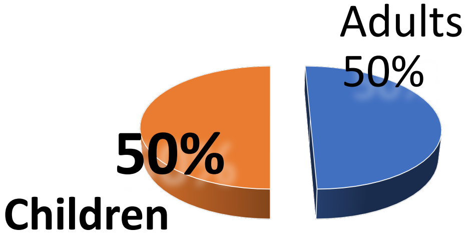
Disease

Medical context

Less than 1 person out of **2000**

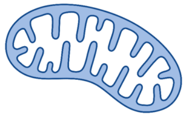
30 millions of people





Rare diseases

responsible for a wide variety of biochemical processes

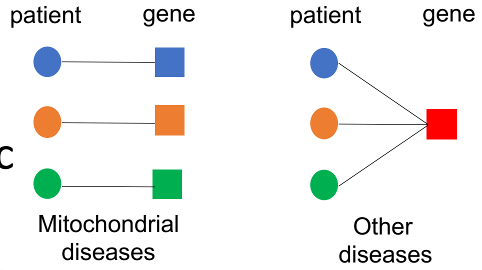


under the double control of mtDNA and nDNA

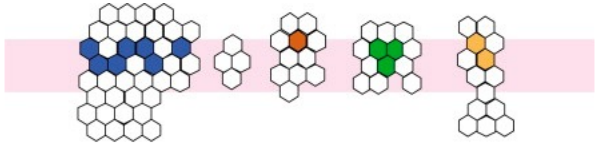
Mitochondria

Mitochondrial diseases

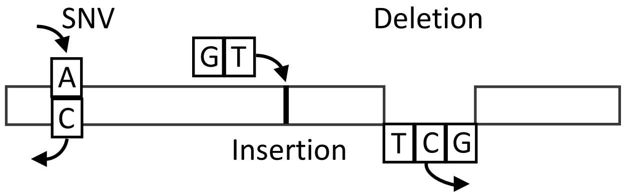
Responsible gene bearing the pathogenic variant



Diagnosis



deficiency of the mitochondrial respiratory chain



due to rare hereditary or spontaneous variants of mtDNA or nDNA

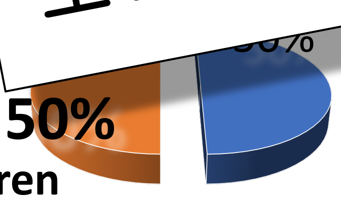
Disease

Medical context

Less than 1 person out of 2000

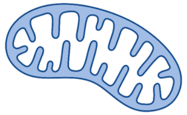
30 millions of people

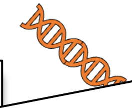
Geneti



Rare diseases

responsible for a wide variety of biochemical processes



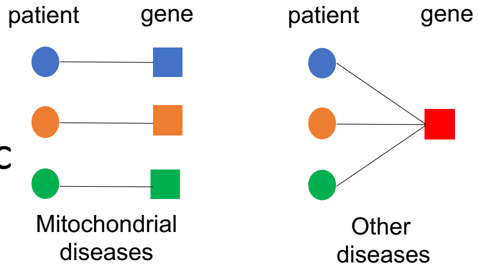


under th

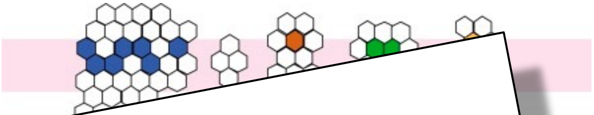
Mitochondria

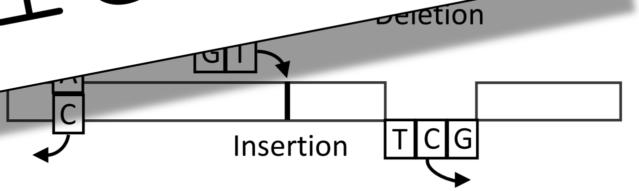
1 patient out of 2 is in diagnostic stalemate

Responsible gene bearing the pathogenic variant



Diagnosis





due to rare hereditary or spontaneous variants of mtDNA or nDNA

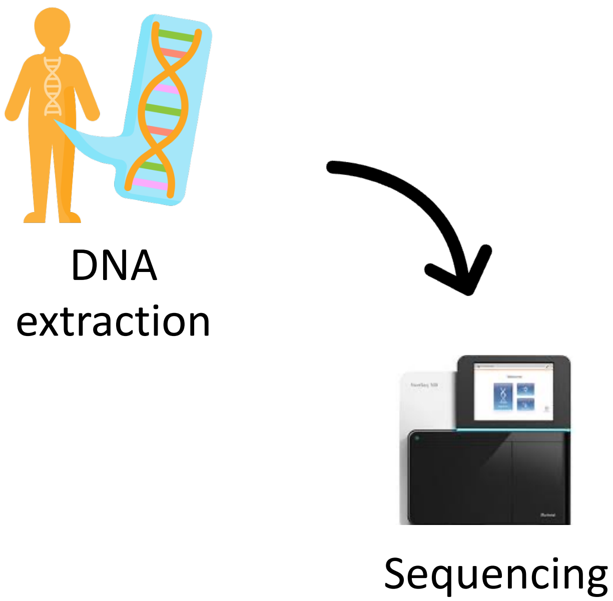
Disease

Diagnosis of Mitochondrial Disease (MD)

Diagnostic = identification of the genetic variant responsible for the disease

Diagnosis of Mitochondrial Disease (MD)

Diagnostic = identification of the genetic variant responsible for the disease

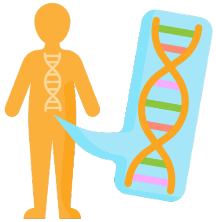


DNA extraction

Sequencing

Diagnosis of Mitochondrial Disease (MD)

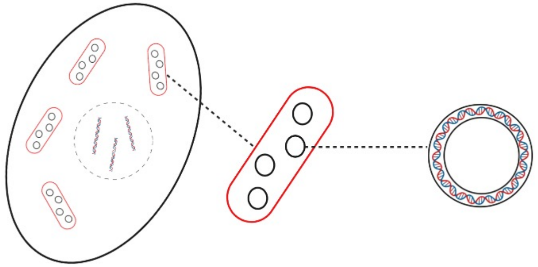
Diagnostic = identification of the genetic variant responsible for the disease



DNA extraction

Sequencing

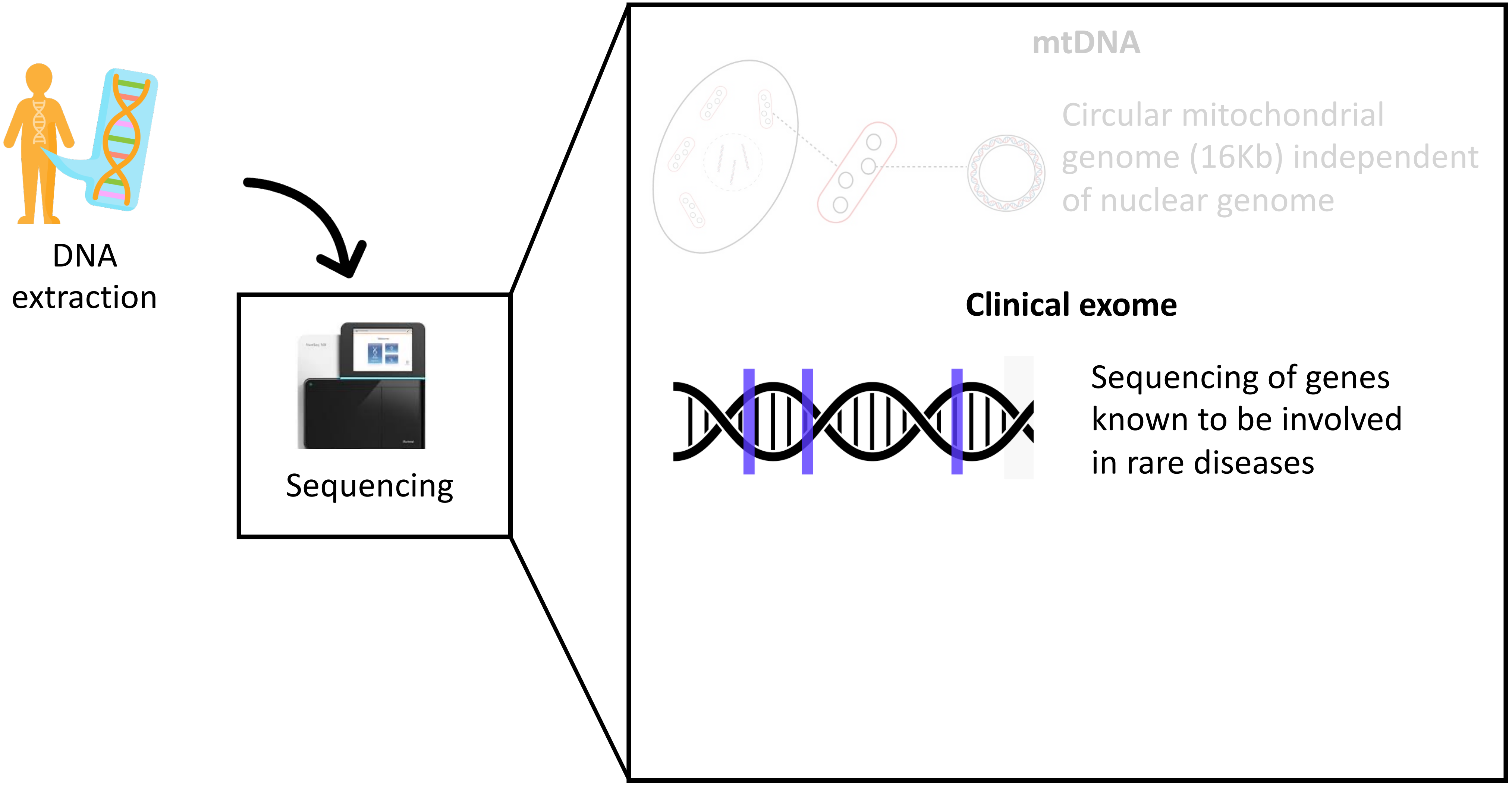
mtDNA



Circular mitochondrial genome (16Kb) independent of nuclear genome

Diagnosis of Mitochondrial Disease (MD)

Diagnostic = identification of the genetic variant responsible for the disease



DNA extraction

Sequencing

mtDNA

Circular mitochondrial genome (16Kb) independent of nuclear genome

Clinical exome

Sequencing of genes known to be involved in rare diseases

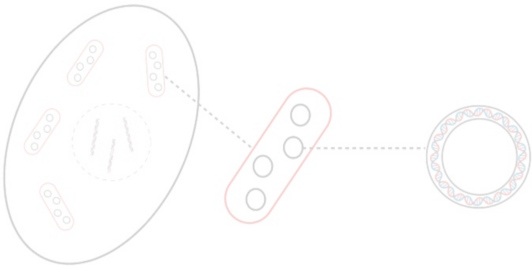
Diagnosis of Mitochondrial Disease (MD)

Diagnostic = identification of the genetic variant responsible for the disease

DNA extraction

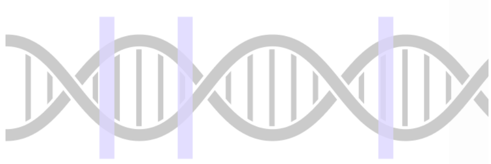
Sequencing

mtDNA



Circular mitochondrial genome (16Kb) independent of nuclear genome

Clinical exome



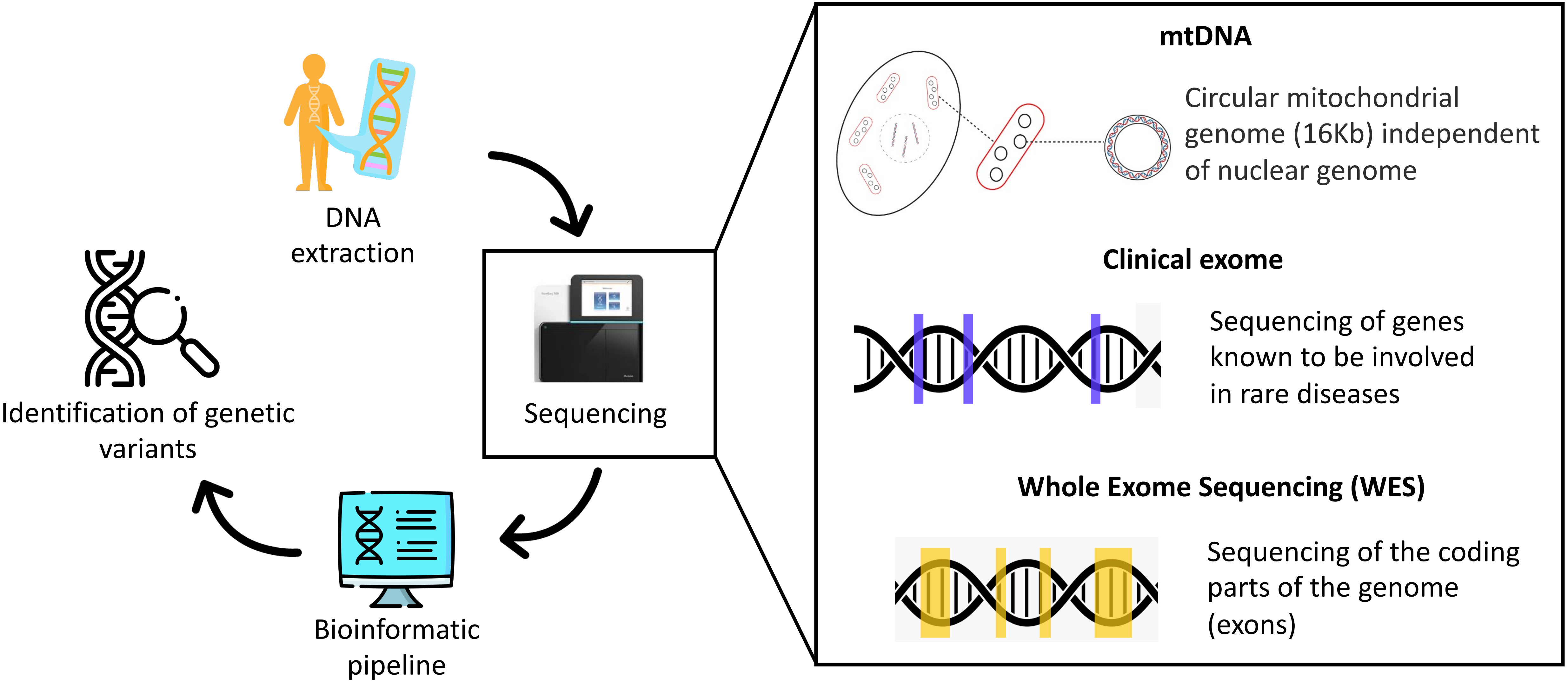
Sequencing of genes known to be involved in rare diseases

Whole Exome Sequencing (WES)

Sequencing of the coding parts of the genome (exons)

Diagnosis of Mitochondrial Disease (MD)

Diagnostic = identification of the genetic variant responsible for the disease



DNA extraction

Sequencing

Bioinformatic pipeline

Identification of genetic variants

mtDNA

Circular mitochondrial genome (16Kb) independent of nuclear genome

Clinical exome

Sequencing of genes known to be involved in rare diseases

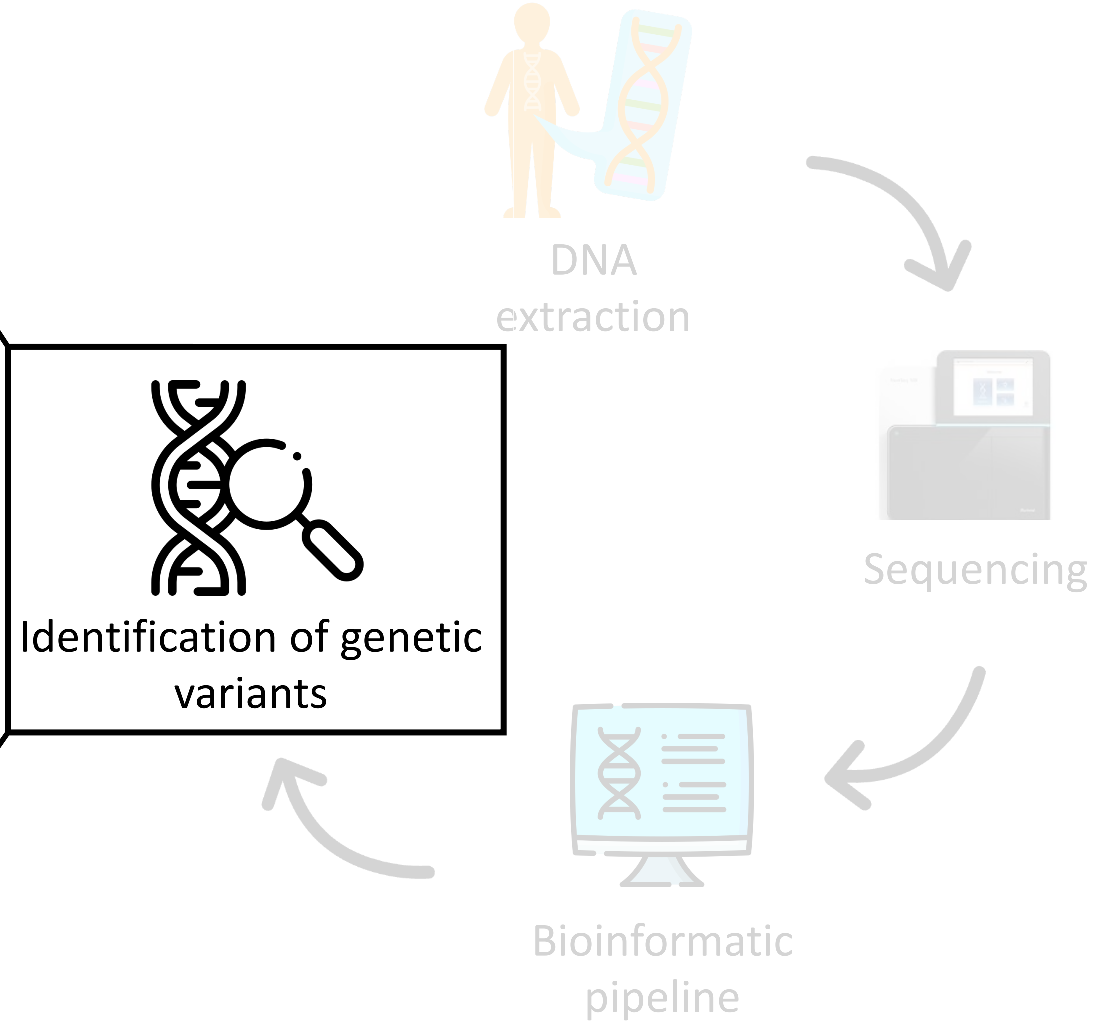
Whole Exome Sequencing (WES)

Sequencing of the coding parts of the genome (exons)

Diagnosis of Mitochondrial Disease (MD)

Challenges :

The variant can be anywhere in genetic sequence (intron, exon, regulatory sequence)



DNA extraction

Sequencing

Bioinformatic pipeline

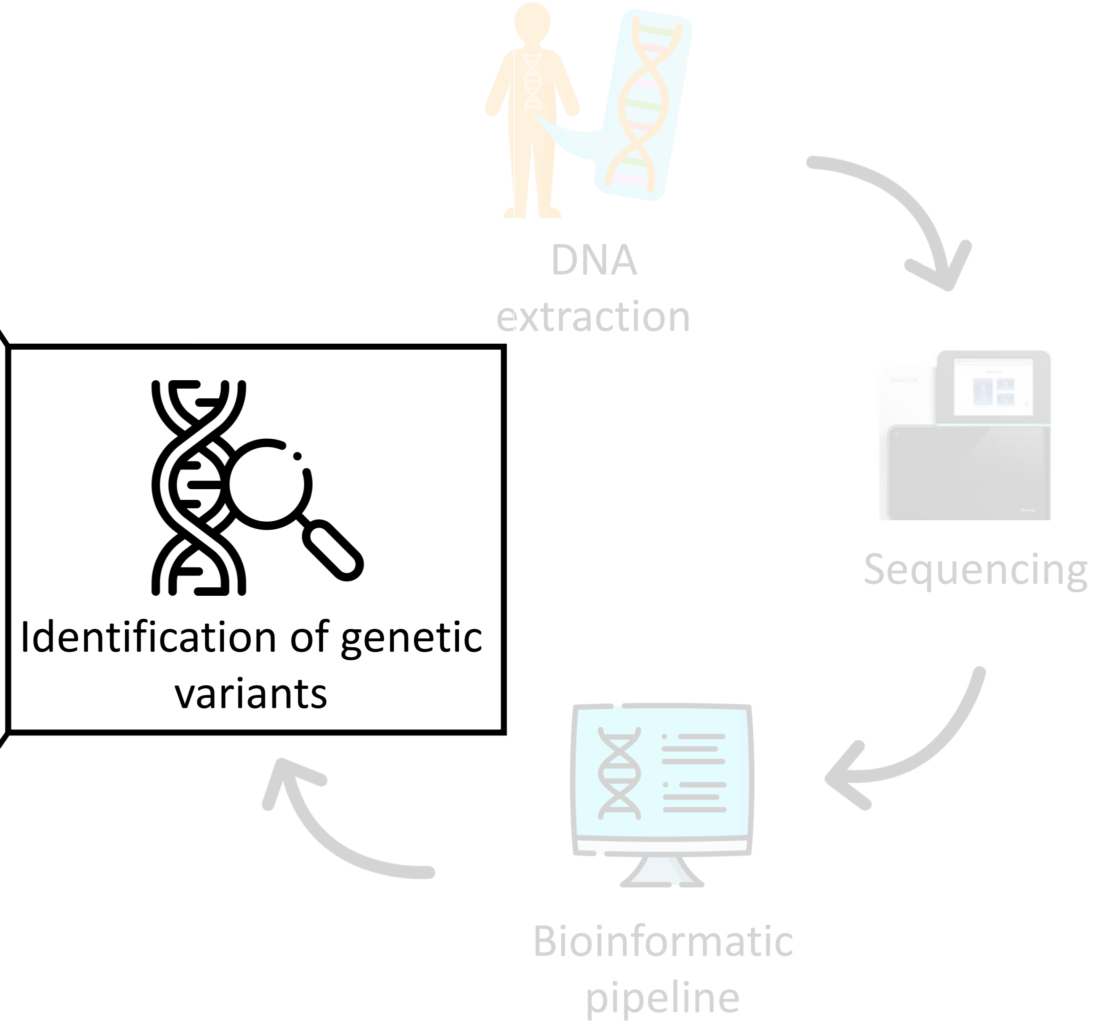
Identification of genetic variants

Diagnosis of Mitochondrial Disease (MD)

Challenges :

The variant can be anywhere in genetic sequence (intron, exon, regulatory sequence)

The variant is rare, i.e is present in less than 1% of the population.



DNA extraction

Sequencing

Bioinformatic pipeline

Identification of genetic variants

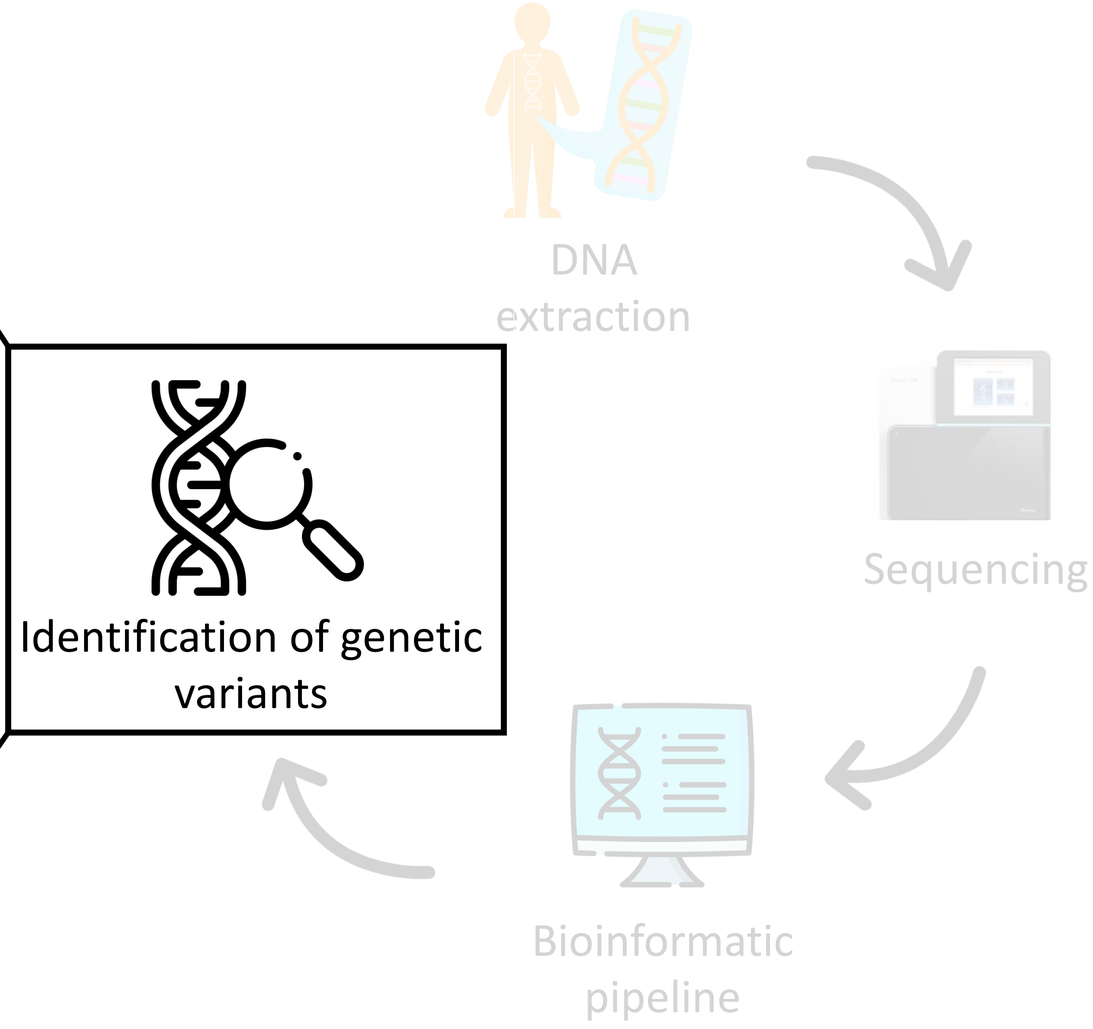
Diagnosis of Mitochondrial Disease (MD)

Challenges :

The variant can be anywhere in genetic sequence (intron, exon, regulatory sequence)

The variant is rare, i.e is present in less than 1% of the population.

Variant poorly characterized in databases



Diagnosis of Mitochondrial Disease (MD)

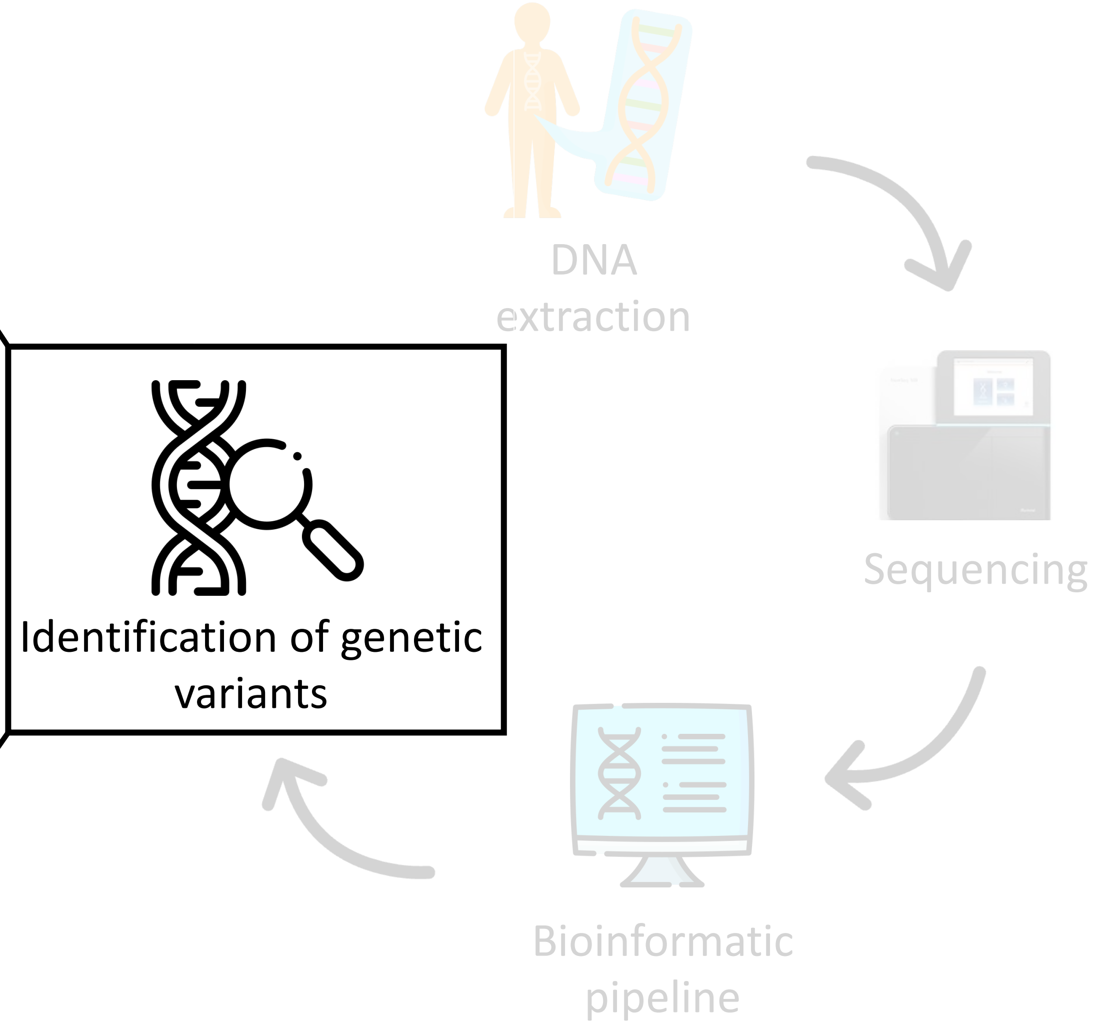
Challenges :

The variant can be anywhere in genetic sequence (intron, exon, regulatory sequence)

The variant is rare, i.e is present in less than 1% of the population.

Variant poorly characterized in databases

1 different responsible variant for each patient



Diagnosis of Mitochondrial Disease (MD)

Challenges :

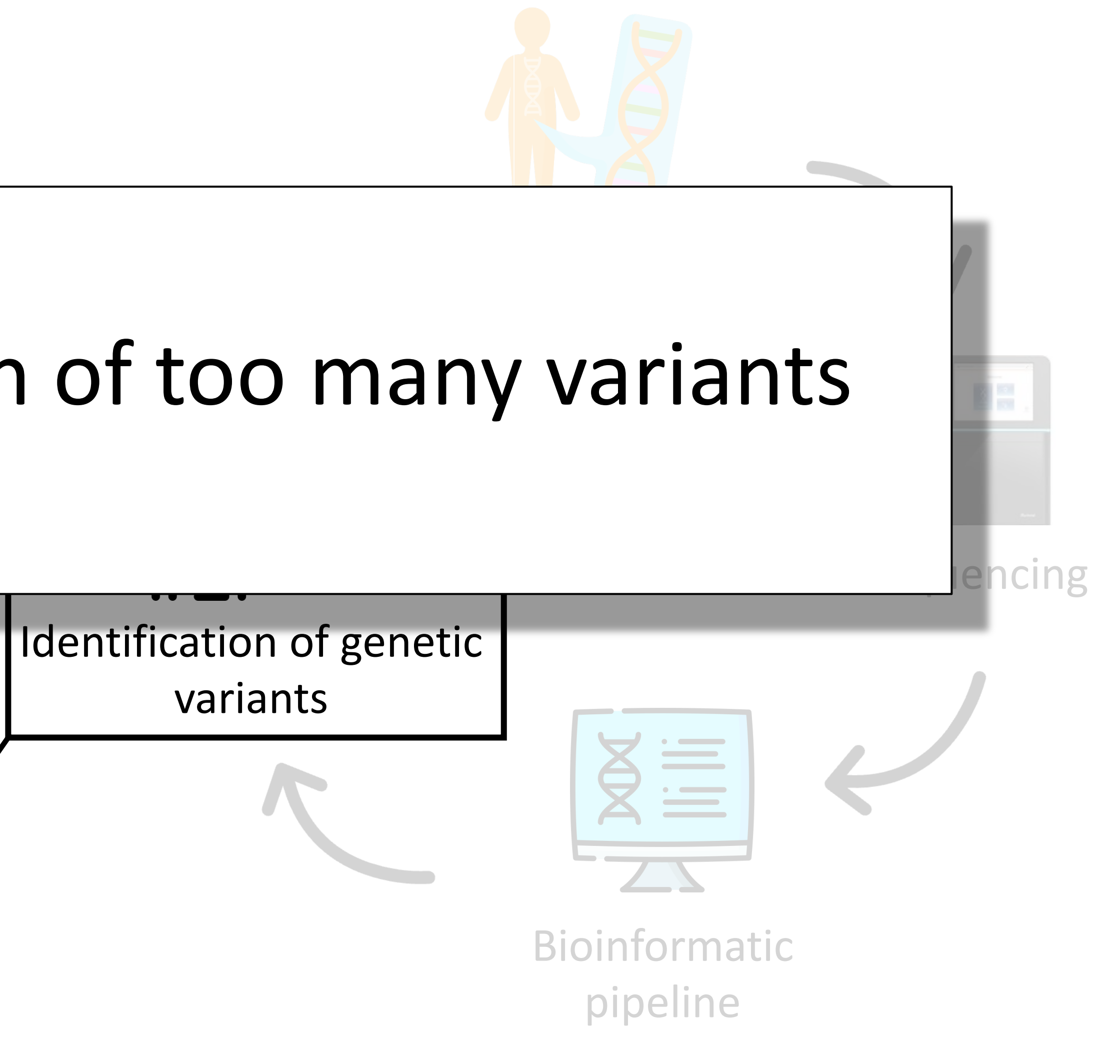
The variant can be anywhere in genetic seque
seque

The va
than 1

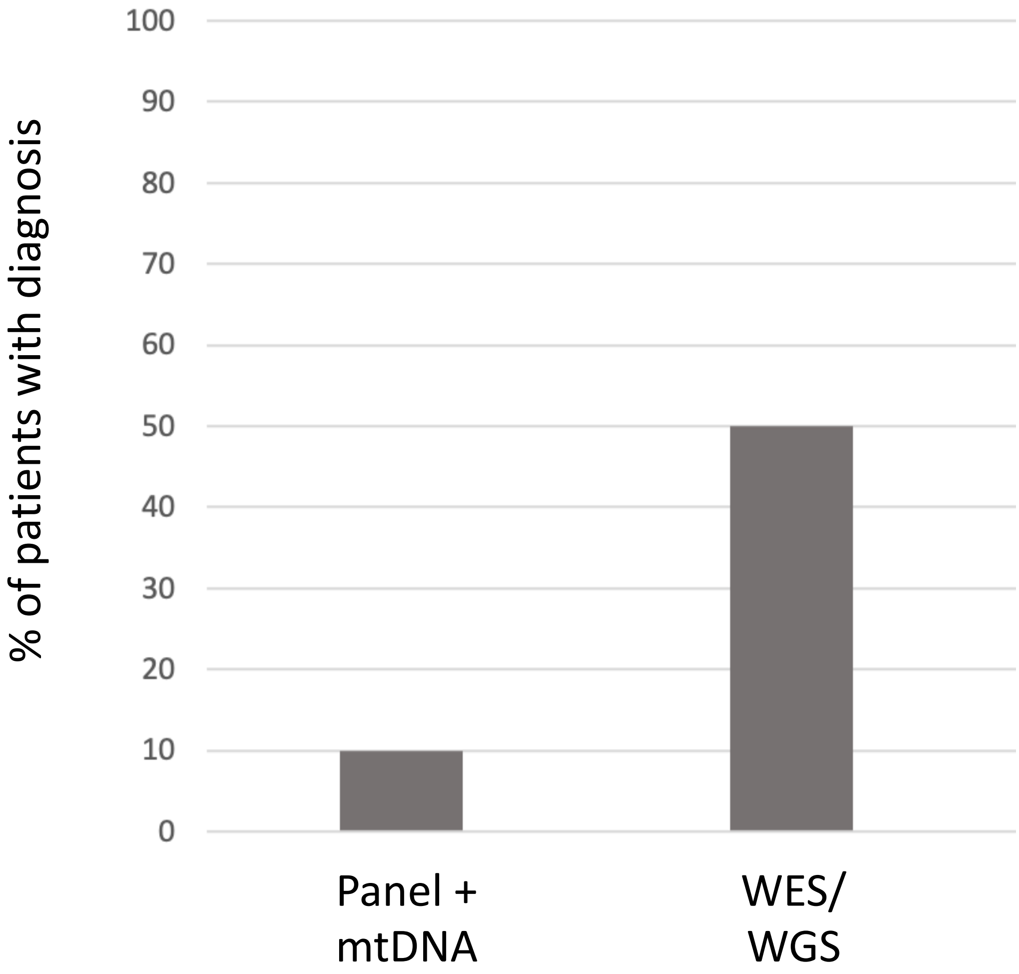
Variant poorly characterized in databases

1 different responsible variant for each patient

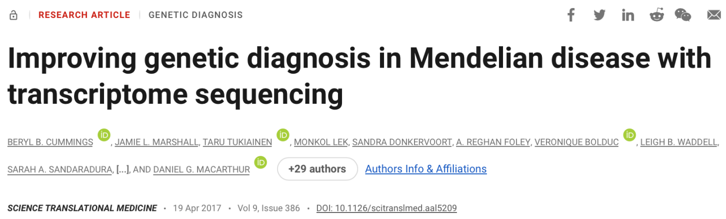
Identification of too many variants



The diagnostic power of MD



RNA-sequencing to improve MD diagnosis

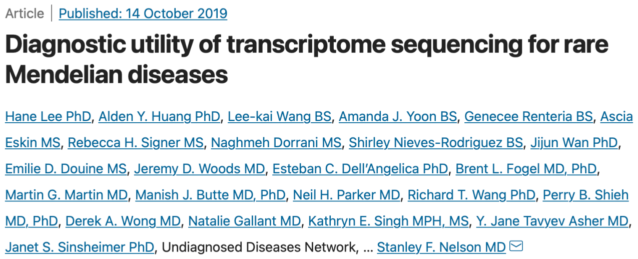


RESEARCH ARTICLE | GENETIC DIAGNOSIS

Improving genetic diagnosis in Mendelian disease with transcriptome sequencing

BERYL B. CUMMINGS, JAMIE L. MARSHALL, TARU TUKIAINEN, MONKOL LEK, SANDRA DONKERVOORT, A. REGHAN FOLEY, VERONIQUE BOLDUC, LEIGH B. WADDELL, SARAH A. SANDARADURA, [...], AND DANIEL G. MACARTHUR +29 authors Authors Info & Affiliations

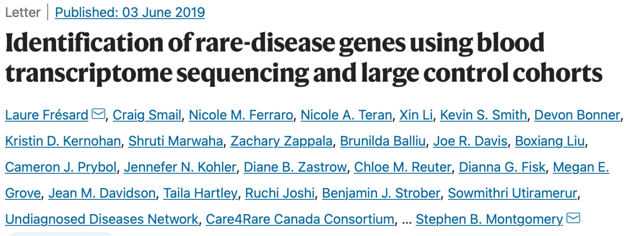
SCIENCE TRANSLATIONAL MEDICINE • 19 Apr 2017 • Vol 9, Issue 386 • DOI: 10.1126/scitranslmed.aal5209



Article | Published: 14 October 2019

Diagnostic utility of transcriptome sequencing for rare Mendelian diseases

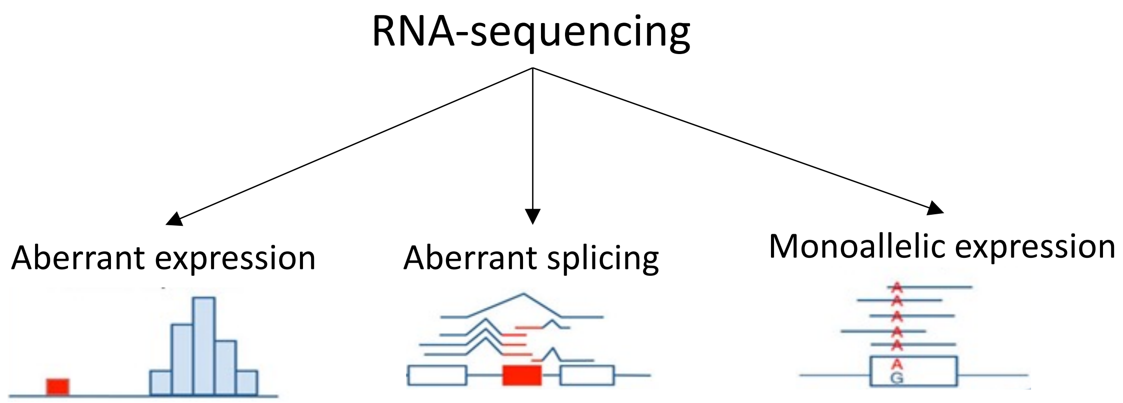
Hane Lee PhD, Alden Y. Huang PhD, Lee-kai Wang BS, Amanda J. Yoon BS, Genecee Renteria BS, Ascia Eskin MS, Rebecca H. Signer MS, Naghmeh Dorrani MS, Shirley Nieves-Rodriguez BS, Jijun Wan PhD, Emilie D. Douine MS, Jeremy D. Woods MD, Esteban C. Dell'Angelica PhD, Brent L. Fogel MD, PhD, Martin G. Martin MD, Manish J. Butte MD, PhD, Neil H. Parker MD, Richard T. Wang PhD, Perry B. Shieh MD, PhD, Derek A. Wong MD, Natalie Gallant MD, Kathryn E. Singh MPH, MS, Y. Jane Tavyev Asher MD, Janet S. Sinsheimer PhD, Undiagnosed Diseases Network, … Stanley F. Nelson MD



Letter | Published: 03 June 2019

Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts

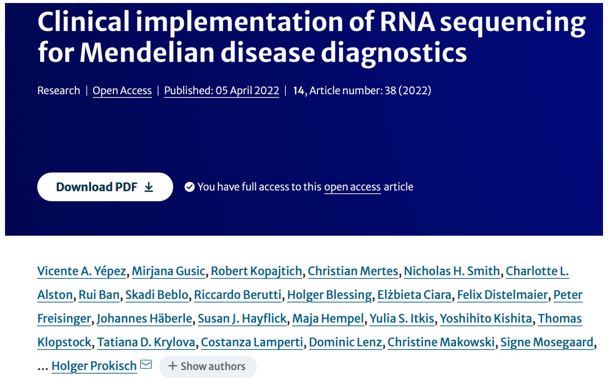
Laure Frésard, Craig Smail, Nicole M. Ferraro, Nicole A. Teran, Xin Li, Kevin S. Smith, Devon Bonner, Kristin D. Kernohan, Shruti Marwaha, Zachary Zappala, Brunilda Balliu, Joe R. Davis, Boxiang Liu, Cameron J. Prybol, Jennefer N. Kohler, Diane B. Zastrow, Chloe M. Reuter, Dianna G. Fisk, Megan E. Grove, Jean M. Davidson, Taila Hartley, Ruchi Joshi, Benjamin J. Strober, Sowmithri Utiramerur, Undiagnosed Diseases Network, Care4Rare Canada Consortium, … Stephen B. Montgomery


RNA-sequencing

Aberrant expression

Aberrant splicing

Monoallelic expression

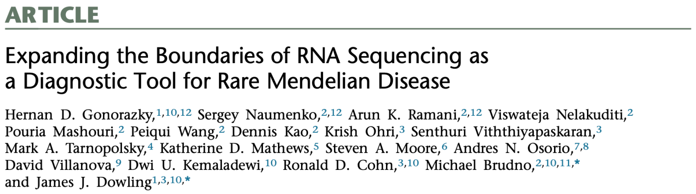


Clinical implementation of RNA sequencing for Mendelian disease diagnostics

Research | Open Access | Published: 05 April 2022 | **14**, Article number: 38 (2022)

Download PDF ✓ You have full access to this open access article

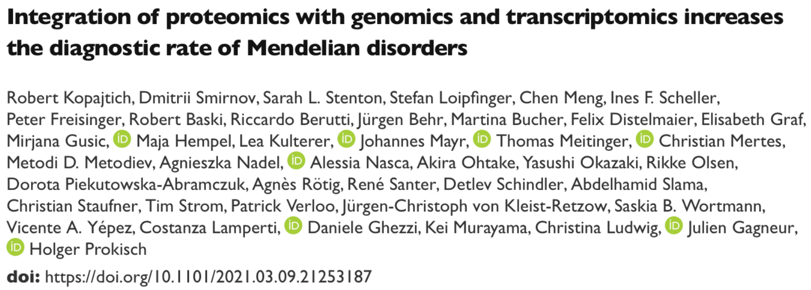
Vicente A. Yépez, Mirjana Gusic, Robert Kopajtich, Christian Mertes, Nicholas H. Smith, Charlotte L. Alston, Rui Ban, Skadi Beblo, Riccardo Berutti, Holger Blessing, Elżbieta Ciara, Felix Distelmaier, Peter Freisinger, Johannes Häberle, Susan J. Hayflick, Maja Hempel, Yulia S. Itkis, Yoshihito Kishita, Thomas Klopstock, Tatiana D. Krylova, Costanza Lamperti, Dominic Lenz, Christine Makowski, Signe Mosegaard, … Holger Prokisch + Show authors



ARTICLE

Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease

Hernan D. Gonorazky,[1,10,12] Sergey Naumenko,[2,12] Arun K. Ramani,[2,12] Viswateja Nelakuditi,[2] Pouria Mashouri,[2] Peiqui Wang,[2] Dennis Kao,[2] Krish Ohri,[3] Senthuri Viththiyapaskaran,[3] Mark A. Tarnopolsky,[4] Katherine D. Mathews,[5] Steven A. Moore,[6] Andres N. Osorio,[7,8] David Villanova,[9] Dwi U. Kemaladewi,[10] Ronald D. Cohn,[3,10] Michael Brudno,[2,10,11,*] and James J. Dowling[1,3,10,*]

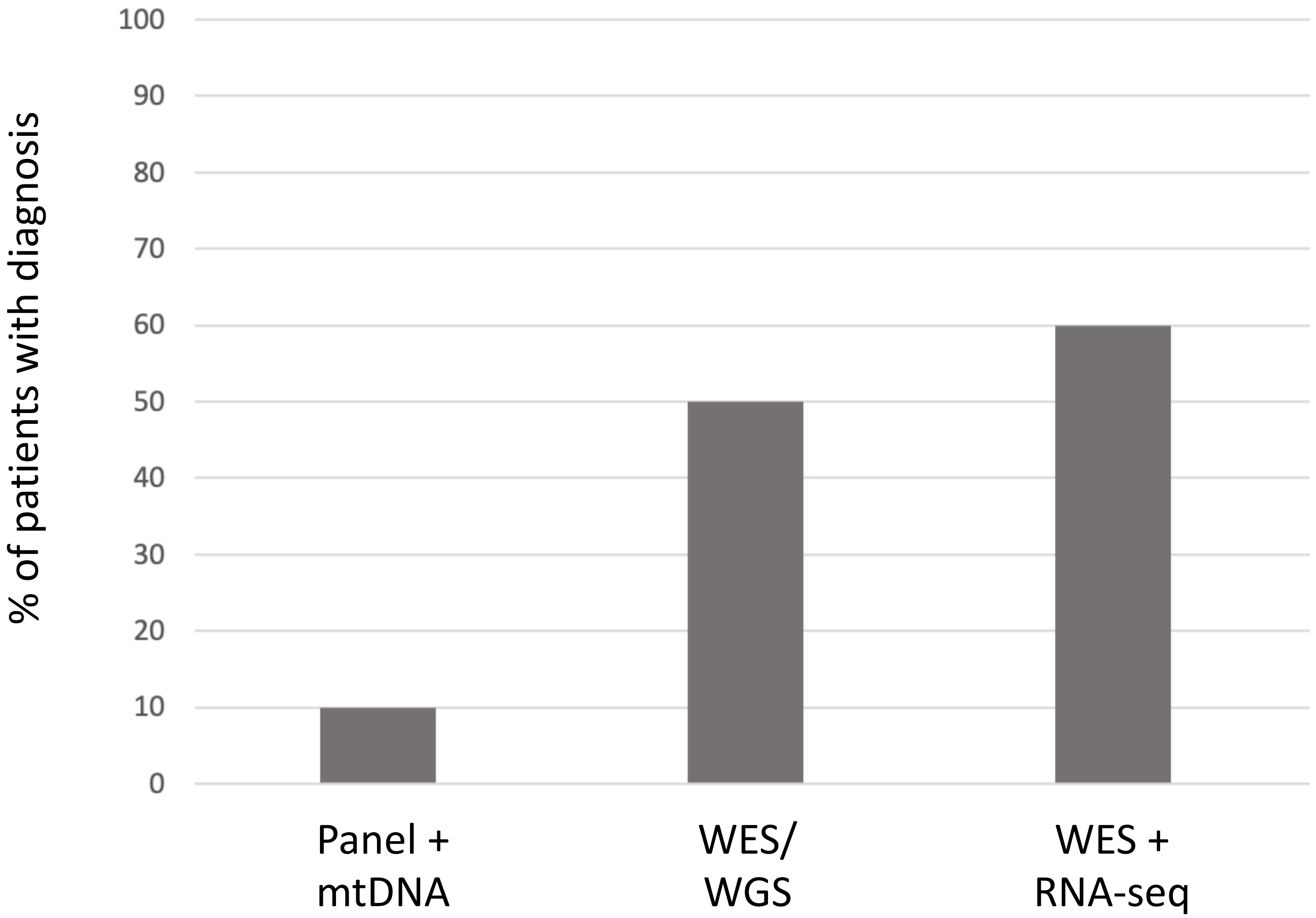


Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders

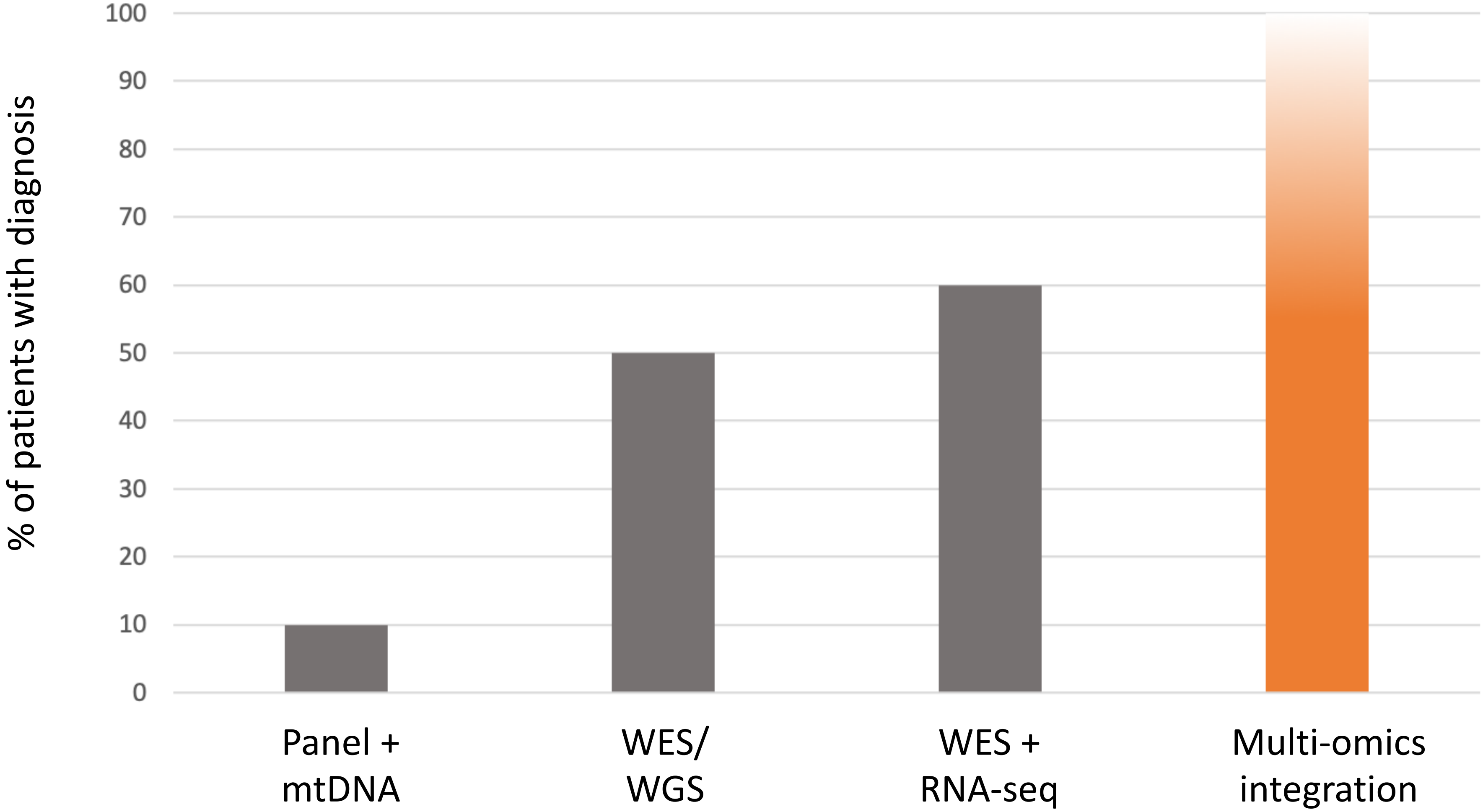
Robert Kopajtich, Dmitrii Smirnov, Sarah L. Stenton, Stefan Loipfinger, Chen Meng, Ines F. Scheller, Peter Freisinger, Robert Baski, Riccardo Berutti, Jürgen Behr, Martina Bucher, Felix Distelmaier, Elisabeth Graf, Mirjana Gusic, Maja Hempel, Lea Kulterer, Johannes Mayr, Thomas Meitinger, Christian Mertes, Metodi D. Metodiev, Agnieszka Nadel, Alessia Nasca, Akira Ohtake, Yasushi Okazaki, Rikke Olsen, Dorota Piekutowska-Abramczuk, Agnès Rötig, René Santer, Detlev Schindler, Abdelhamid Slama, Christian Staufner, Tim Strom, Patrick Verloo, Jürgen-Christoph von Kleist-Retzow, Saskia B. Wortmann, Vicente A. Yépez, Costanza Lamperti, Daniele Ghezzi, Kei Murayama, Christina Ludwig, Julien Gagneur, Holger Prokisch

doi: https://doi.org/10.1101/2021.03.09.21253187

The diagnostic power of MD

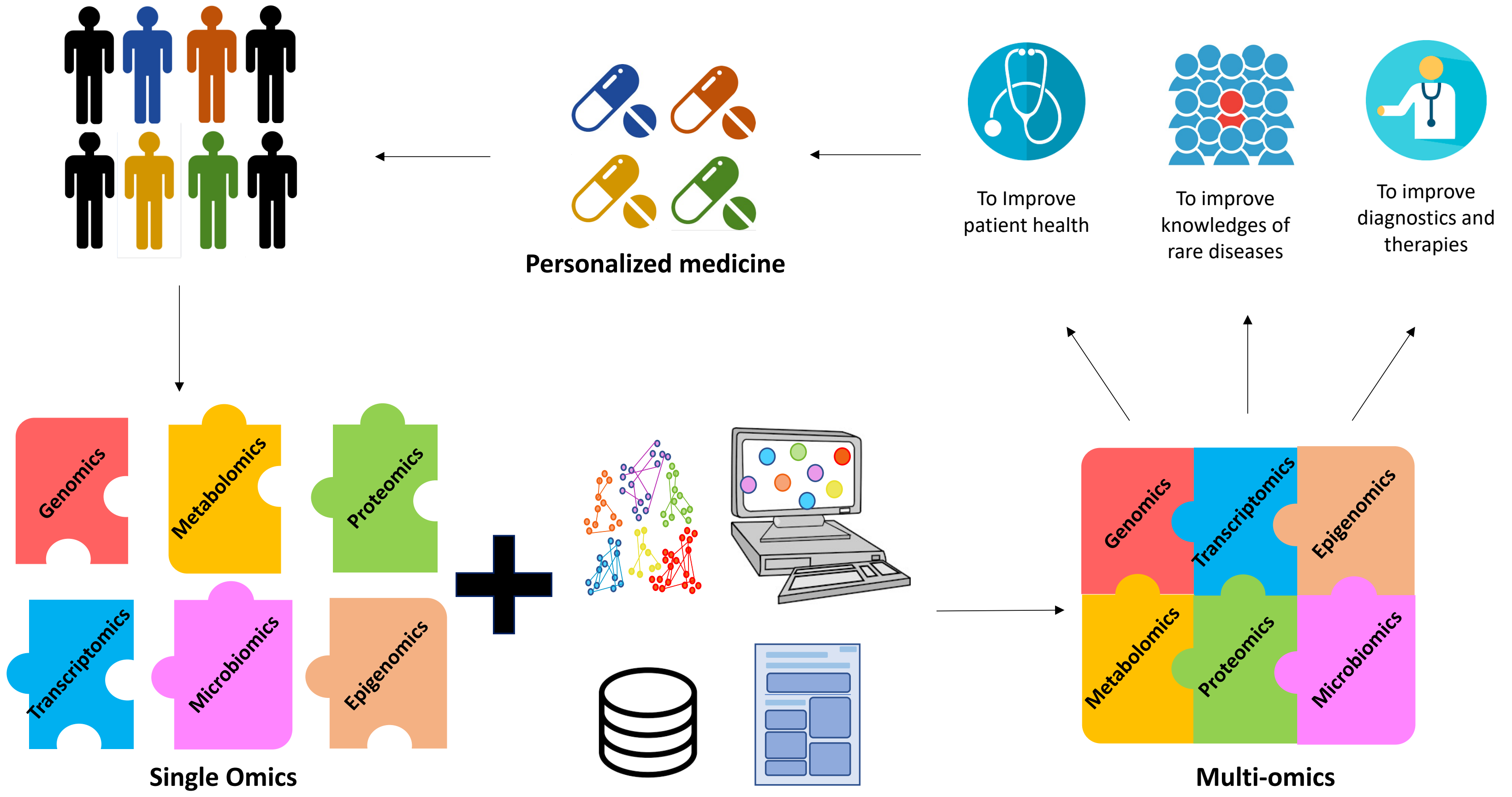


The diagnostic power of MD



Objectives

How to improve the diagnosis of mitochondrial diseases ?



Personalized medicine

To Improve patient health

To improve knowledges of rare diseases

To improve diagnostics and therapies

Single Omics

Multi-omics

From Labory et al. *Front Mol Biosci*.2020

Table of contents

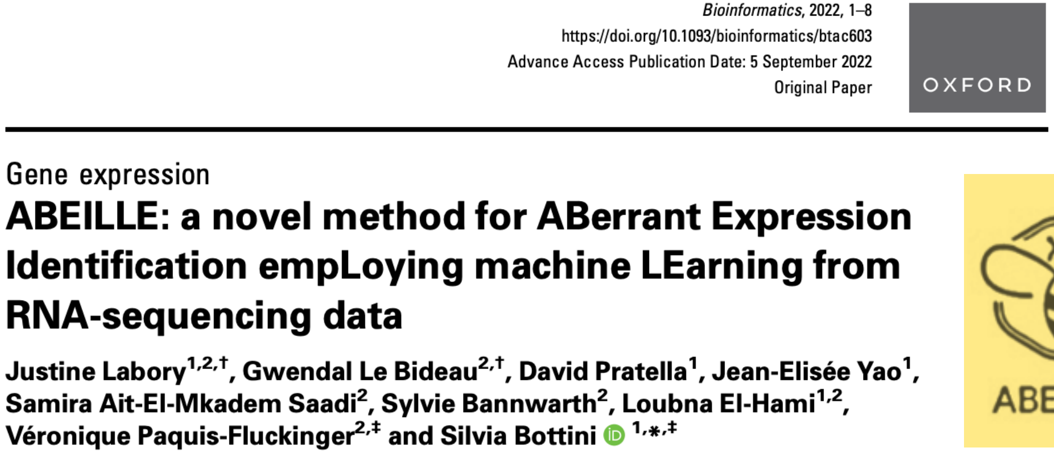
1 ABEILLE (ABerrant Expression Identification empLoying machine LEarning) to find candidate Aberrant Gene expression (AGEs)

2 VIOLA (Variant prIOritization using LAtent space) to find candidate pathogenic genetic variants

Table of contents

1

ABEILLE (ABerrant Expression Identification empLoying machine LEarning) to find candidate Aberrant Gene expression (AGEs)



Bioinformatics, 2022, 1–8
https://doi.org/10.1093/bioinformatics/btac603
Advance Access Publication Date: 5 September 2022
Original Paper

OXFORD

Gene expression

ABEILLE: a novel method for ABerrant Expression Identification empLoying machine LEarning from RNA-sequencing data

Justine Labory[1,2,†], Gwendal Le Bideau[2,†], David Pratella[1], Jean-Elisée Yao[1], Samira Ait-El-Mkadem Saadi[2], Sylvie Bannwarth[2], Loubna El-Hami[1,2], Véronique Paquis-Fluckinger[2,‡] and Silvia Bottini [1,*,‡]

ABEILLE

https://github.com/UCA-MSI/ABEILLE

Context : RNA-seq questions

2 approaches

Context : RNA-seq questions

2 approaches

Which genes are differentially expressed between 2 groups ?

Differential Expression (DE)

	Control 1	Control 2	Control 3	Patient 1	Patient 2	Patient 3
Gene A	🟦	🟦	🟦	🟥	🟥	🟥
Gene B	🟦	🟦	🟦	🟦	🟦	🟦
Gene C	🟦	🟦	🟦	🟥	🟥	🟦

→ Gene A = **DE**

→ Gene B = Normal gene

→ Gene C = **DE**

Control group

Replicates

Tool: DESeq2[1]

[1] Love et al. *Genome Biology.*2014

Context : RNA-seq questions

2 approaches

Which genes are differentially expressed between 2 groups ?

Differential Expression (DE)

	Control 1	Control 2	Control 3	Patient 1	Patient 2	Patient 3
Gene A	🟦	🟦	🟦	🟧	🟥	🟥
Gene B	🟦	🟦	🟦	🟦	🟦	🟦
Gene C	🟦	🟦	🟦	🟥	🟧	🟦

→ Gene A = **DE**

→ Gene B = Normal gene

→ Gene C = **DE**

⚠ Control group

⚠ Replicates

Tool: DESeq2[1]

Which genes are AGEs for each patient ?

Aberrant Gene Expression (AGE)

	Patient 1	Patient 2	Patient 3
Gene A	🟧	🟥	🟥
Gene B	🟦	🟦	🟦
Gene C	🟥	🟧	🟦

→ [Gene A;Patient 2] = **AGE**

→ Gene B = Normal gene

→ [Gene C;Patient 1] = **AGE**

No control group

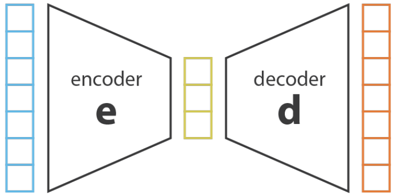
No replicates

[1] Love et al. *Genome Biology.*2014

Methods to identify AGEs

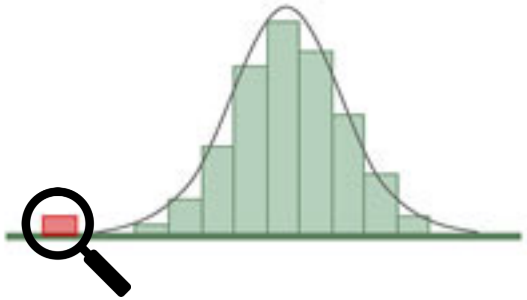
OUTRIDER[1]

Autoencoder



+

Statistical test

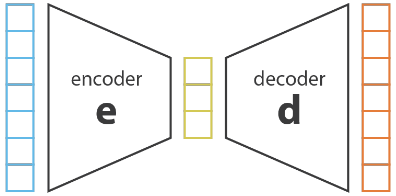


[1] Brechtmann et al. *Am. J. Hum. Genet.* 2018

Methods to identify AGEs

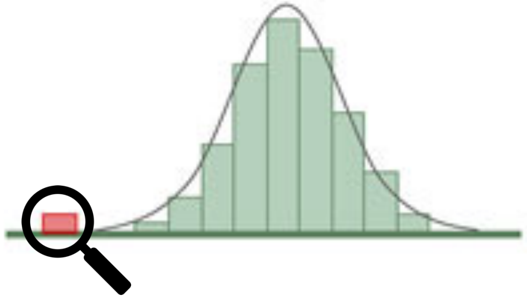
OUTRIDER[1]

Autoencoder



+

Statistical test



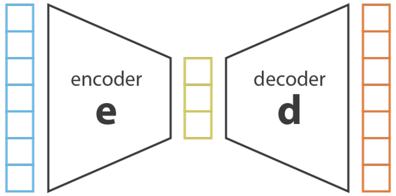
➔ OUTRIDER does not work on small data

[1] Brechtmann et al. *Am. J. Hum. Genet.* 2018

Methods to identify AGEs

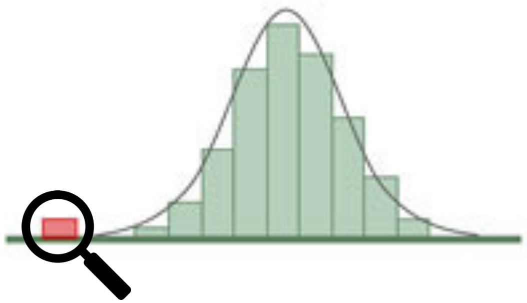
OUTRIDER[1]

Autoencoder



+

Statistical test



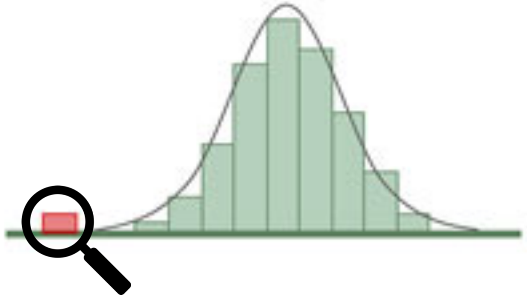
➔ OUTRIDER does not work on small data

OutPyR[2]

Bayesian model

+

Statistical test



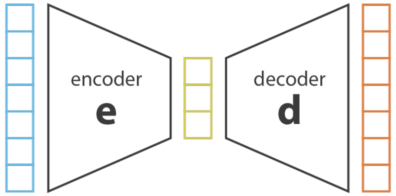
[1] Brechtmann et al. *Am. J. Hum. Genet.* 2018

[2] Salkovic et al. *Journal of Computational Science,*.2020

Methods to identify AGEs

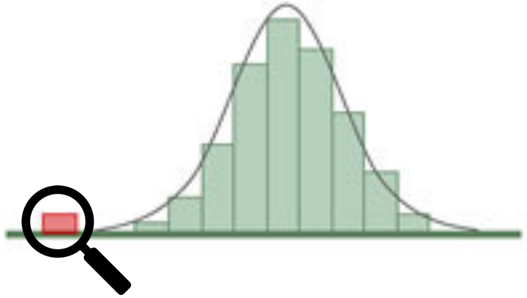
OUTRIDER[1]

Autoencoder



+

Statistical test



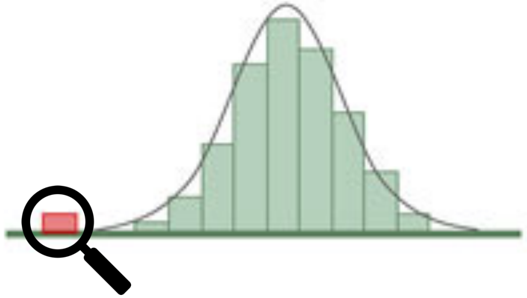
➔ OUTRIDER does not work on small data

OutPyR[2]

Bayesian model

+

Statistical test



➔ Tested only on a subset of real data

[1] Brechtmann et al. *Am. J. Hum. Genet.* 2018

[2] Salkovic et al. *Journal of Computational Science,*.2020

Methods to identify AGEs

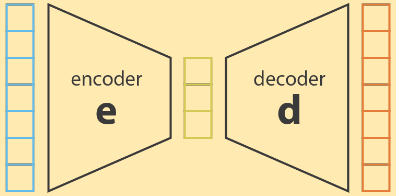
How to identify AGEs for small cohorts ?

Methods to identify AGEs

How to identify AGEs for small cohorts ?

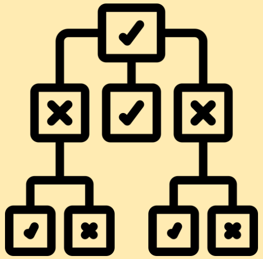
ABEILLE[1]

Variational Autoencoder



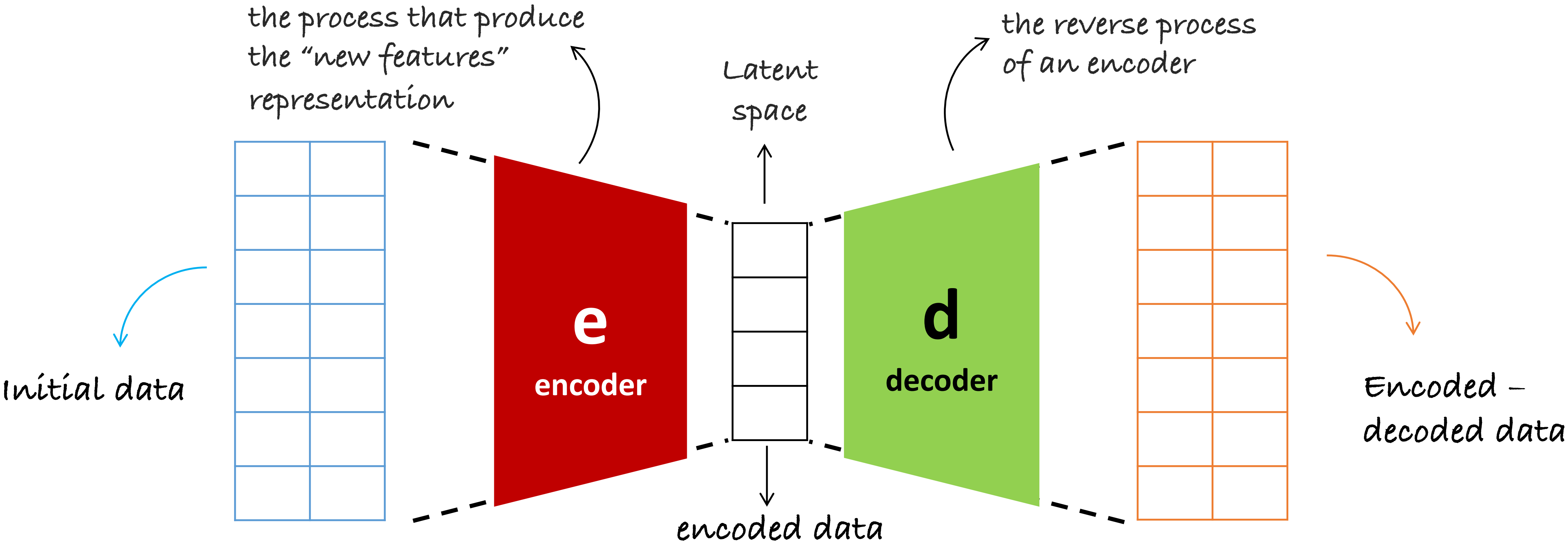
+

Decision tree



[1] **Labory et al.** *Bioinformatics* 2022

The autoencoder


the process that produce the "new features" representation
Latent space
the reverse process of an encoder
Initial data
e
encoder
d
decoder
Encoded – decoded data
encoded data

Dimensionality Reduction

Denoising

Sequence to sequence prediction

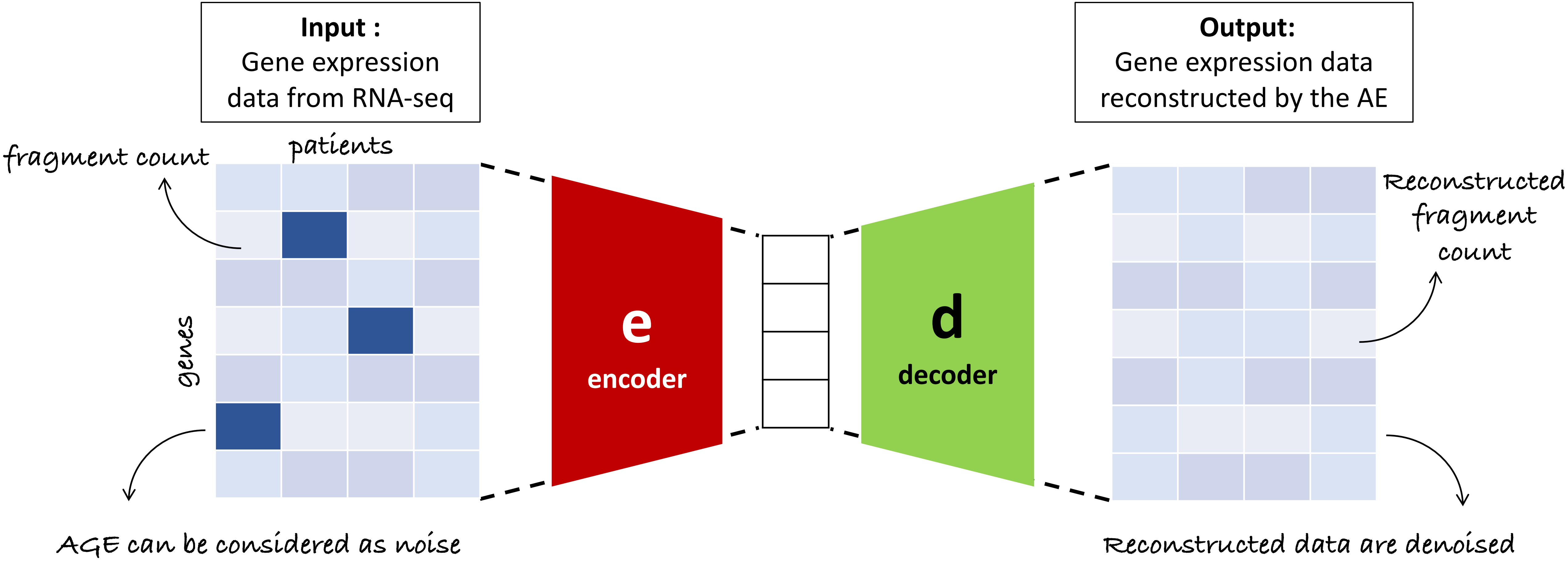
Compression

Feature Extraction

Generation

Recommendation system

How to use AE to identify AGEs ?



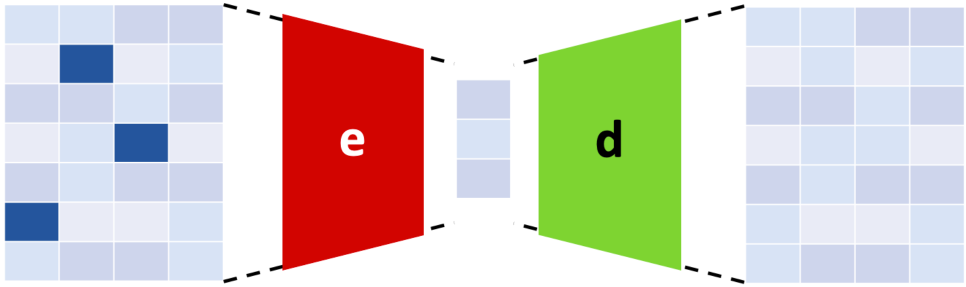
Input :
Gene expression
data from RNA-seq

Output:
Gene expression data
reconstructed by the AE

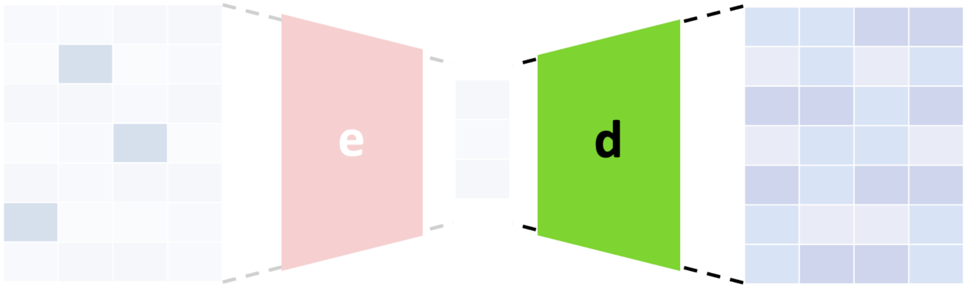
AGE can be considered as noise

Reconstructed data are denoised

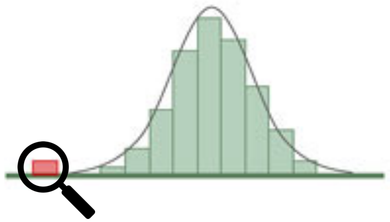
Difference between ABEILLE and OUTRIDER



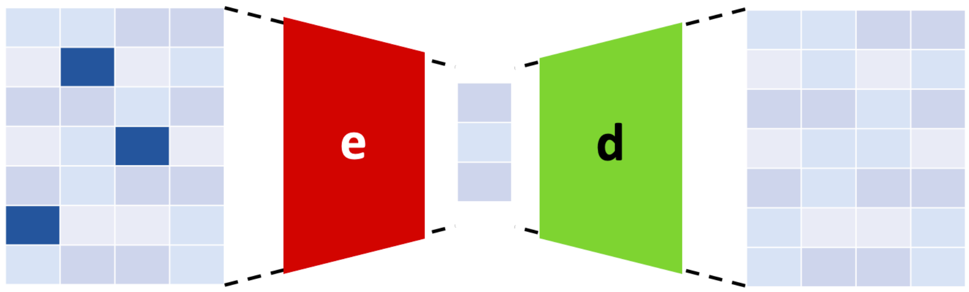
Difference between ABEILLE and OUTRIDER



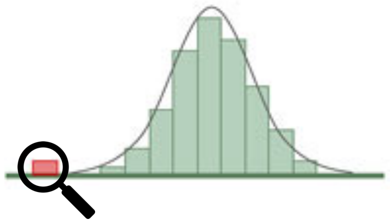
→ OUTRIDER ➡ Statistical test to find AGEs



Difference between ABEILLE and OUTRIDER



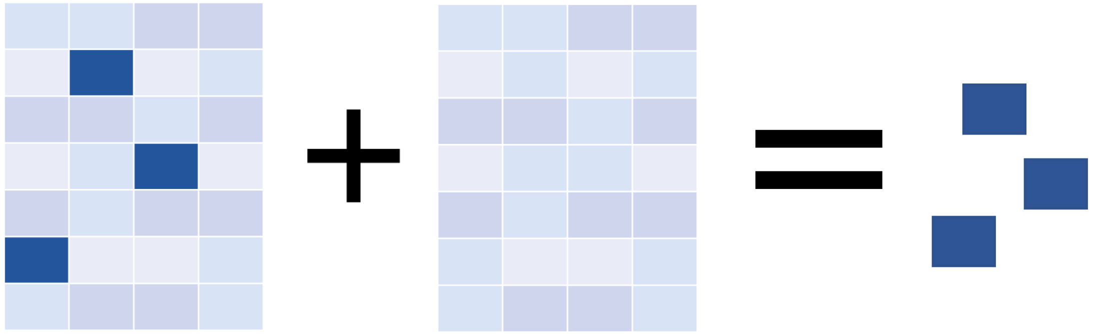
OUTRIDER ➡ Statistical test to find AGEs



ABEILLE

⬇

To compare input and output of the VAE to find AGEs

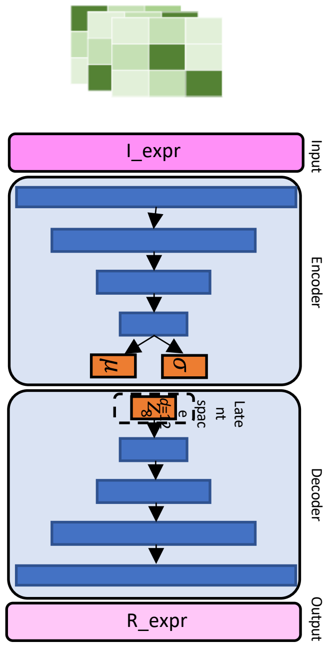


ABEILLE workflow

ABEILLE workflow

1 **VAE**

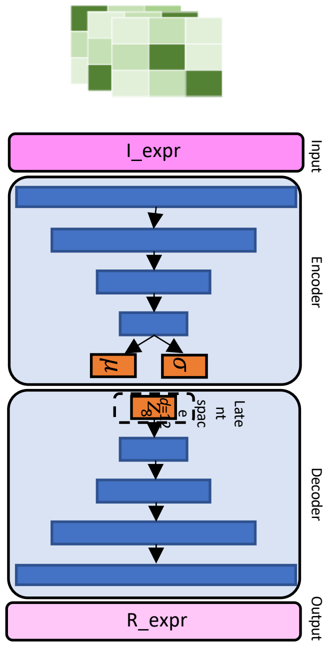
To use VAE to generate reconstructed denoised counts



ABEILLE workflow

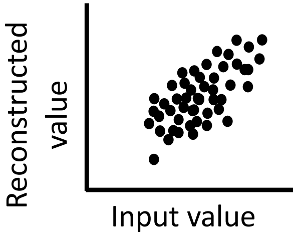
1 VAE

To use VAE to generate reconstructed denoised counts

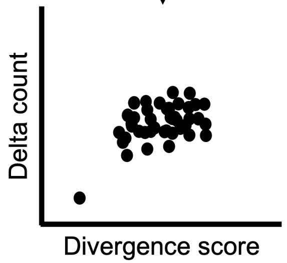


2 Definition of metrics

To compute metrics to assess the reconstruction fidelity



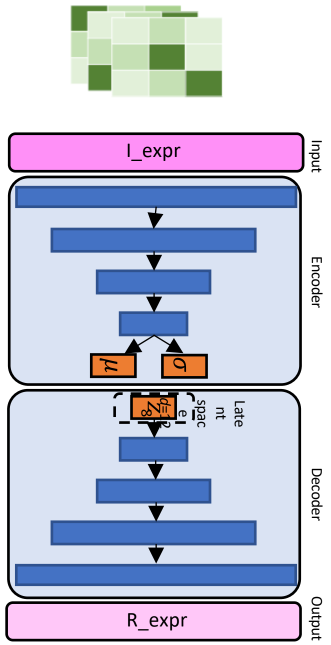
Definition of new scores to compare I_expr and R_expr



ABEILLE workflow

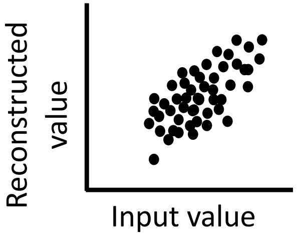
1 VAE

To use VAE to generate reconstructed denoised counts

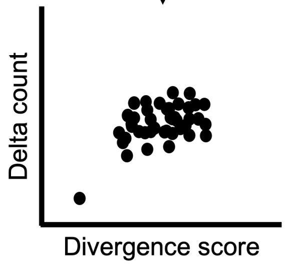


2 Definition of metrics

To compute metrics to assess the reconstruction fidelity



Definition of new scores to compare I_expr and R_expr



3 Classification

To classify gene expressions as AGEs or no AGEs

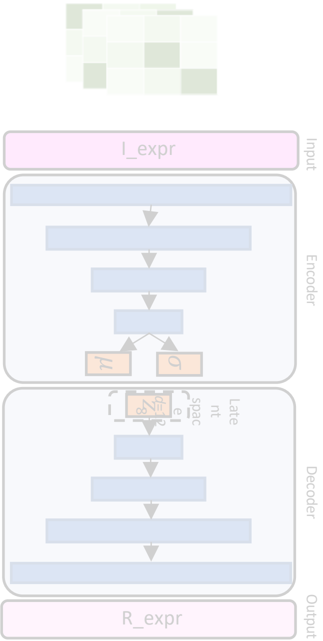
Supervised → Semi-synthetics data → Decision tree

Unsupervised → Real data

ABEILLE workflow

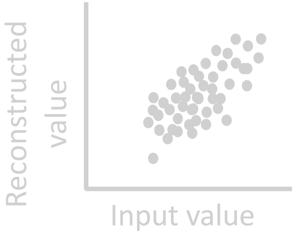
1 VAE

To use VAE to generate reconstructed denoised counts

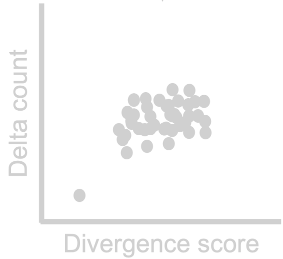


2 Definition of metrics

To compute metrics to assess the reconstruction fidelity



Definition of new scores to compare I_expr and R_expr



3 Classification

To classify gene expressions as AGEs or no AGEs

Supervised → Semi-synthetics data → Decision tree

Unsupervised → Real data

Supervised phase – Creation of semi-synthetics datasets

GTEx

54 tissue

1000 individuals

56 200 transcripts

Supervised phase – Creation of semi-synthetics datasets

GTEx

1 tissue

504 individuals

56 200 transcripts

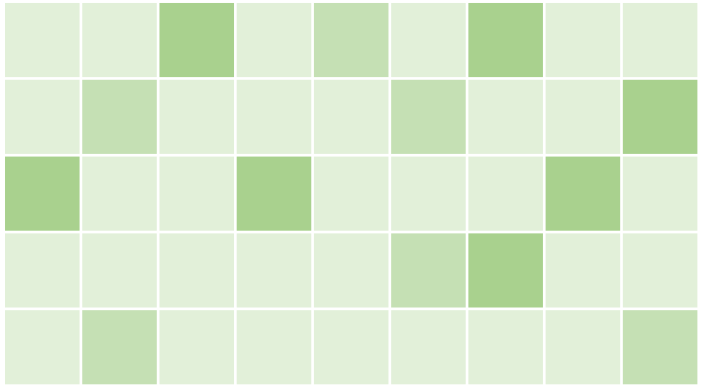
Supervised phase – Creation of semi-synthetics datasets

GTEx

1 tissue

504 individuals

56 200 transcripts



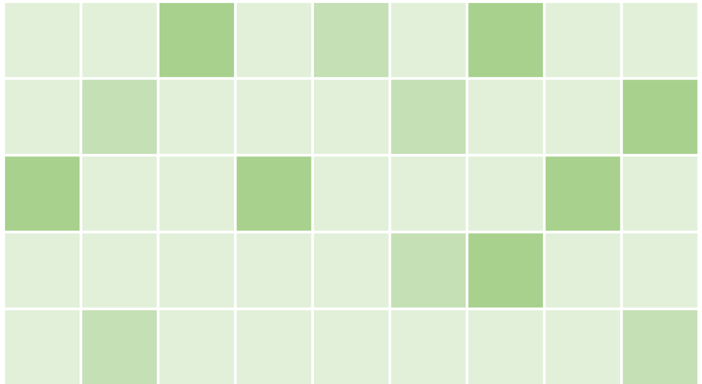
Supervised phase – Creation of semi-synthetics datasets

GTEx

1 tissue

504 individuals

56 200 transcripts



🟩

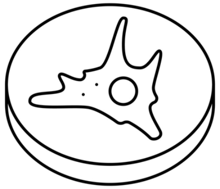
Generate computational AGEs by replacing randomly 10 000 expression values

$$k_{ij}^{O} = \mathrm{round}\left(s_i 2^{\mu_j^u \pm \exp(N)\sigma_j^u}\right)$$

Supervised phase – Creation of semi-synthetics datasets

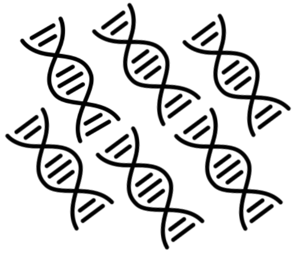
GTEx

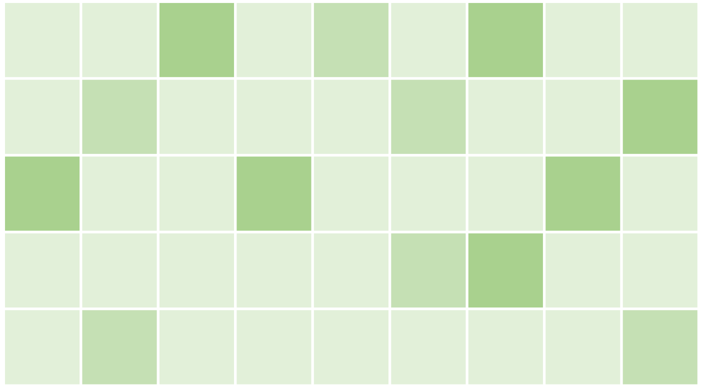
1 tissue



504 individuals

56 200 transcripts

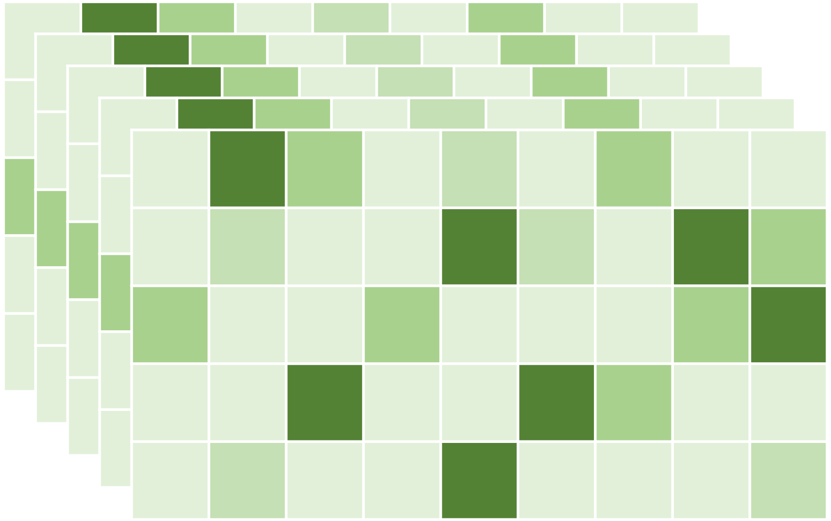




🟩

Generate computational AGEs by replacing randomly 10 000 expression values

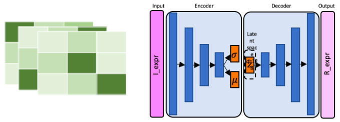
$$k_{ij}^{O} = \text{round}\left(s_i 2^{\mu_j^{u} \pm \exp(N)\sigma_j^{u}}\right)$$



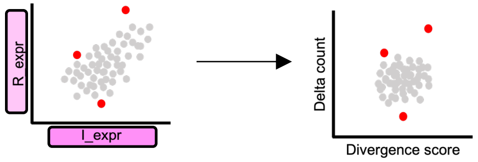
Repeat the process 20 times

Supervised phase – To obtain the decision tree

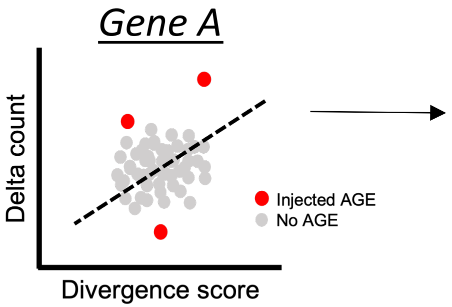
1 **To use VAE to generate reconstructed denoised counts**



2 **To compute metrics to assess the reconstruction fidelity**



3 To create a decision tree and identify thresholds for gene expression classification



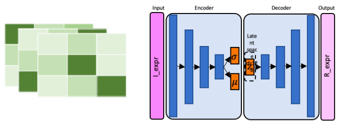
Linear regression

Parameters calculated on each linear regression :

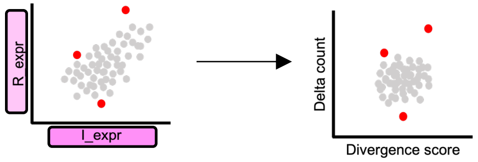
- Dfbetas
- Hat
- Type error

Supervised phase – To obtain the decision tree

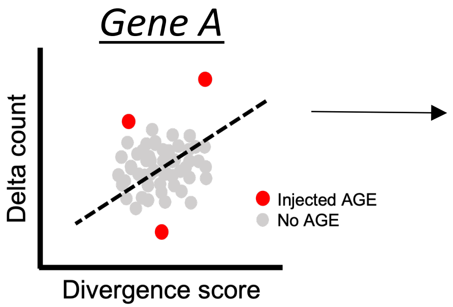
1 To use VAE to generate reconstructed denoised counts



2 To compute metrics to assess the reconstruction fidelity



3 To create a decision tree and identify thresholds for gene expression classification



Parameters calculated on each linear regression :

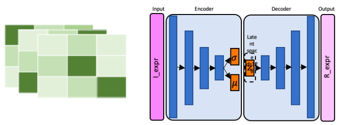
- Dfbetas
- Hat
- Type error

Linear regression

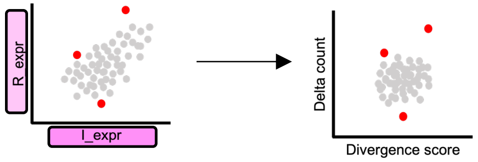
To be done for each gene

Supervised phase – To obtain the decision tree

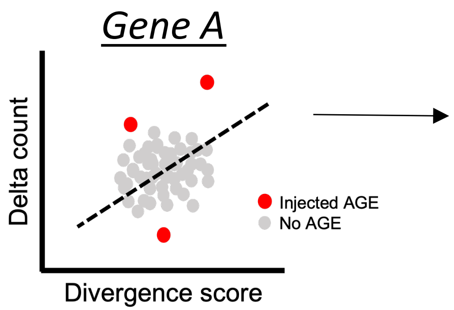
1 To use VAE to generate reconstructed denoised counts



2 To compute metrics to assess the reconstruction fidelity



3 To create a decision tree and identify thresholds for gene expression classification



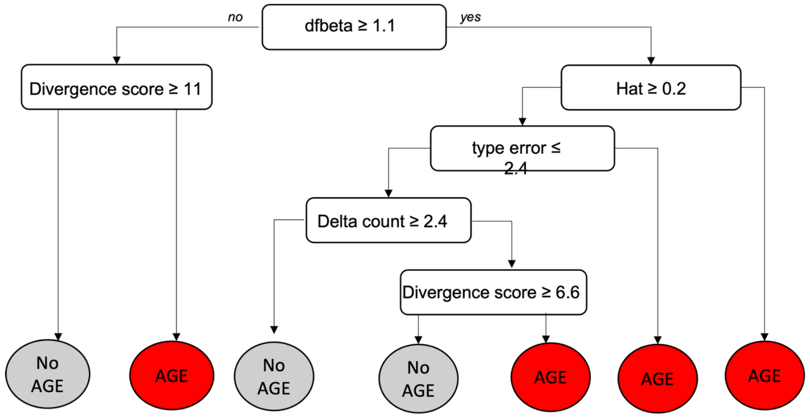
Parameters calculated on each linear regression :

- Dfbetas
- Hat
- Type error

Linear regression

To be done for each gene

Parameters calculated for all genes in all patients are used to feed a decision tree



dfbeta ≥ 1.1

no

yes

Divergence score ≥ 11

Hat ≥ 0.2

type error ≤ 2.4

Delta count ≥ 2.4

Divergence score ≥ 6.6

No AGE

AGE

No AGE

No AGE

AGE

AGE

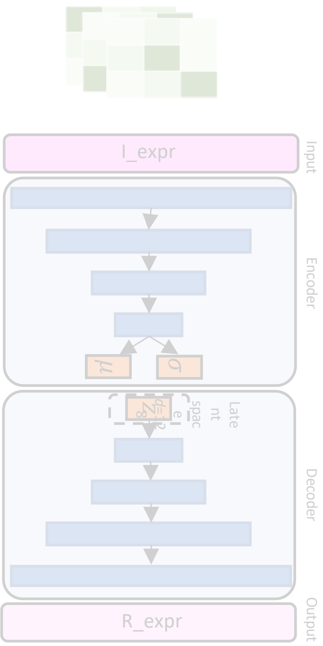
AGE

This decision tree will be used in the unsupervised phase for gene expression classification

ABEILLE workflow

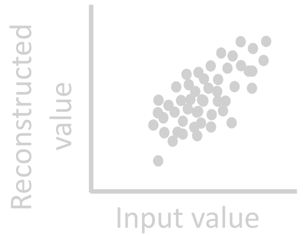
1 VAE

To use VAE to generate reconstructed denoised counts



2 Definition of metrics

To compute metrics to assess the reconstruction fidelity

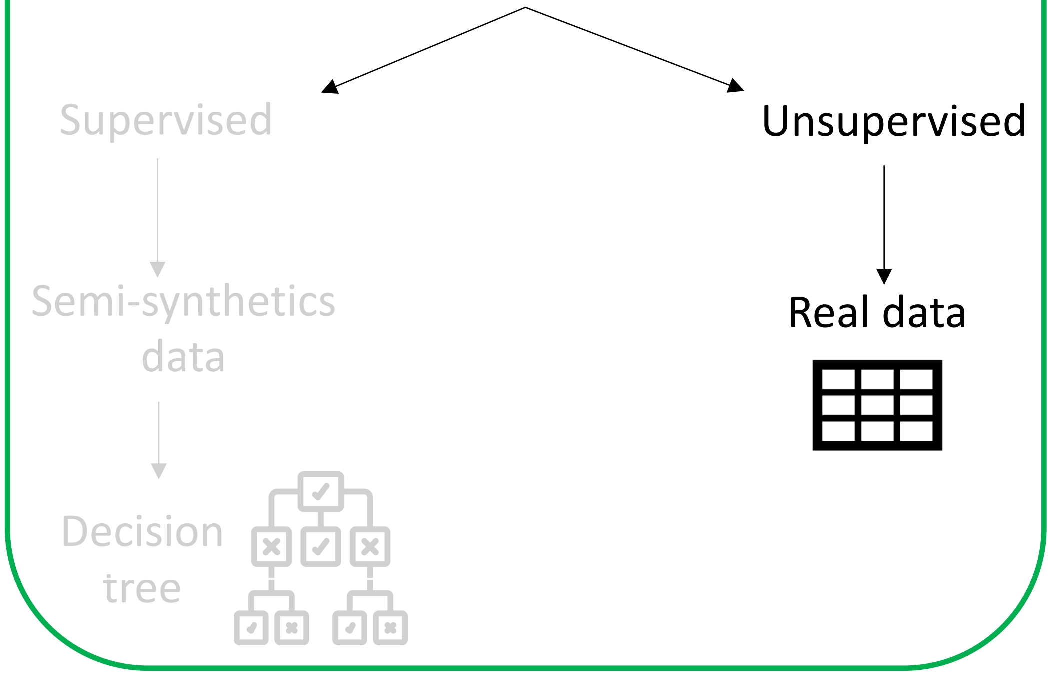


Definition of new scores to compare I_expr and R_expr

Delta count
Divergence score

3 Classification

To classify gene expressions as AGEs or no AGEs



Supervised

Semi-synthetics data

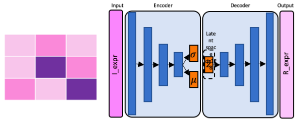
Decision tree

Unsupervised

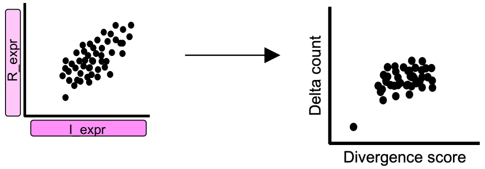
Real data

Unsupervised phase – gene expression classification

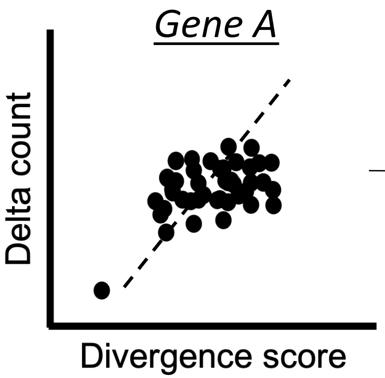
1 **To use VAE to generate reconstructed denoised counts**



2 **To compute metrics to assess the reconstruction fidelity**

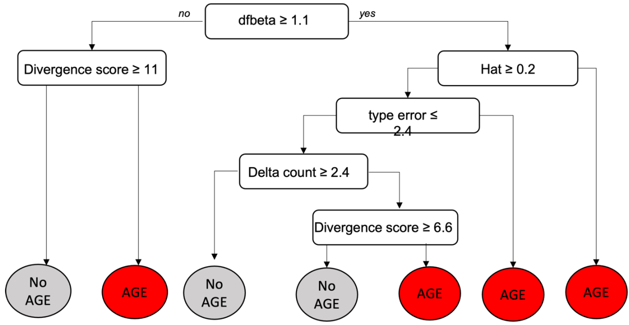


3 Classification of gene expressions as AGEs or no AGEs



Linear regression

Parameters calculated on each linear regression



dfbeta ≥ 1.1

no

yes

Divergence score ≥ 11

Hat ≥ 0.2

type error ≤ 2.4

Delta count ≥ 2.4

Divergence score ≥ 6.6

No AGE

AGE

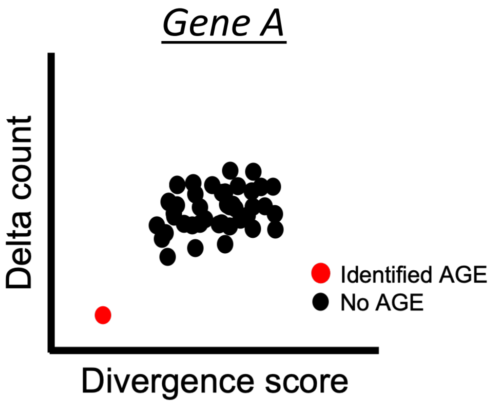
No AGE

No AGE

AGE

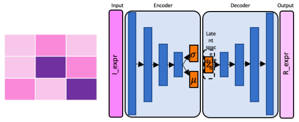
AGE

AGE

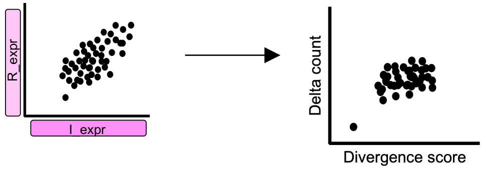


Unsupervised phase – gene expression classification

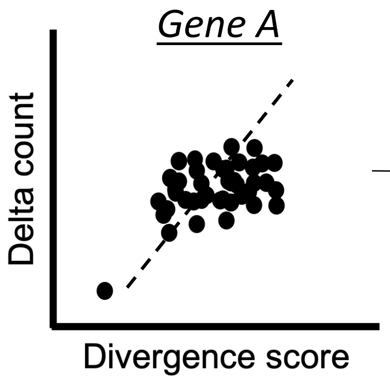
1 To use VAE to generate reconstructed denoised counts



2 To compute metrics to assess the reconstruction fidelity

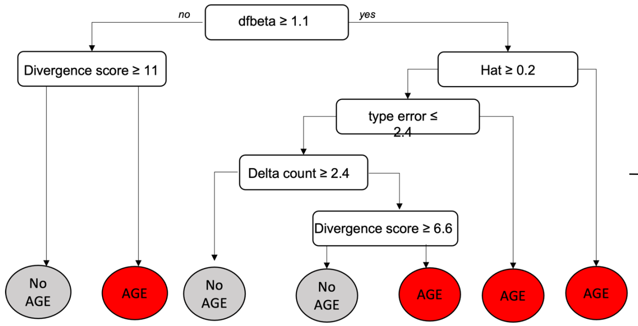


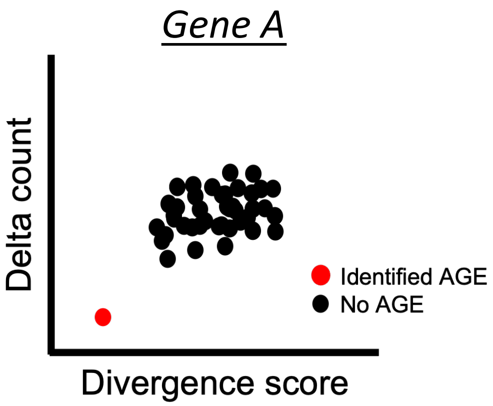
3 Classification of gene expressions as AGEs or no AGEs



Linear regression

Parameters calculated on each linear regression

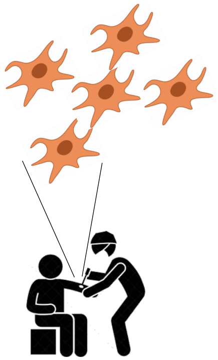




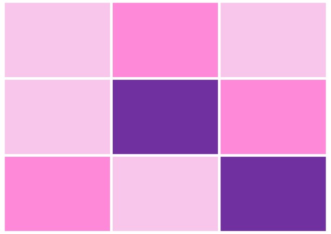
To be done for each gene

Case study

119 patients with MD suspicion
(from Kremer et al. *Nat Comm* 2017)



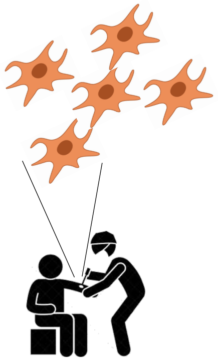
RNA-seq



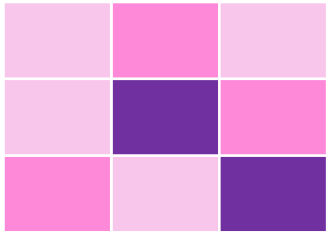
Validation of 5 candidate genes in 6 patients

Case study

119 patients with MD suspicion
(from Kremer et al. *Nat Comm* 2017)



RNA-seq



Validation of 5 candidate genes in 6 patients

Goal : Compare ABEILLE to other methods

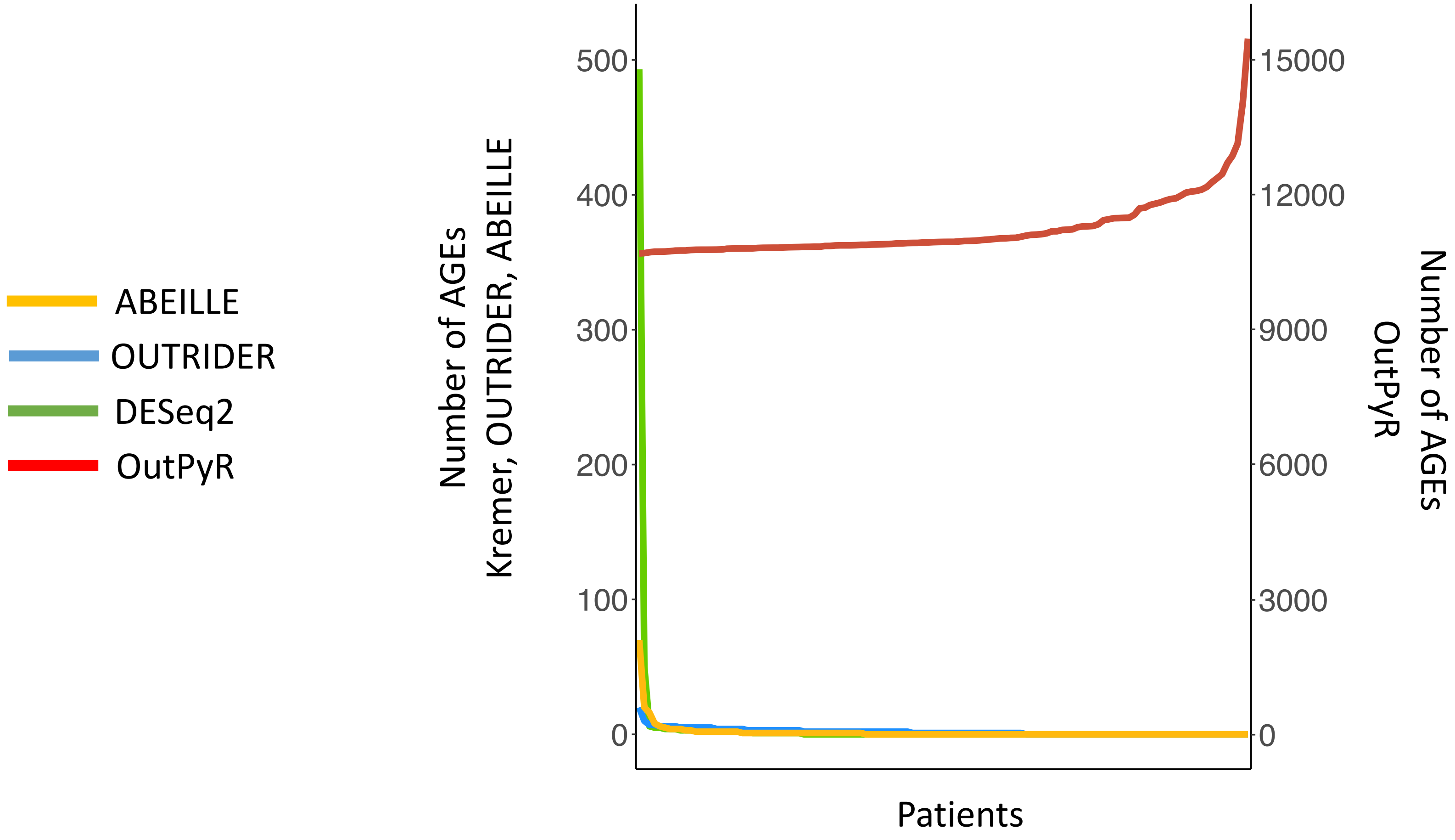
🟨 ABEILLE

🟦 OUTRIDER

🟥 OutPyR

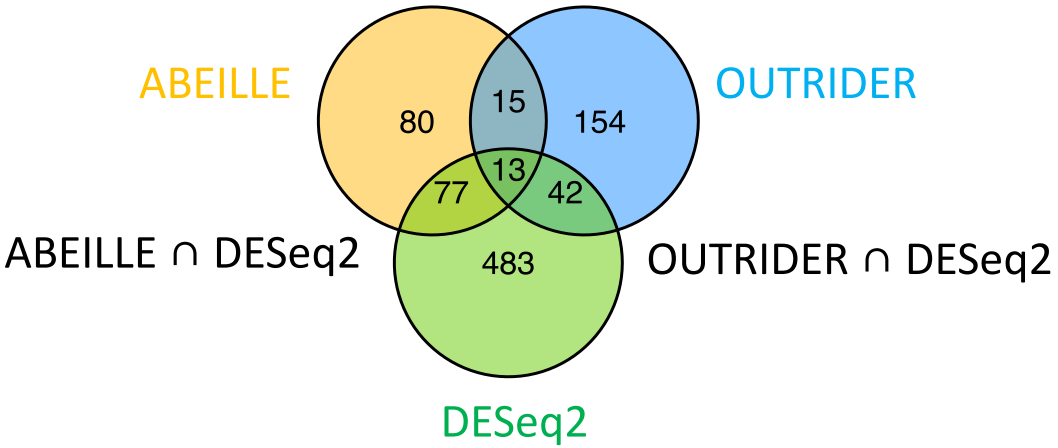
🟩 DESeq2

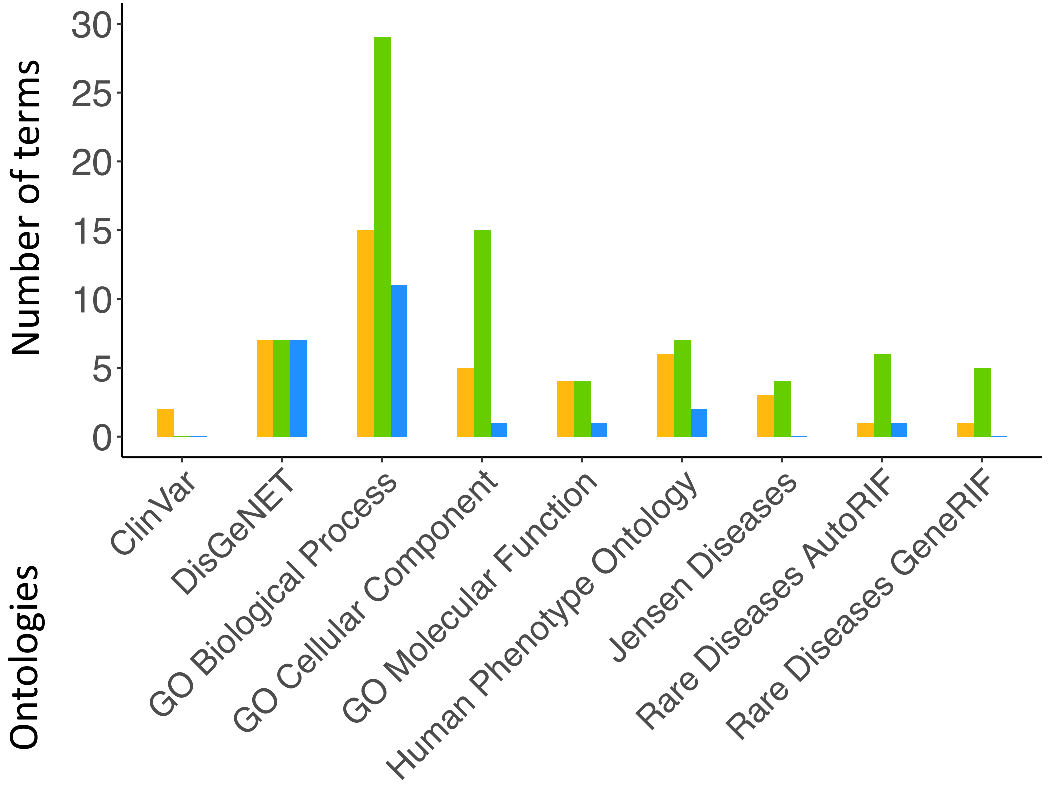
Performances of the four tools on real dataset



These observations rule out OutPyR as a tool for AGE identification in this context.

Performances of ABEILLE and OUTRIDER



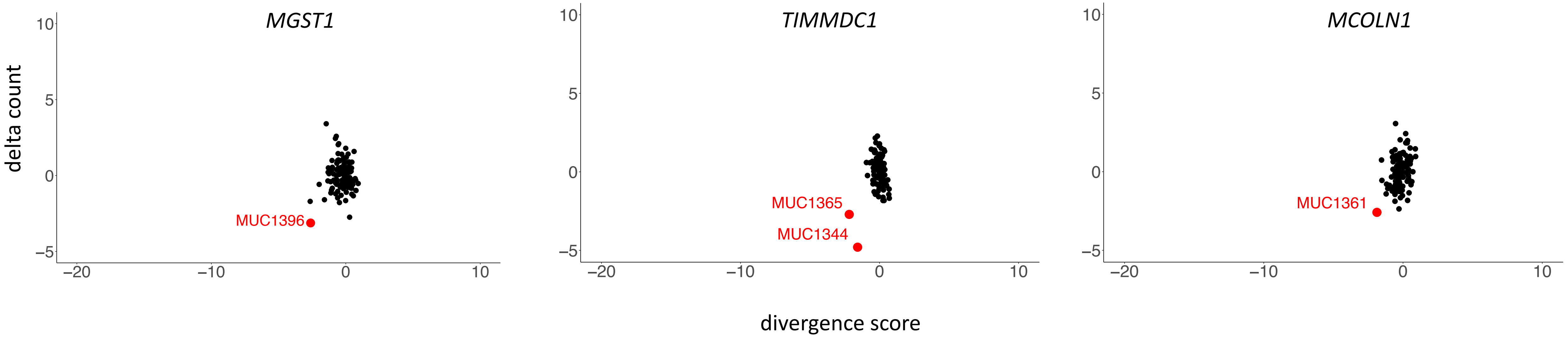


AGEs found by ABEILLE are more enriched in terms related to mitochondrial biology than the AGEs found by OUTRIDER.

Validated pathogenic genes

Validated pathogenic genes	Detected by	ABEILLE	OUTRIDER
MGST1	AGE	✓	✓
TIMMDC1	AGE	✓	✓
MCOLN1	AGE	✓	✓

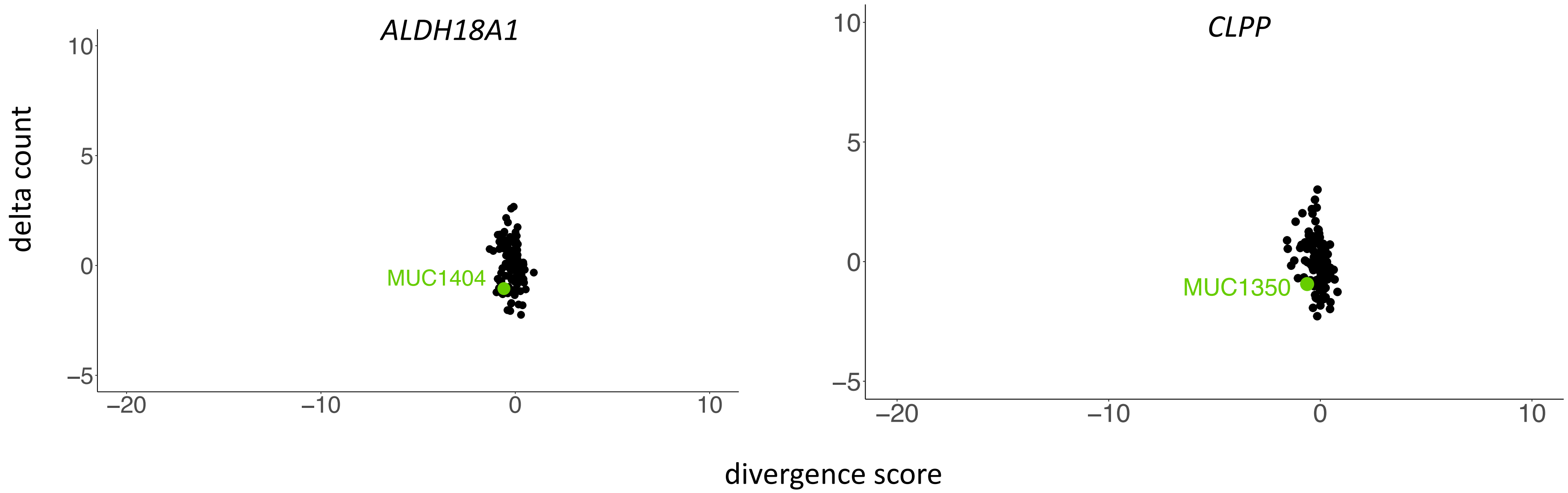
ABEILLE & OUTRIDER correctly classify the pathogenic genes as AGEs



Validated pathogenic genes

Validated pathogenic genes	**Detected by**	**ABEILLE**	**OUTRIDER**
ALDH18A1	MAE	x	✓
CLPP	AS	x	✓

OUTRIDER classifies as AGEs two pathogenic genes that do not show aberrant expression (putative false positives)



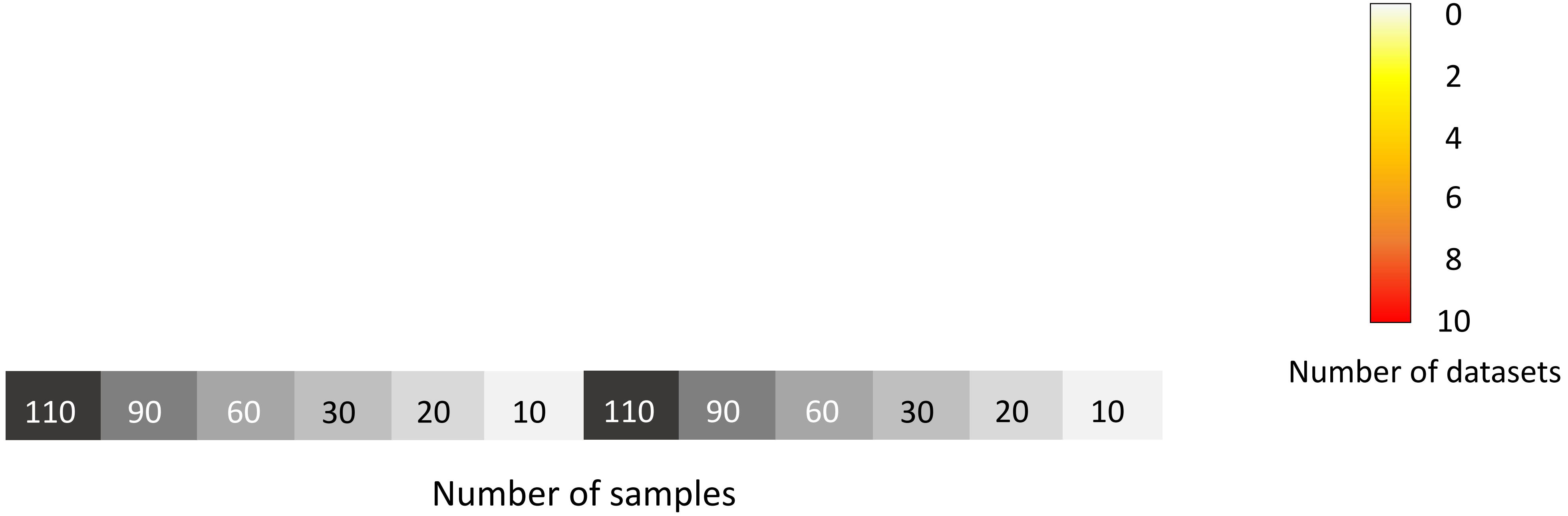
AGE detection on small dataset size

AGE detection on small dataset size

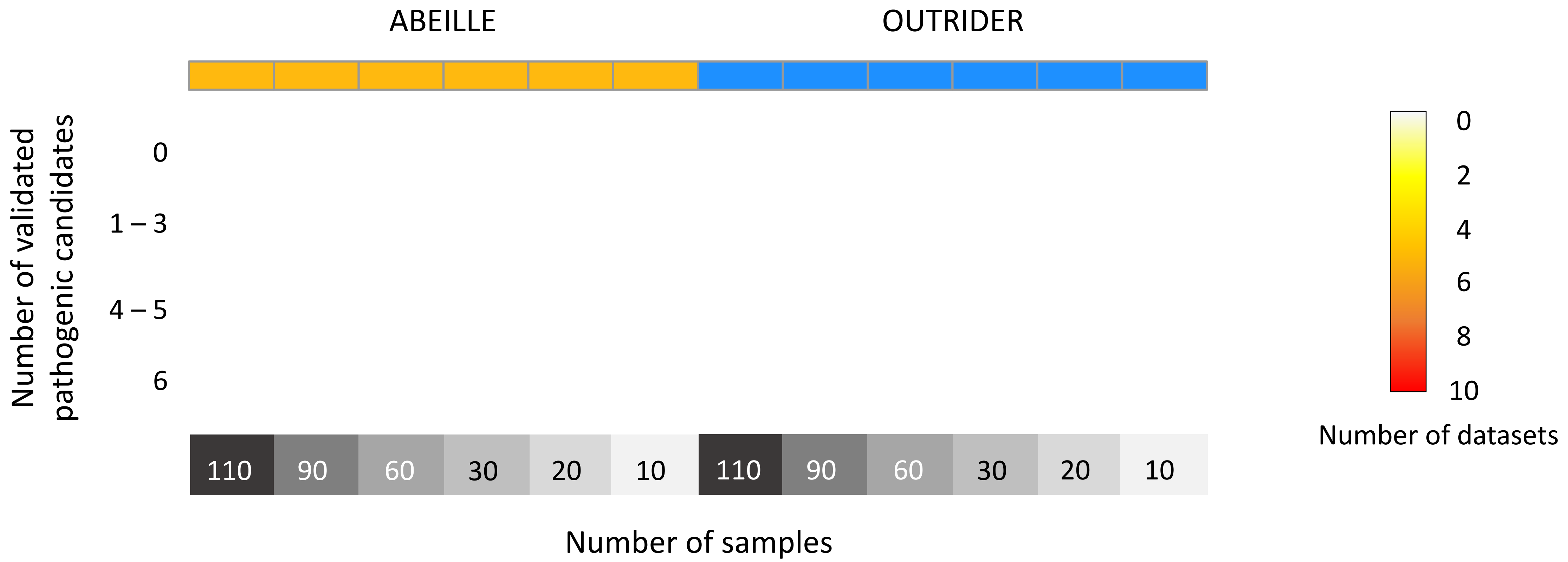
110	90	60	30	20	10	110	90	60	30	20	10

Number of samples

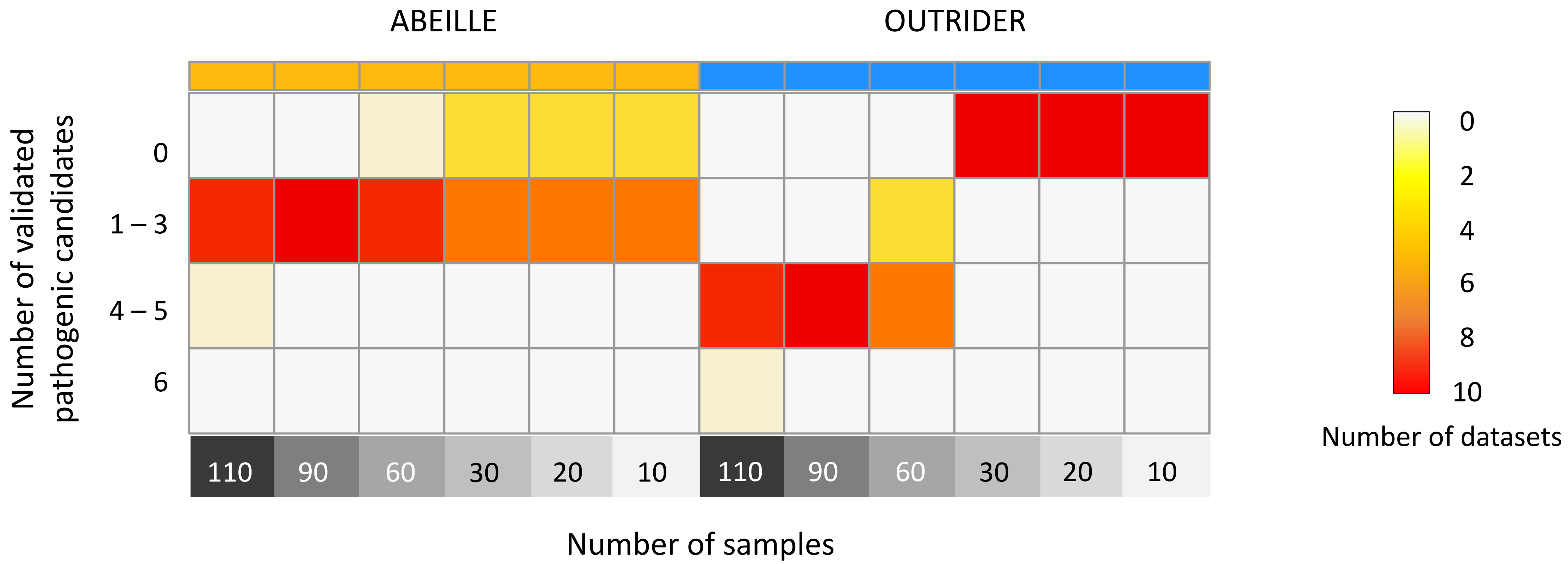
AGE detection on small dataset size



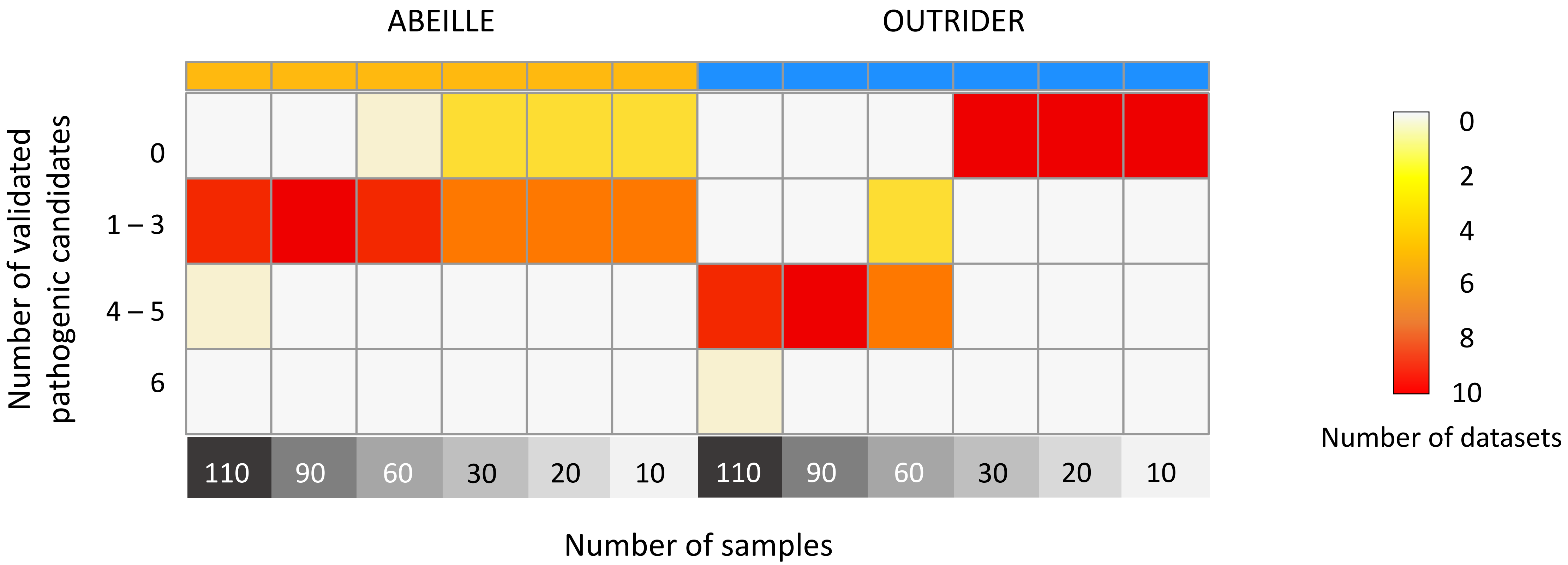
AGE detection on small dataset size



AGE detection on small dataset size



AGE detection on small dataset size



The performances of ABEILLE do not depend on the number of samples

Conclusion of part 1

ADVANTAGES

- ABEILLE identifies AGEs from RNA-seq data without the need of replicates
- ABEILLE showed good performances on small datasets

LIMITATIONS

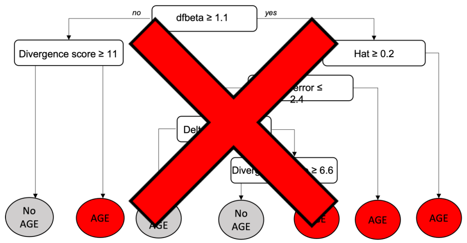
- The decision tree must be trained for each different type of data
- The choice of semi-synthetics datasets to feed the decision tree

PERSPECTIVES

- Use a flexible model to work on any type of data

Perspectives

We are developing a version 2 of ABEILLE :



DBSCAN : density based model

Multi-omics analyses are now possible

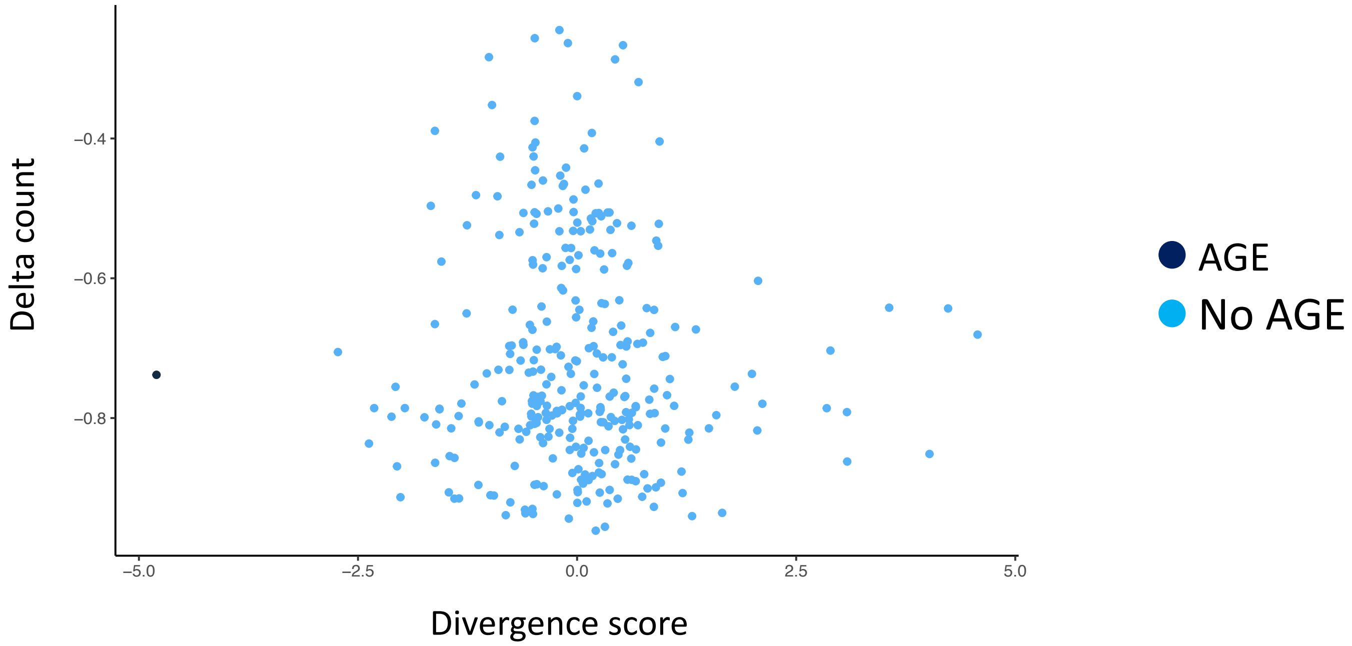
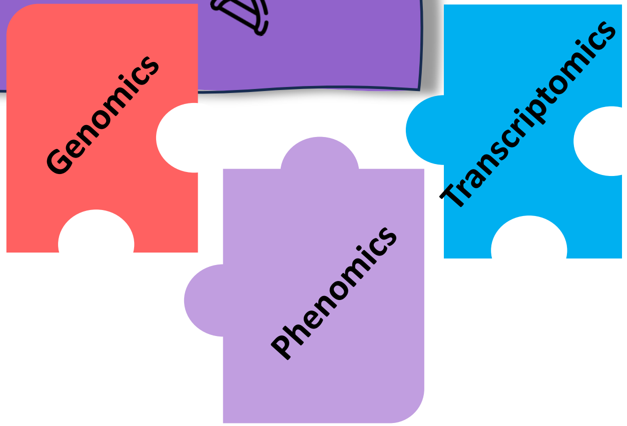


Table of contents

1 ABEILLE (ABerrant Expression Identification empLoying machine LEarning) to find candidate Aberrant Gene expression (AGEs)

2 VIOLA (Variant prIOritization using LAtent space) to find candidate pathogenic genetic variants



Diagnosis of Mitochondrial Disease (MD)

Challenges :

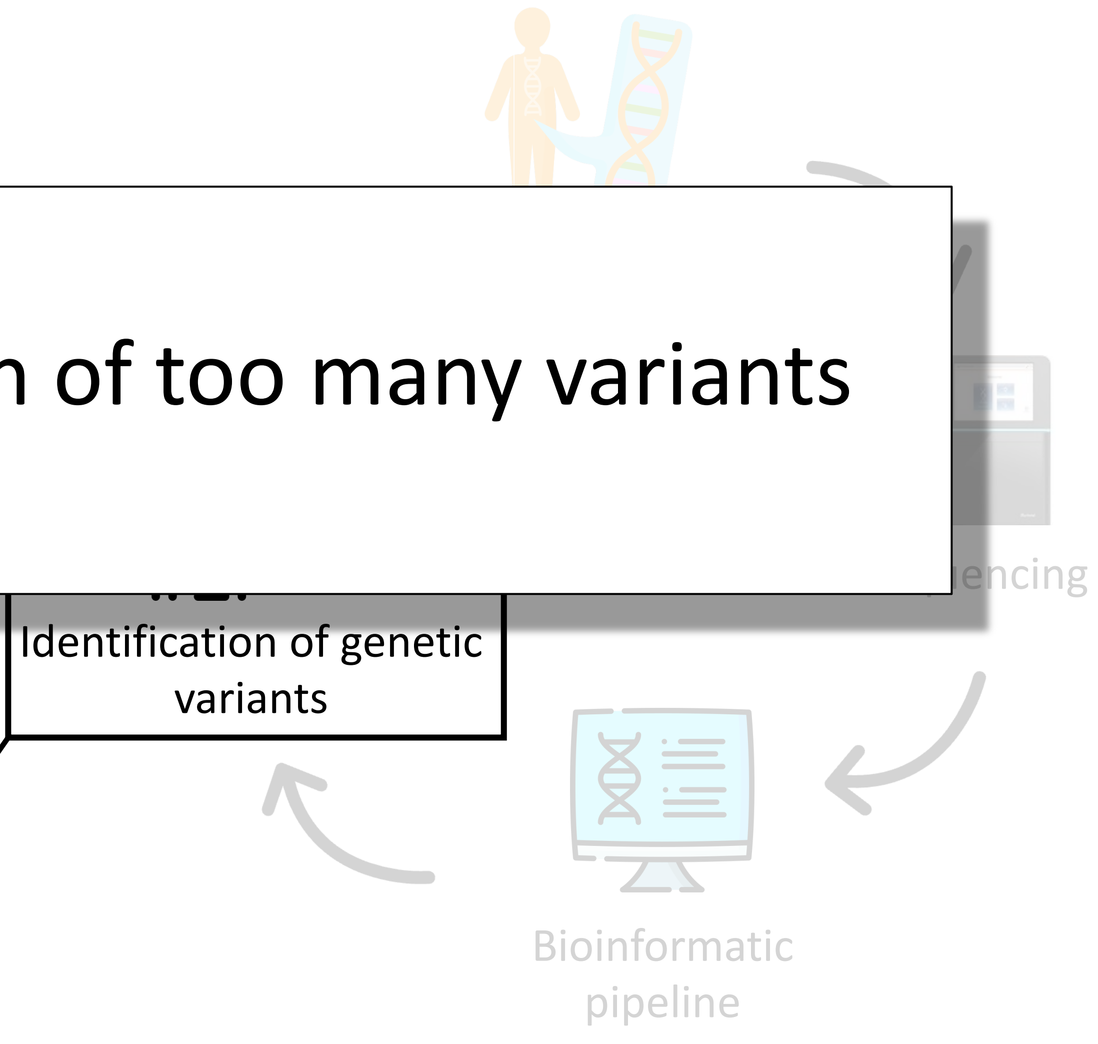
The variant can be anywhere in genetic
sequen
sequen

The va
than 1

Variant poorly characterized in databases

1 different responsible variant for each patient

Identification of too many variants



Diagnosis of Mitochondrial Disease (MD)

Challenges :

The variant can be anywhere in genetic sequen

sequen

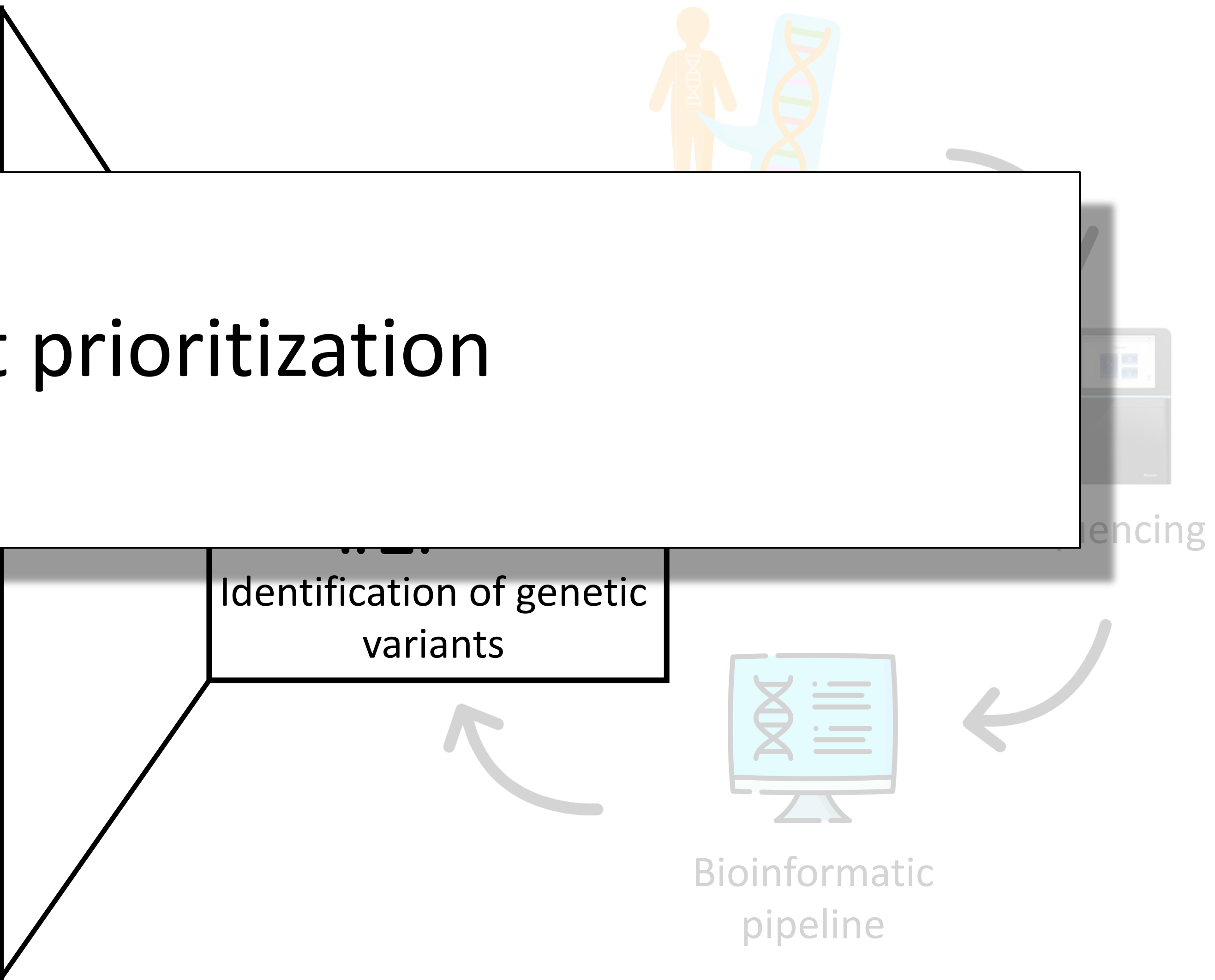
The va

than 1

Variant poorly characterized in databases

1 different responsible variant for each patient

Variant prioritization



Variant prioritization

Process of **selecting** and **ranking** genetic variants based on their potential **significance** or relevance to a specific **phenotype** or condition.

State-of-the-art tool : EXOMISER
(SOTAT)

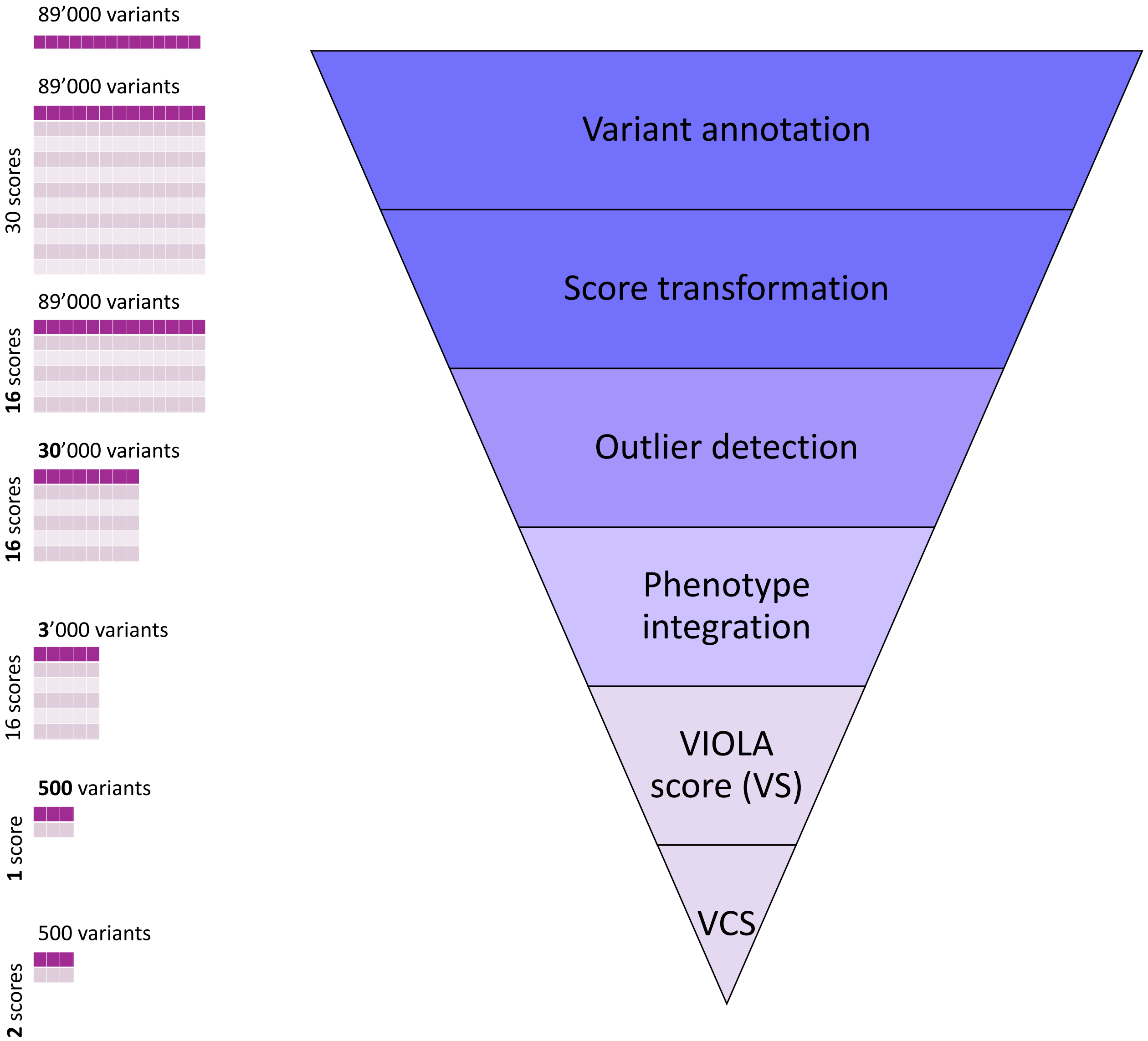
- Exomiser ranks genetic variants according to a combination of criteria :
  - variant frequency
  - predicted pathogenicity
  - known disease associations
  - conservation
  - functional impact
  - phenotypic information

- Drawbacks :
  - variants of a same gene have the same rank in Exomiser results
  - Exomiser is trained on large databases

VIOLA's hypothesis

The disease-responsible variant(s) are **patient-specific** and **rare**.
→ unique combination of properties different from the rest of the patient variants.
The putative disease variants for MD are **outliers** of each patient variants' distribution.

VIOLA’s workflow



89’000 variants

Variant annotation

89’000 variants

30 scores

Score transformation

89’000 variants

16 scores

Outlier detection

30’000 variants

16 scores

Phenotype integration

3’000 variants

16 scores

VIOLA score (VS)

500 variants

1 score

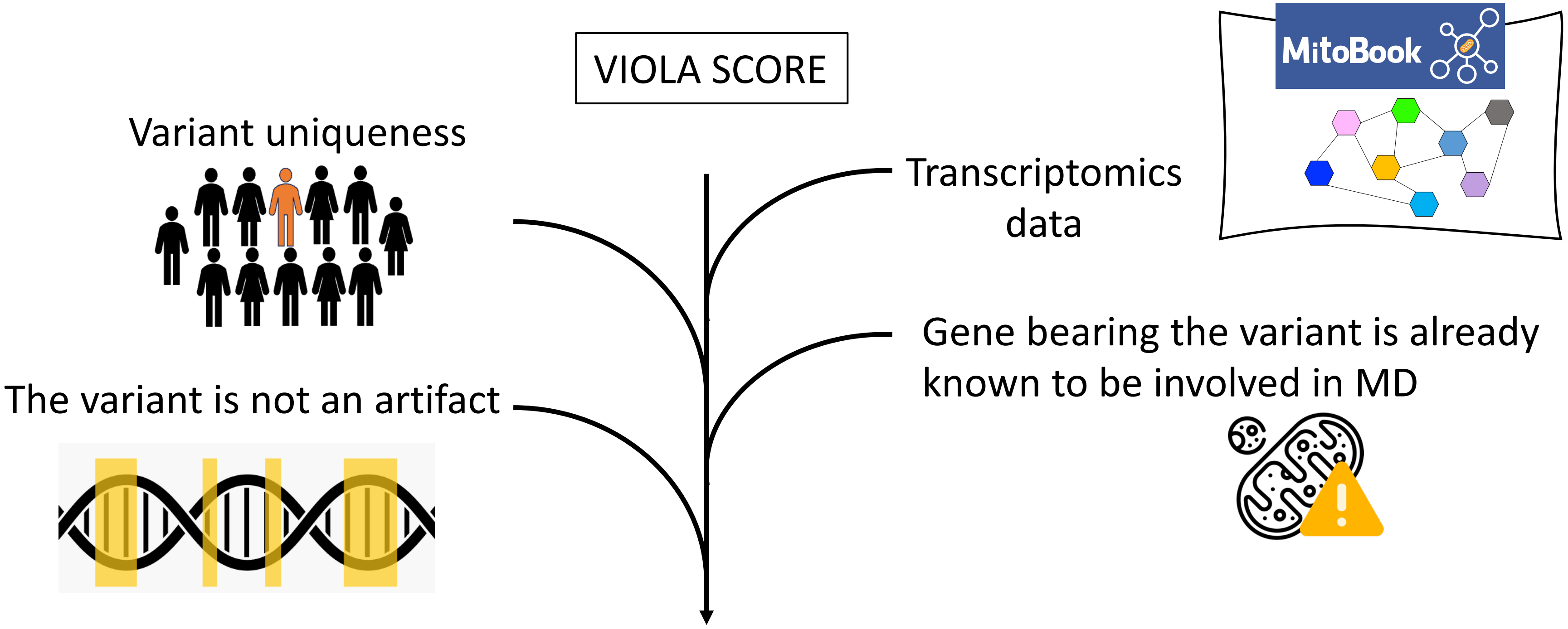
VCS

500 variants

2 scores

Creation of the VIOLA combined score (VCS)

<u>Goal</u> : Incorporate knowledge of mitochondrial diseases into VIOLA score



VIOLA SCORE

Variant uniqueness

The variant is not an artifact

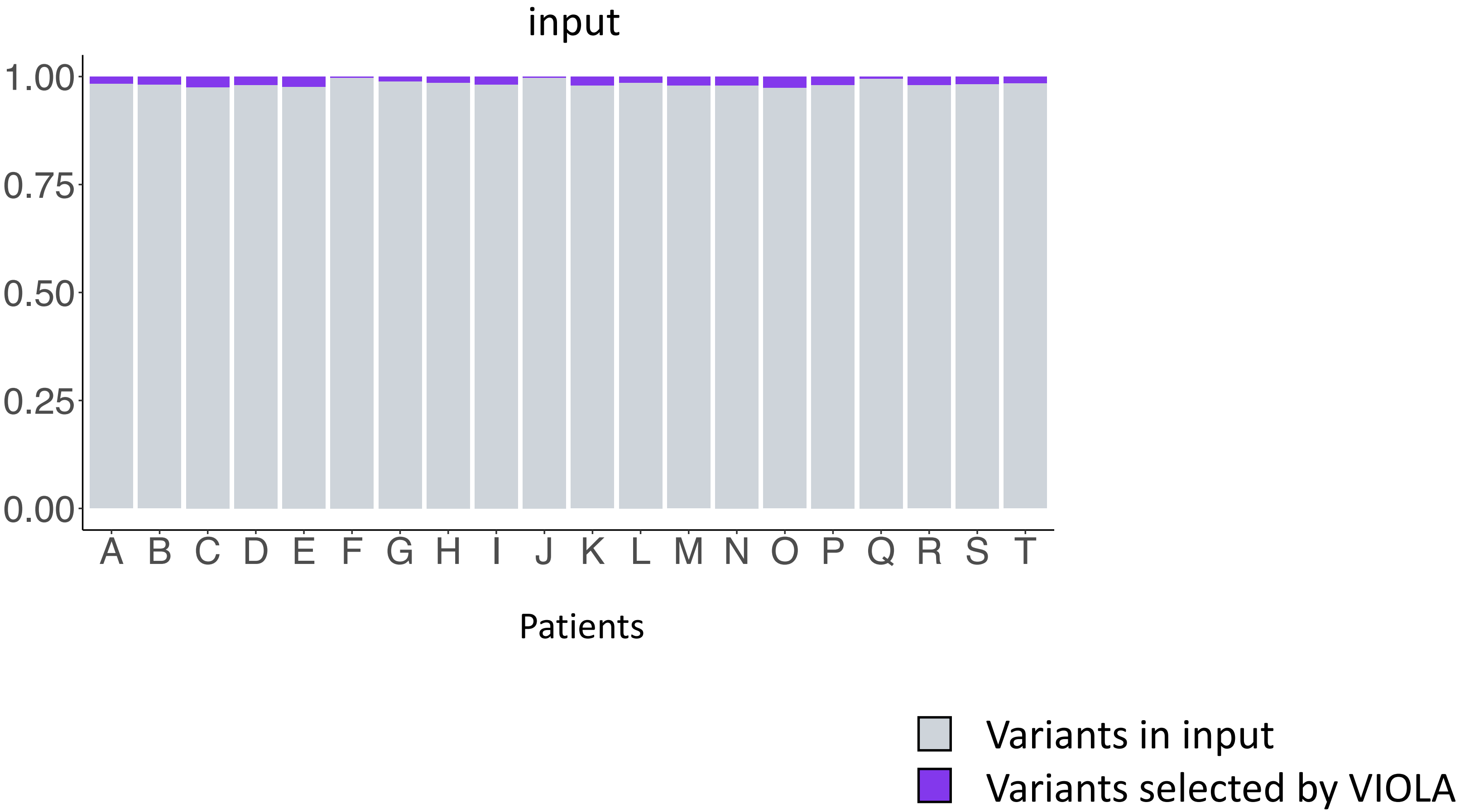
Transcriptomics data

Gene bearing the variant is already known to be involved in MD

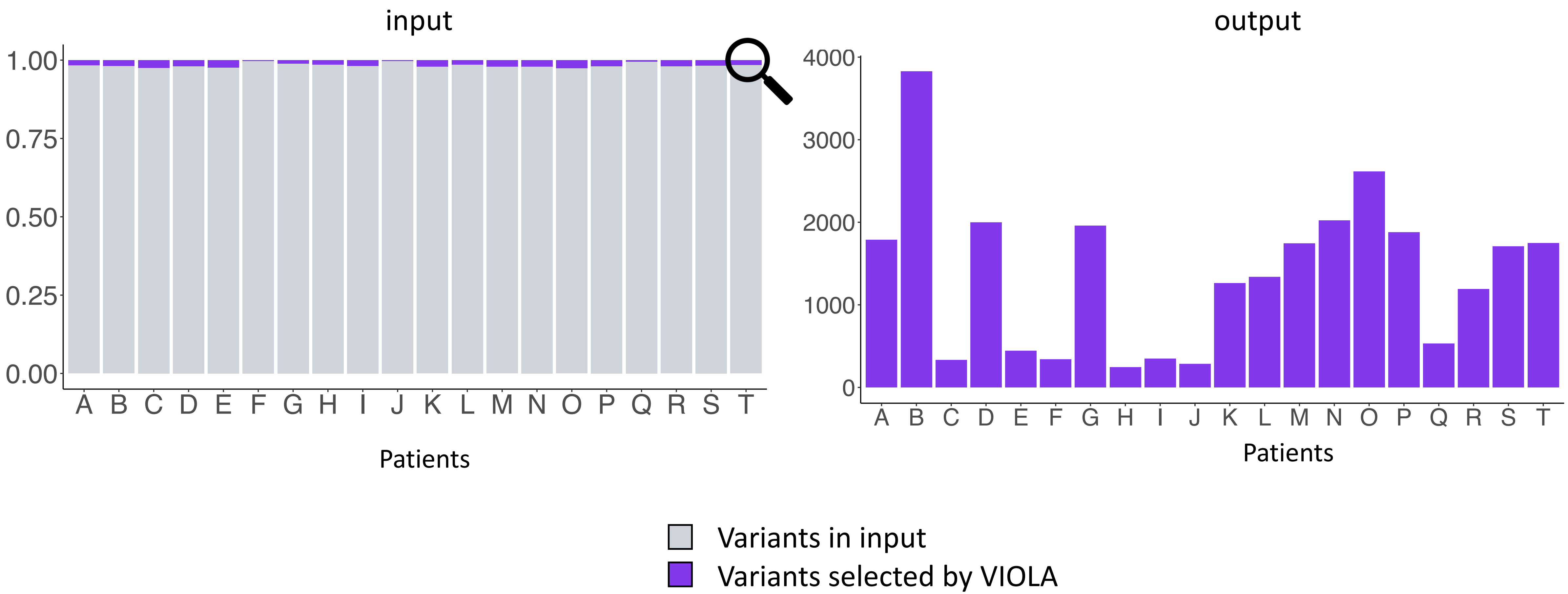
VCS = 0.5 (VS + transcriptomics + uniqueness) + 0.01 (known gene + artifact)

Results on in-house cohort

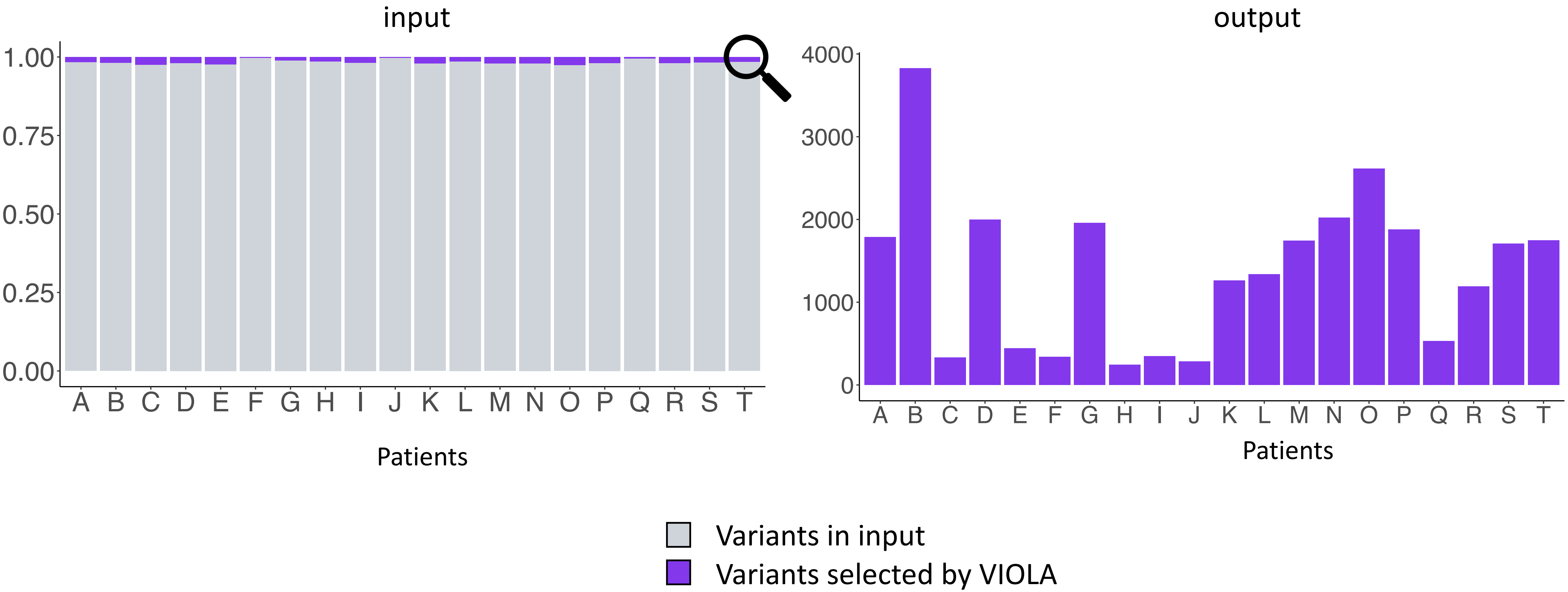
VIOLA results on the other patients



VIOLA results on the other patients



VIOLA results on the other patients



VIOLA selects 1% of input variants as potential candidates for MD

Can VIOLA find the responsible variant for positive patients?

Ranking

🟧 1st - 10th

🟨 11th - 40th

🟦 41th and +

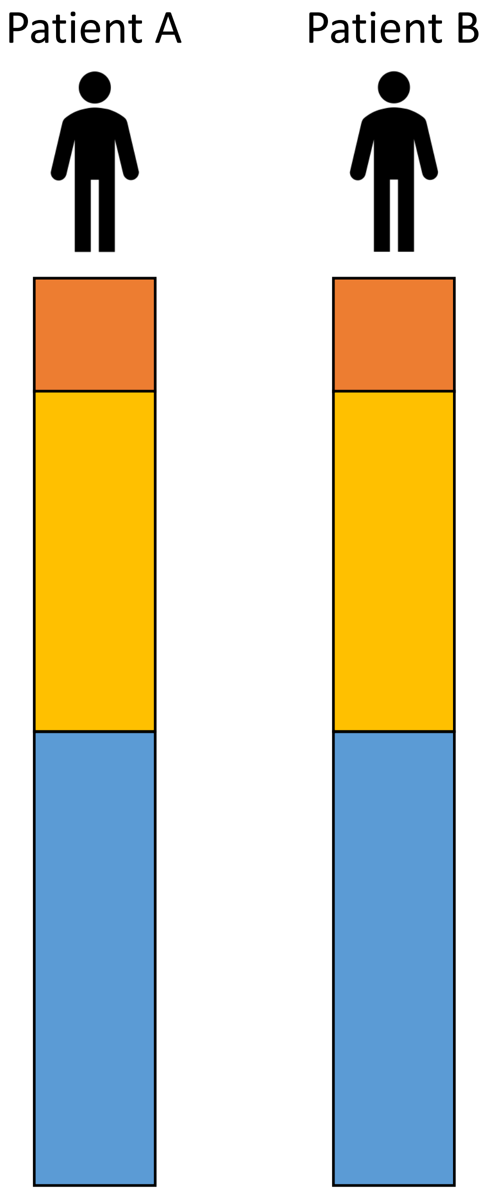
Scores

VIOLA Score ●

VIOLA Combined Score ★

State Of the Art Tool ▲

Can VIOLA find the responsible variant for positive patients?



Ranking

🟧 1st - 10th

🟨 11th - 40th

🟦 41th and +

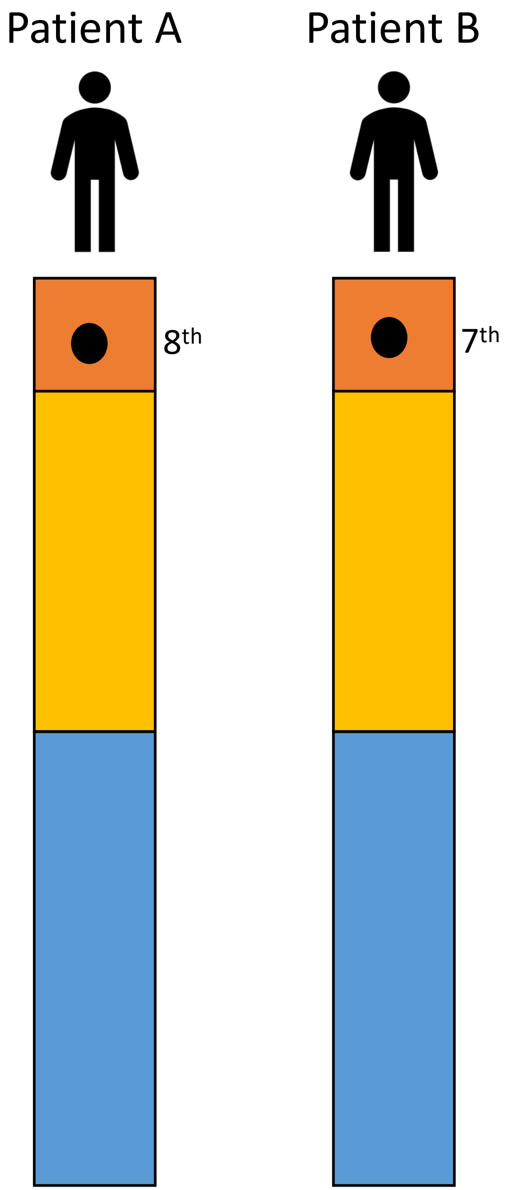
Scores

VIOLA Score ●

VIOLA Combined Score ★

State Of the Art Tool ▲

Can VIOLA find the responsible variant for positive patients?



Ranking

🟧 1st - 10th

🟨 11th - 40th

🟦 41th and +

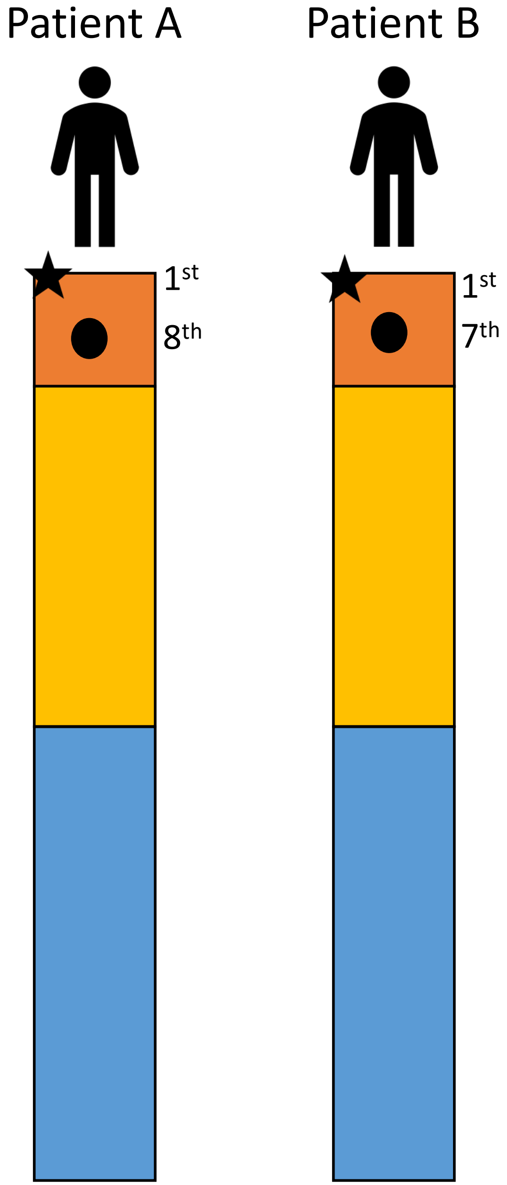
Scores

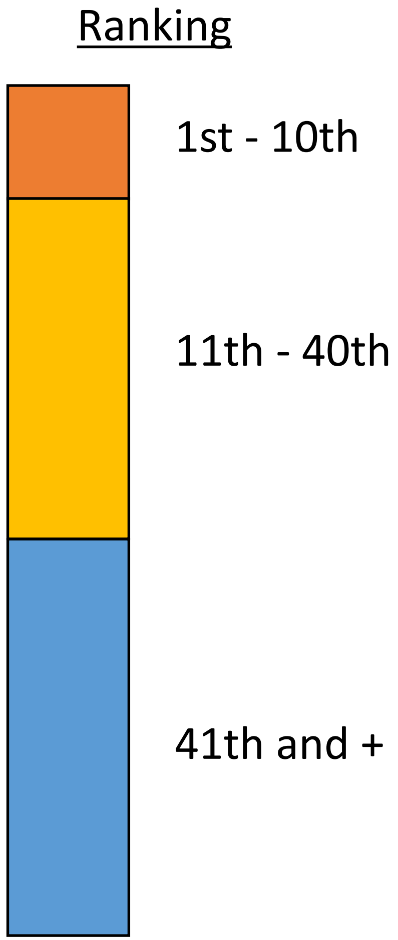
VIOLA Score ●

VIOLA Combined Score ★

State Of the Art Tool ▲

Can VIOLA find the responsible variant for positive patients?





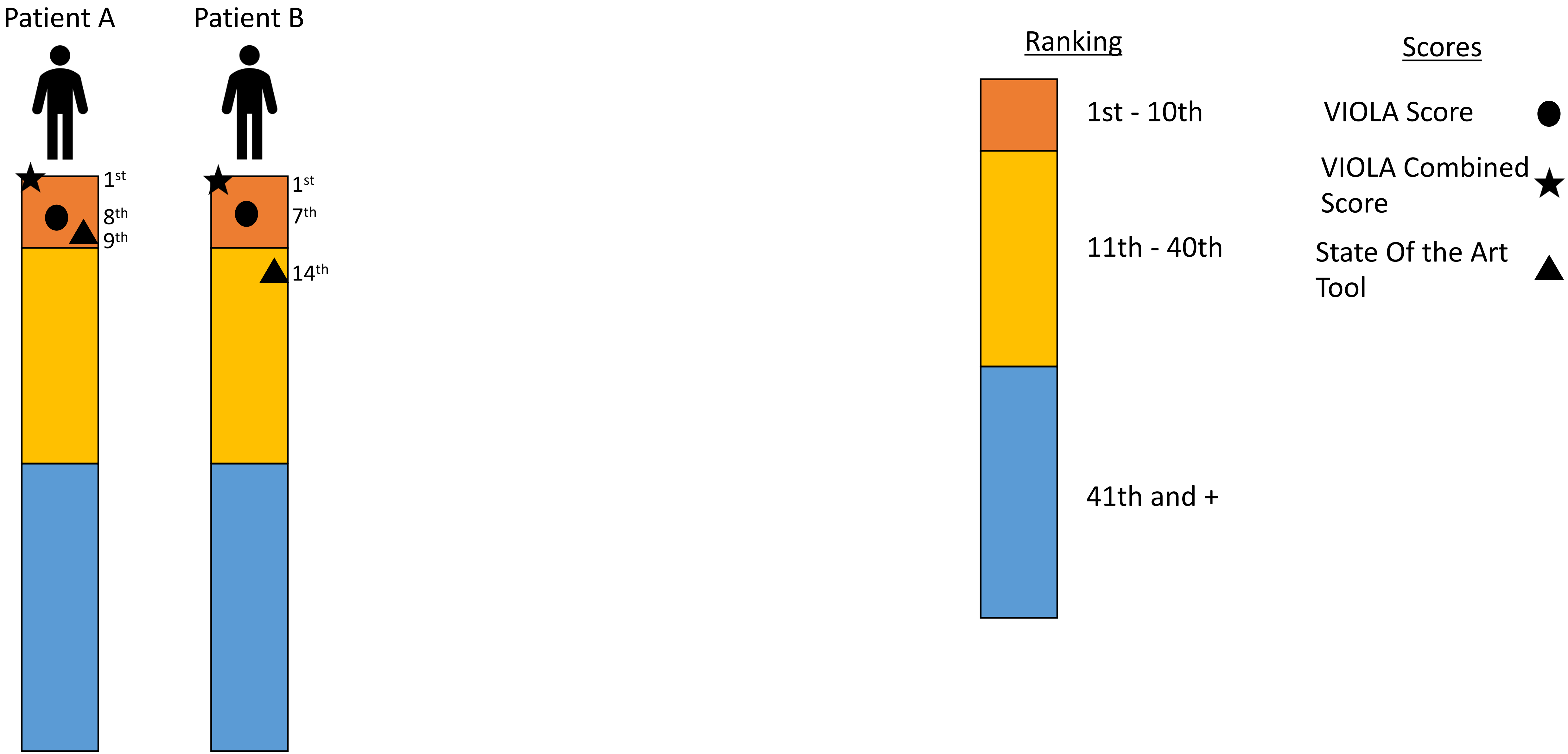
Scores

VIOLA Score ●

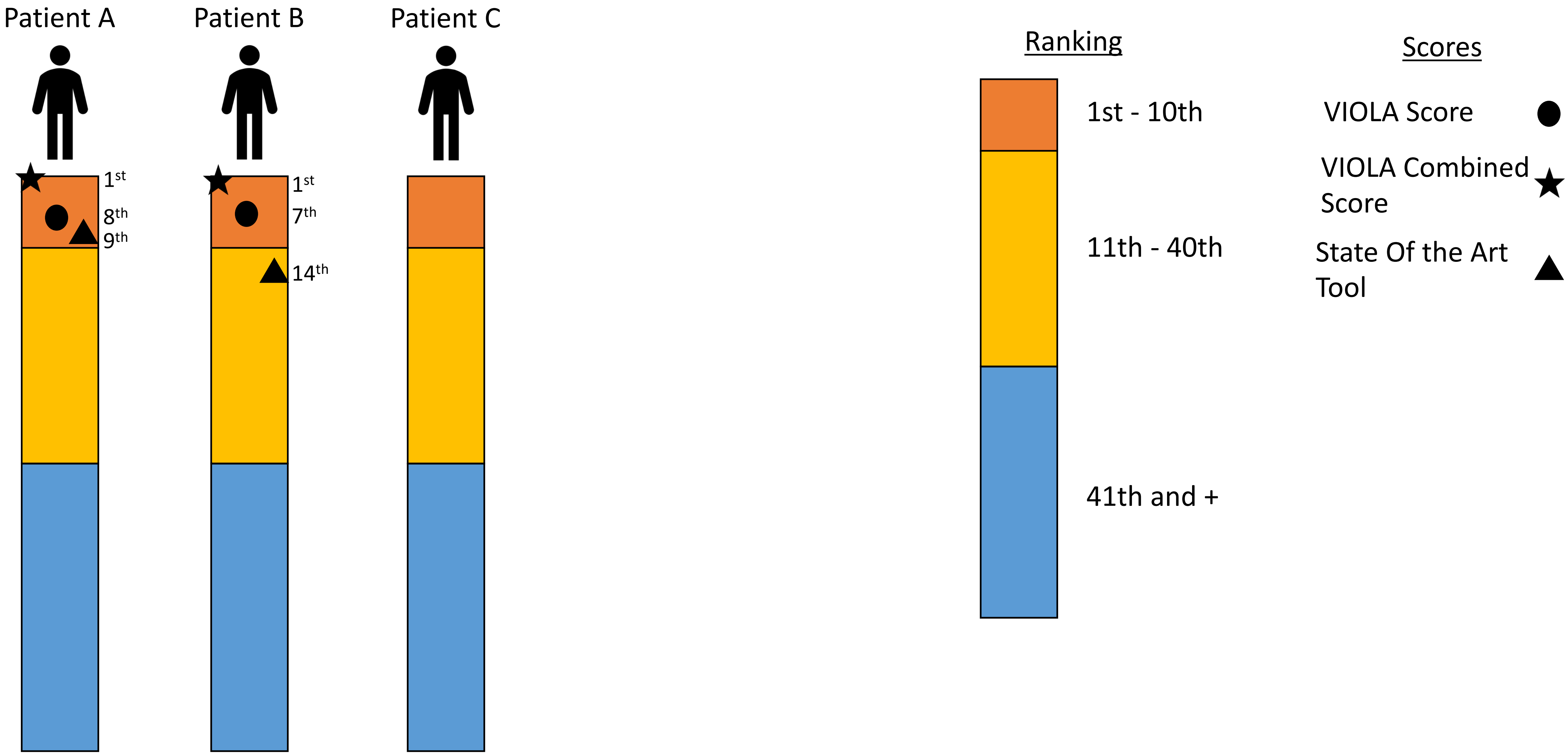
VIOLA Combined Score ★

State Of the Art Tool ▲

Can VIOLA find the responsible variant for positive patients?



Can VIOLA find the responsible variant for positive patients?



Ranking

1st - 10th

11th - 40th

41th and +

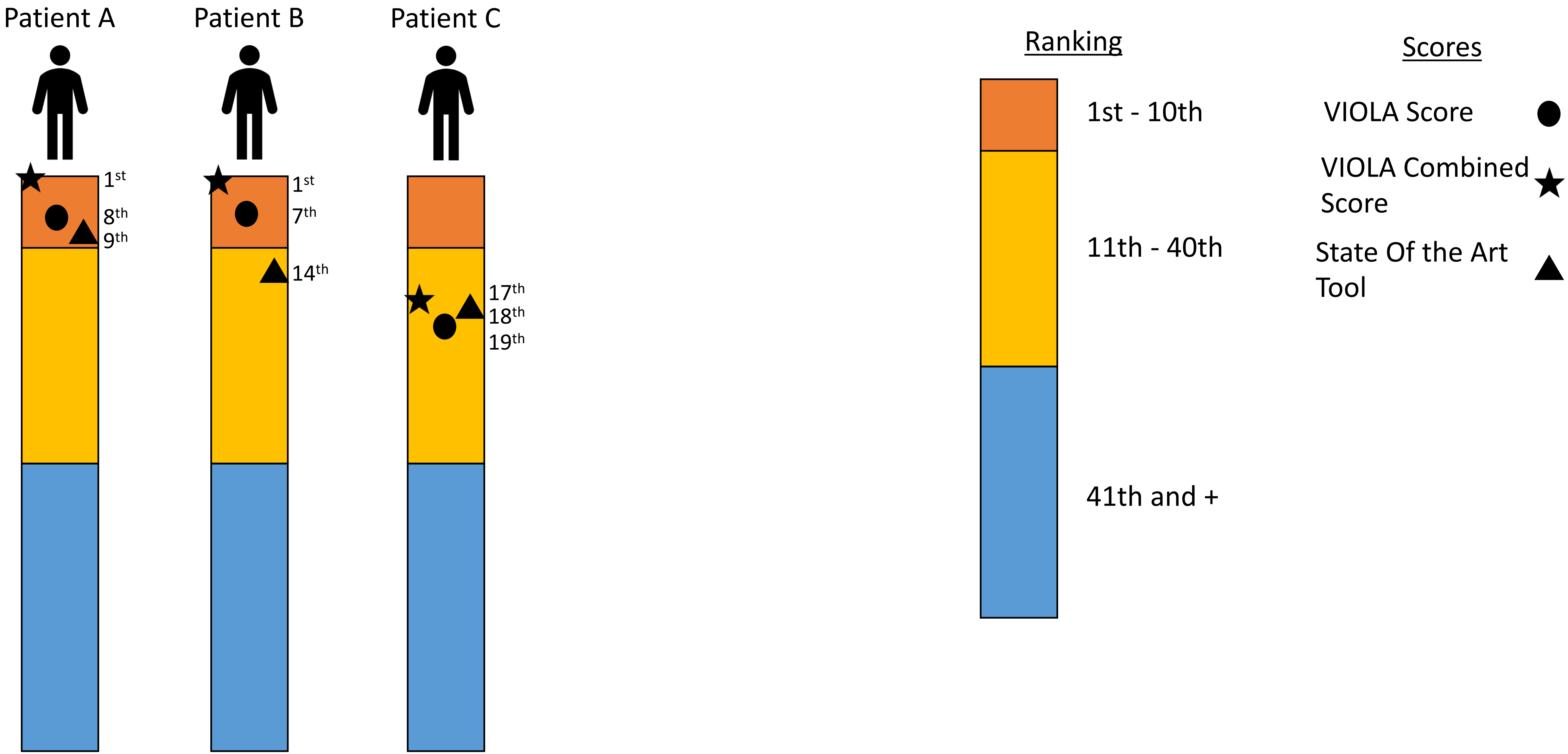
Scores

VIOLA Score ●

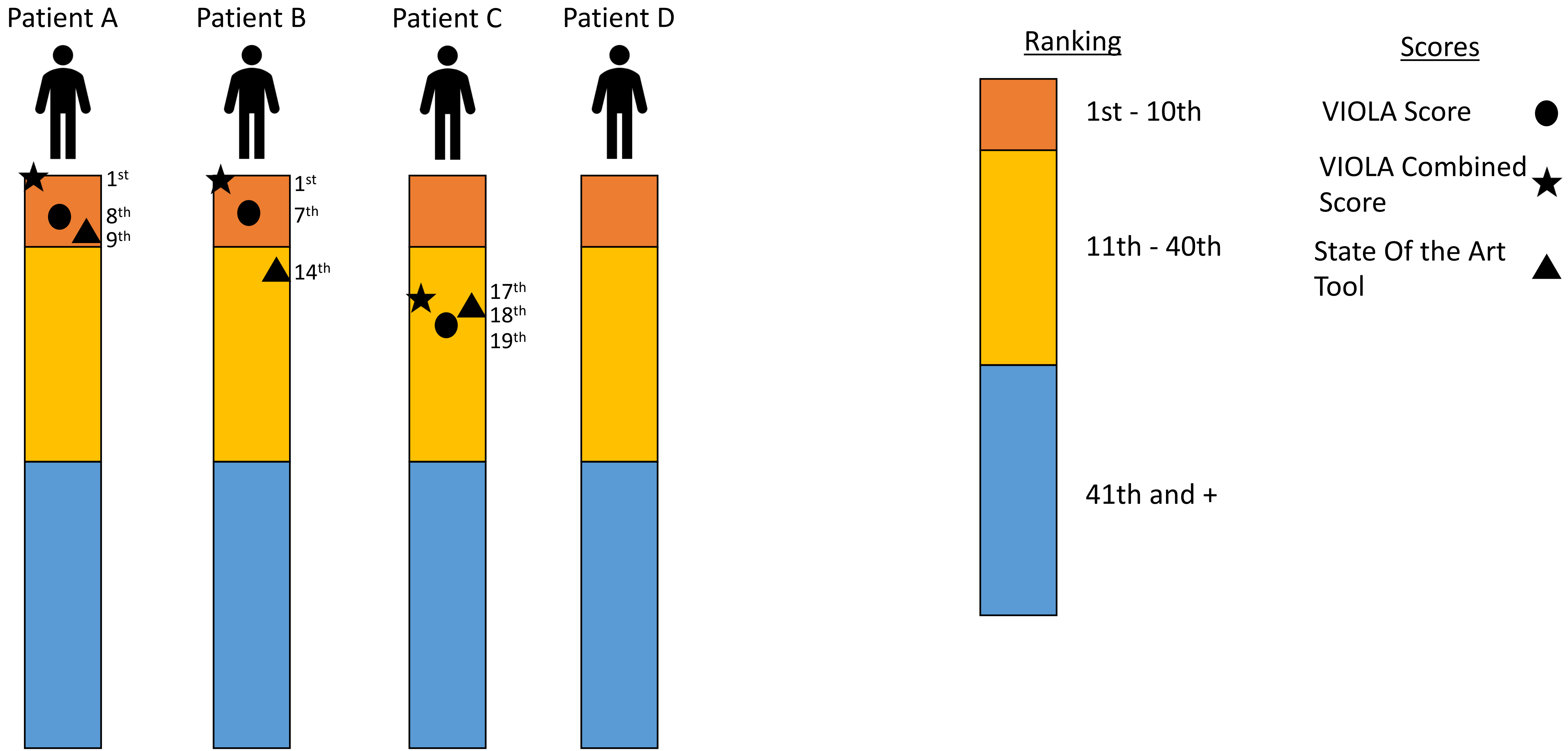
VIOLA Combined Score ★

State Of the Art Tool ▲

Can VIOLA find the responsible variant for positive patients?



Can VIOLA find the responsible variant for positive patients?



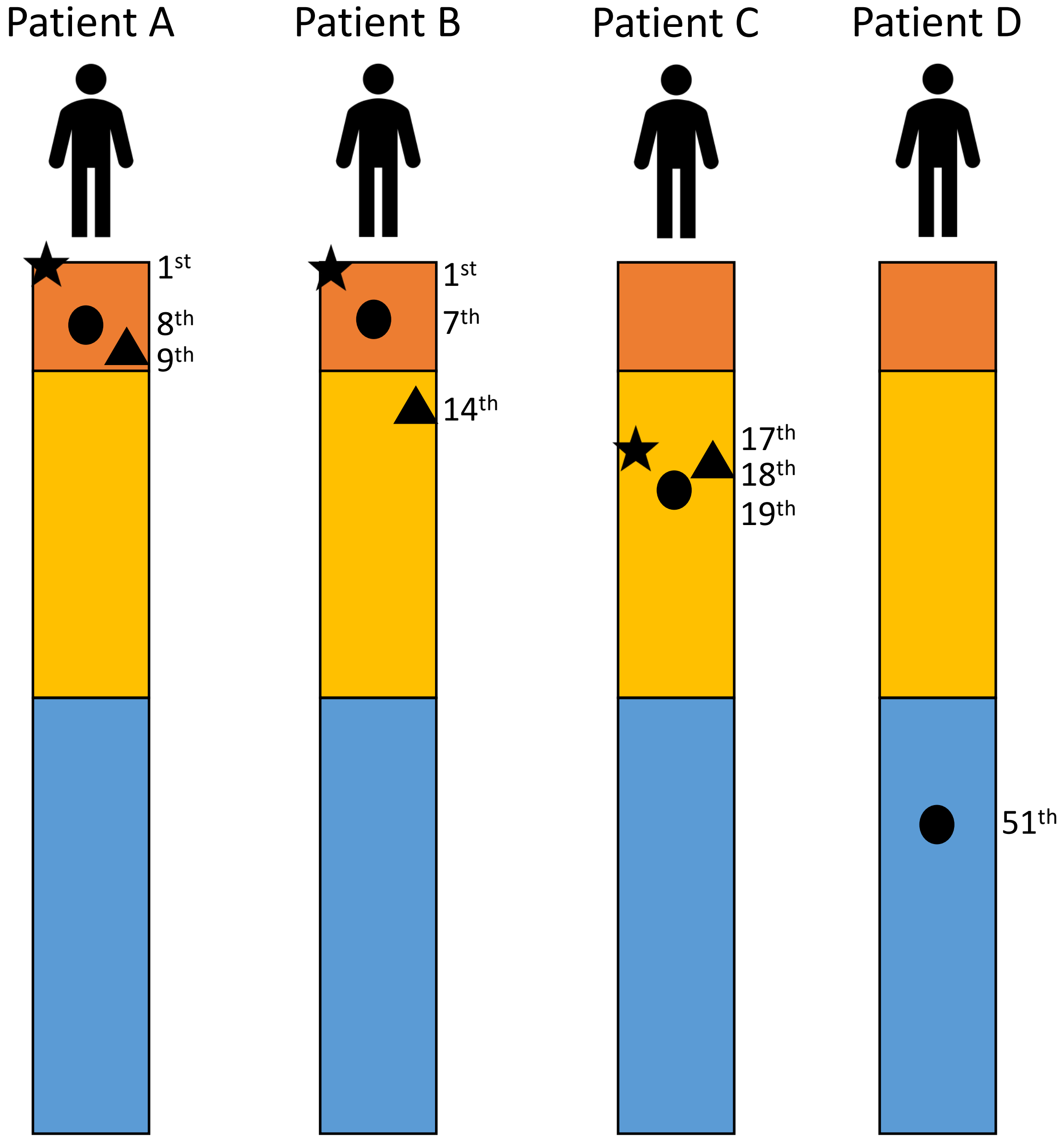
Scores

VIOLA Score ●

VIOLA Combined Score ★

State Of the Art Tool ▲

Can VIOLA find the responsible variant for positive patients?



Ranking

🟧 1st - 10th

🟨 11th - 40th

🟦 41th and +

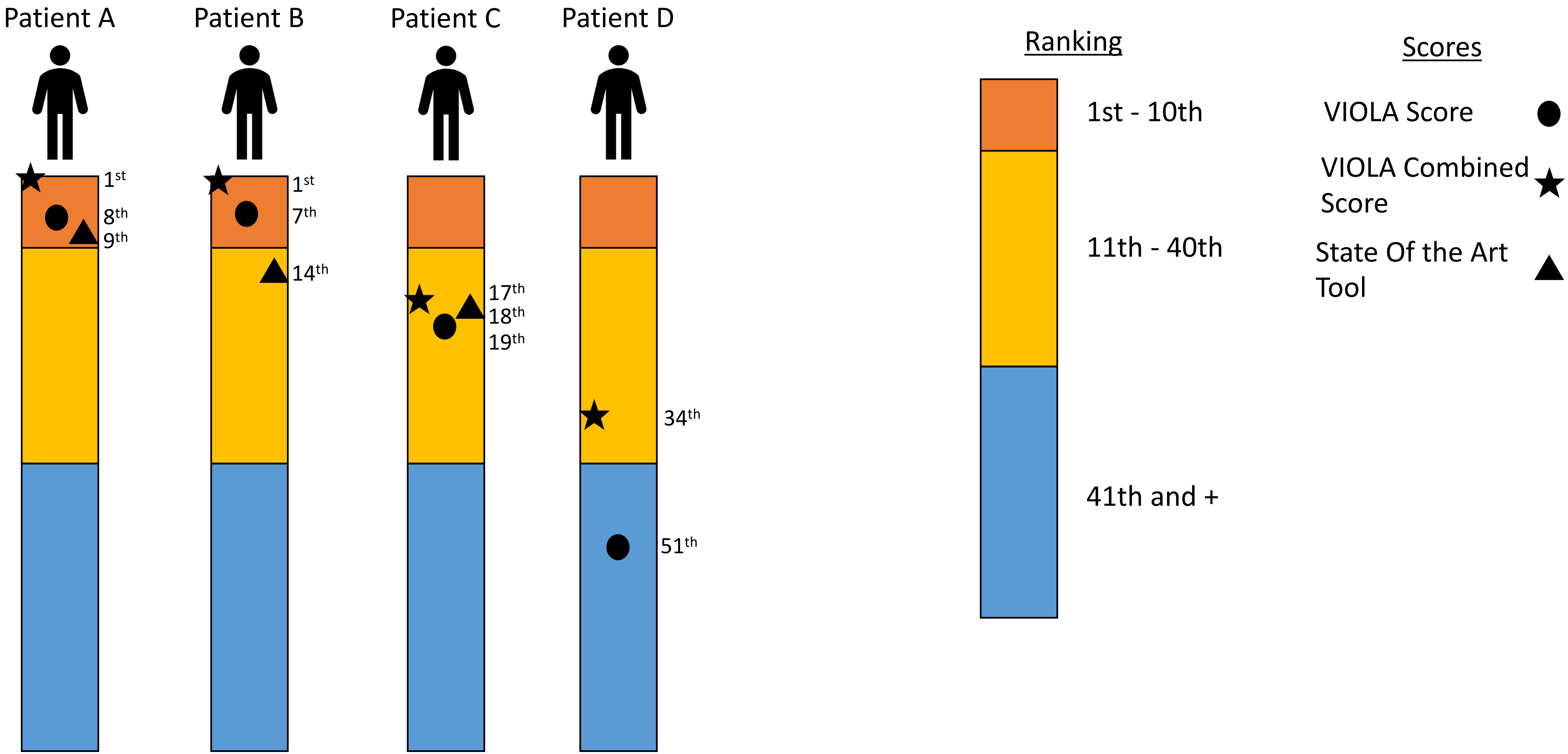
Scores

VIOLA Score ●

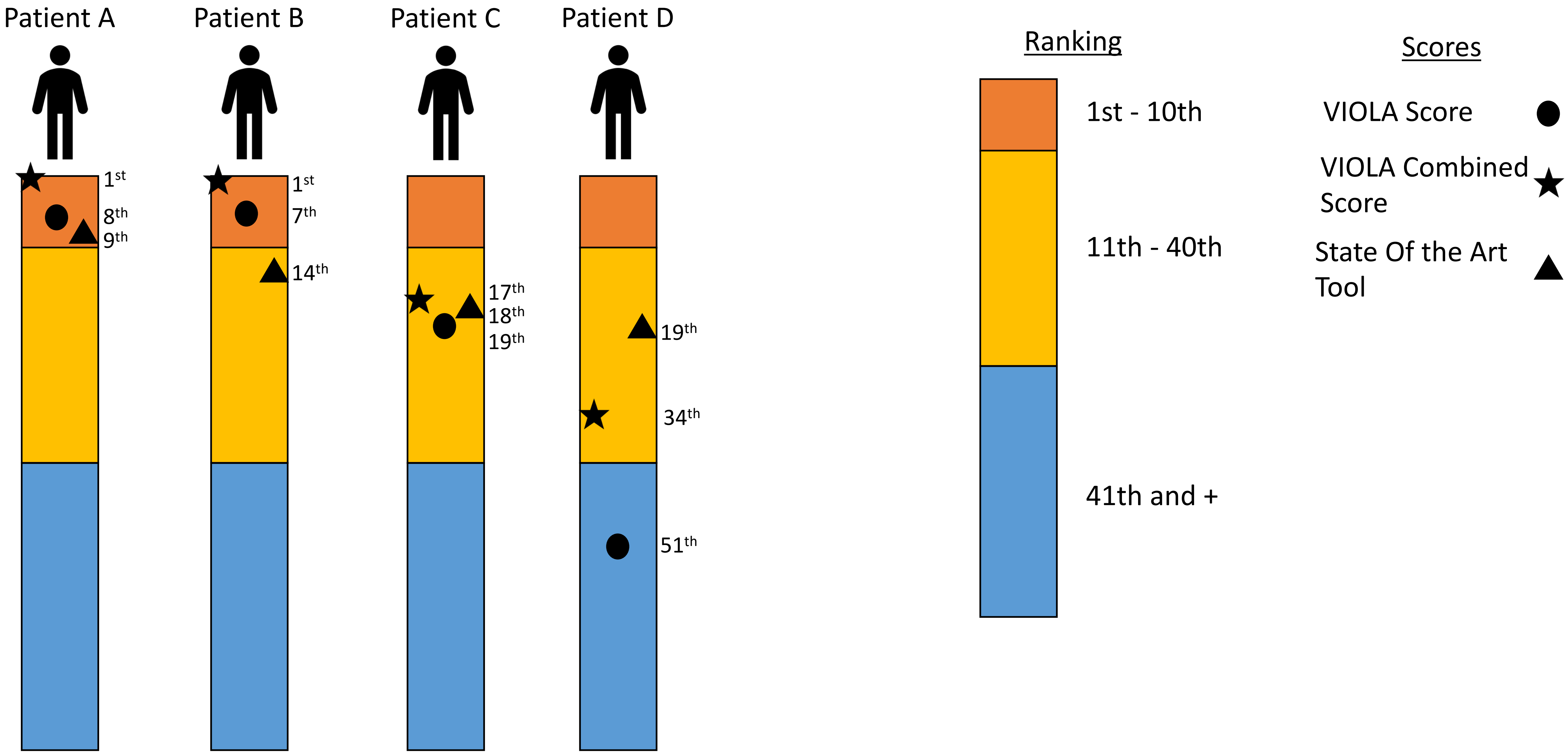
VIOLA Combined Score ★

State Of the Art Tool ▲

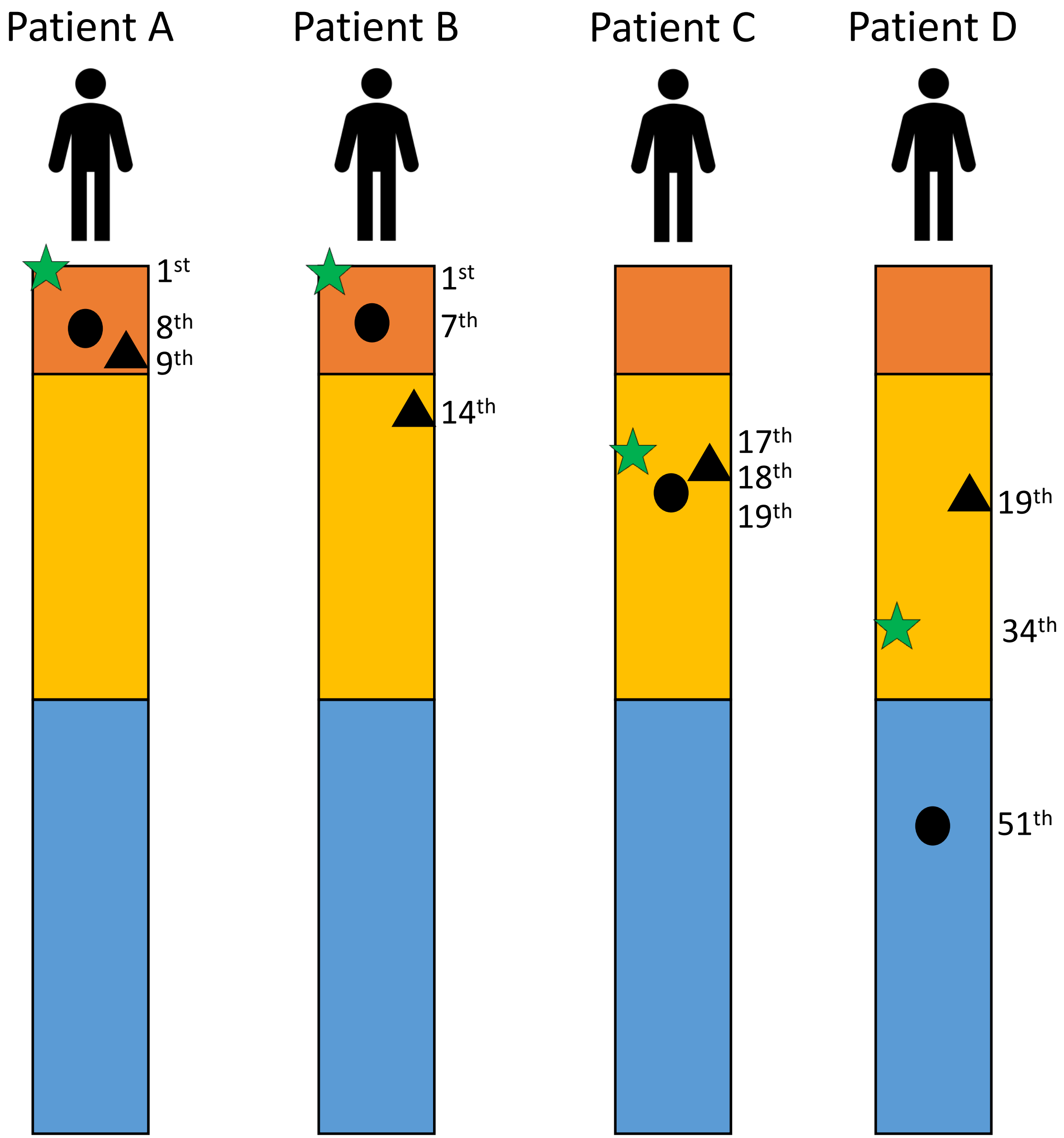
Can VIOLA find the responsible variant for positive patients?

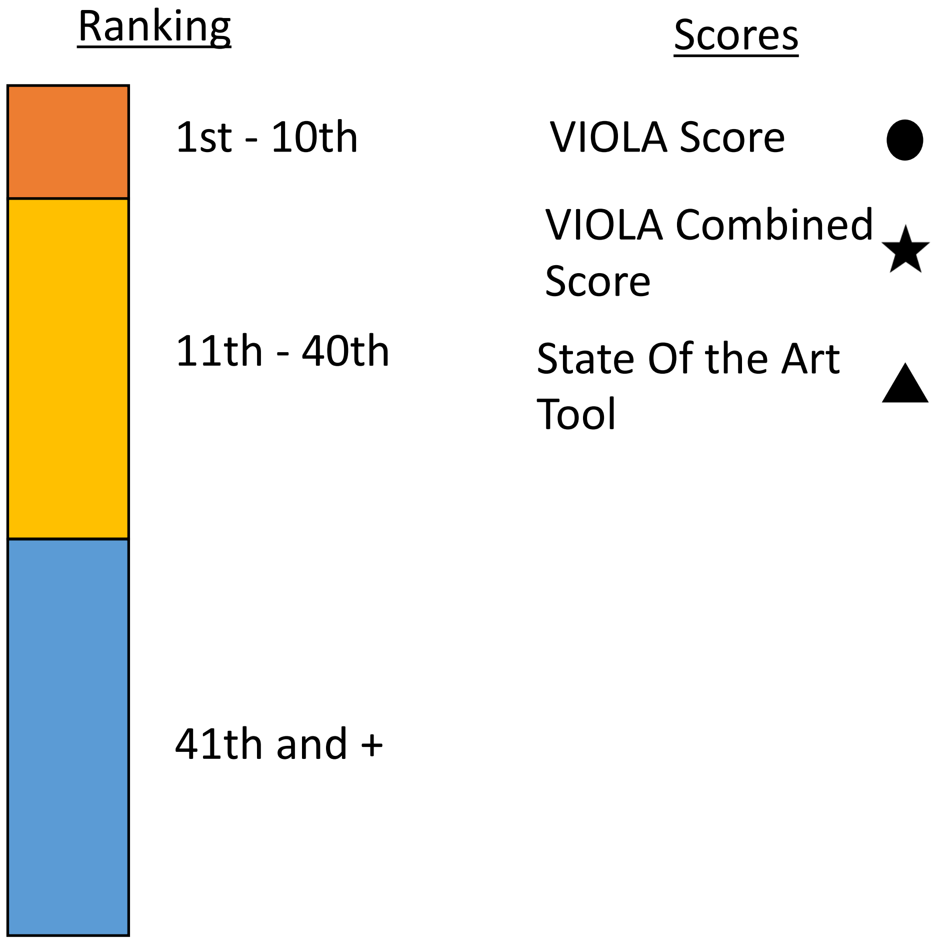


Can VIOLA find the responsible variant for positive patients?



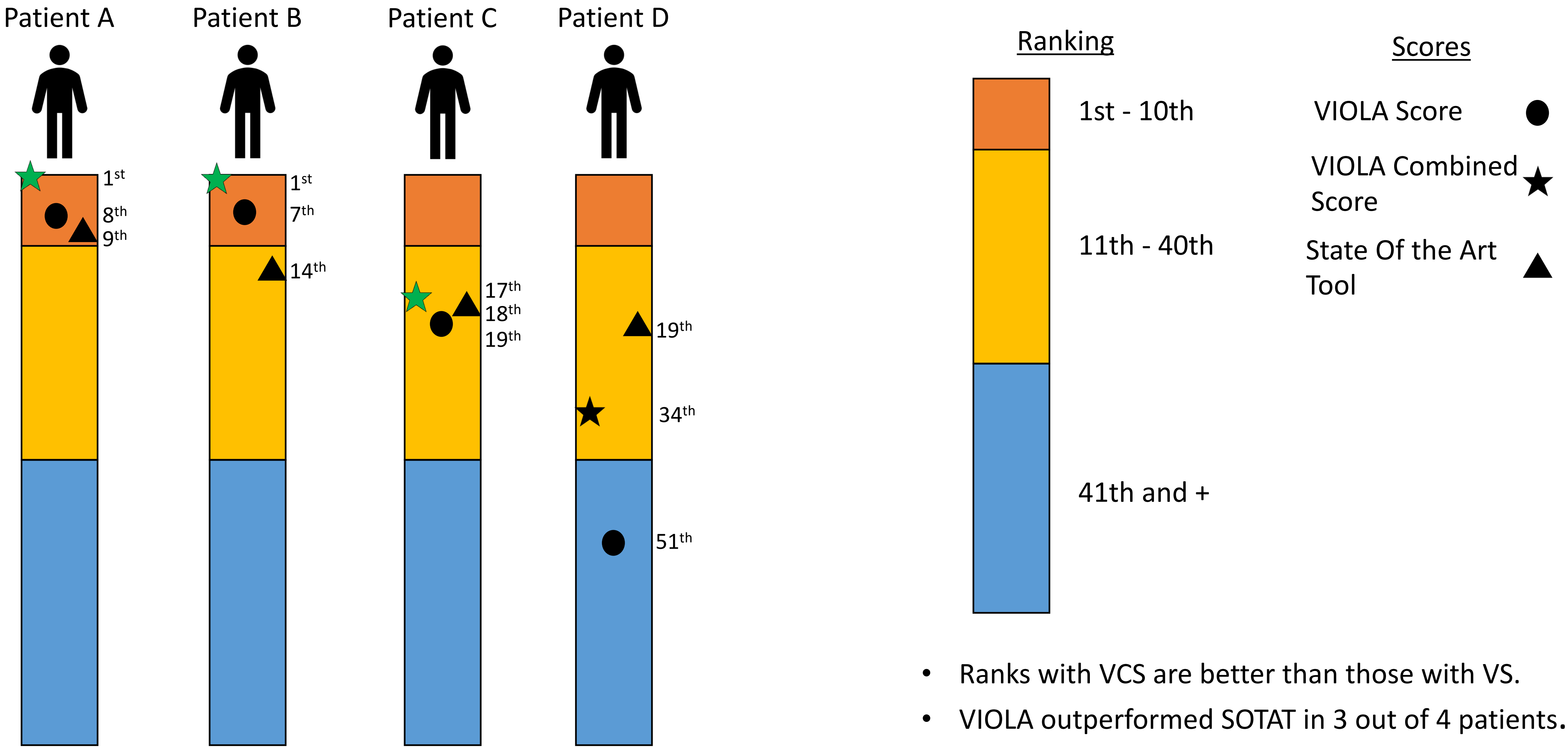
Can VIOLA find the responsible variant for positive patients?





- Ranks with VCS are better than those with VS.

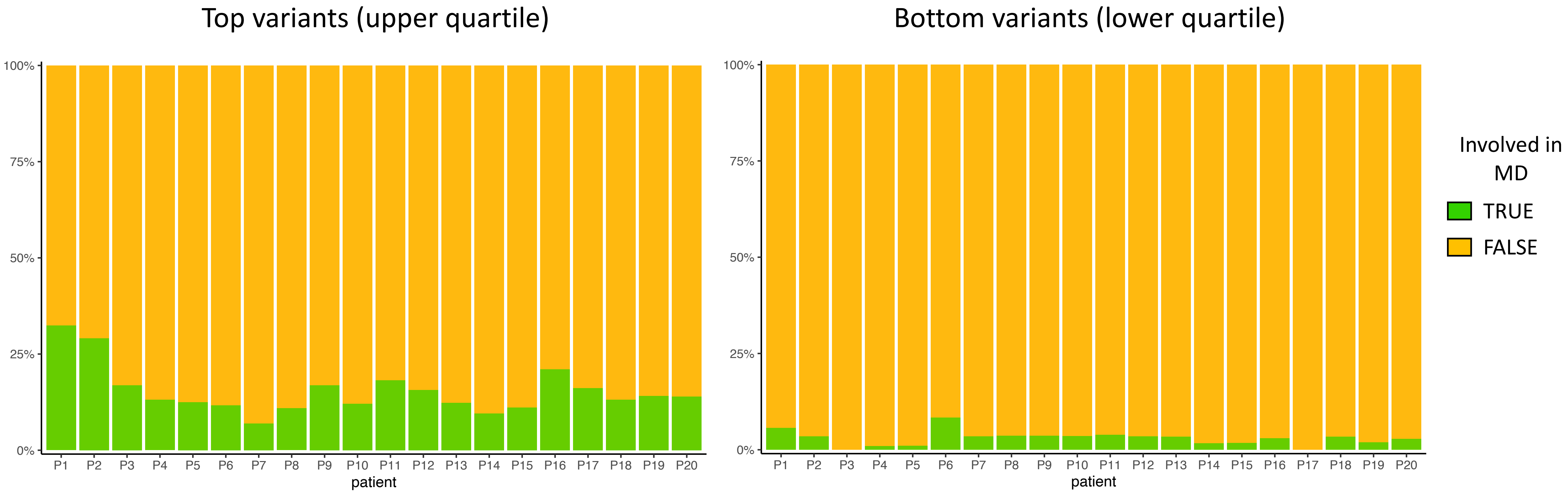
Can VIOLA find the responsible variant for positive patients?



- Ranks with VCS are better than those with VS.
- VIOLA outperformed SOTAT in 3 out of 4 patients.

VIOLA results

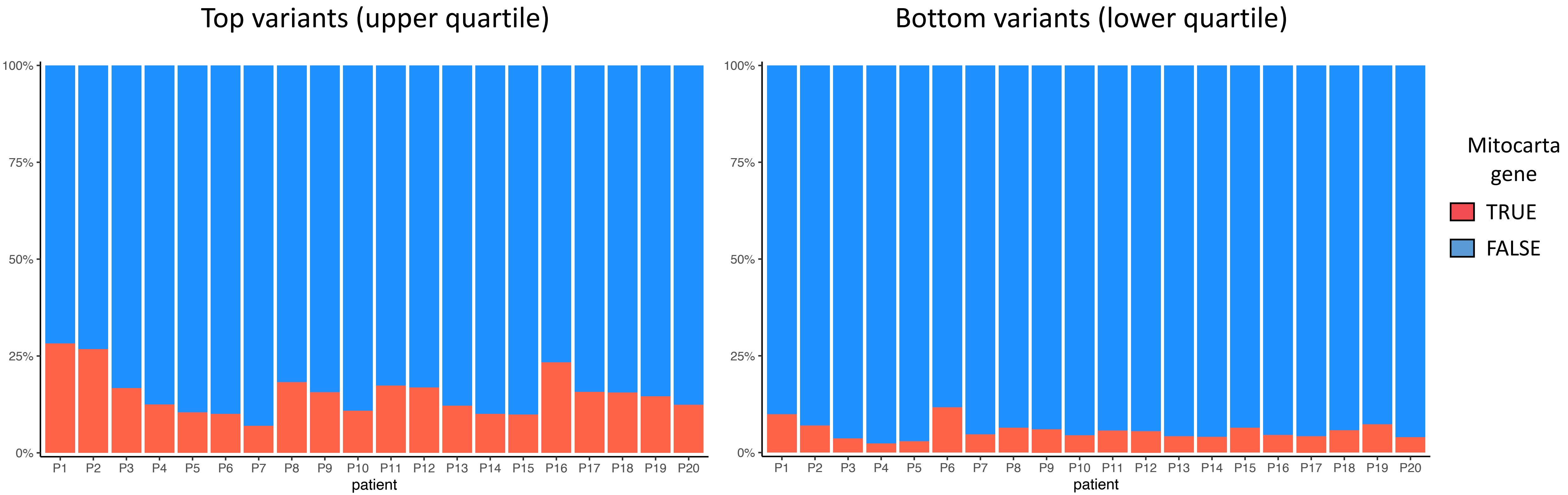
- Enrichment in genes already known to be involved in Mitochondrial disease (MD)



Genes bearing top variants are more enriched in genes already known to be involved in MD than genes bearing bottom variants

VIOLA results

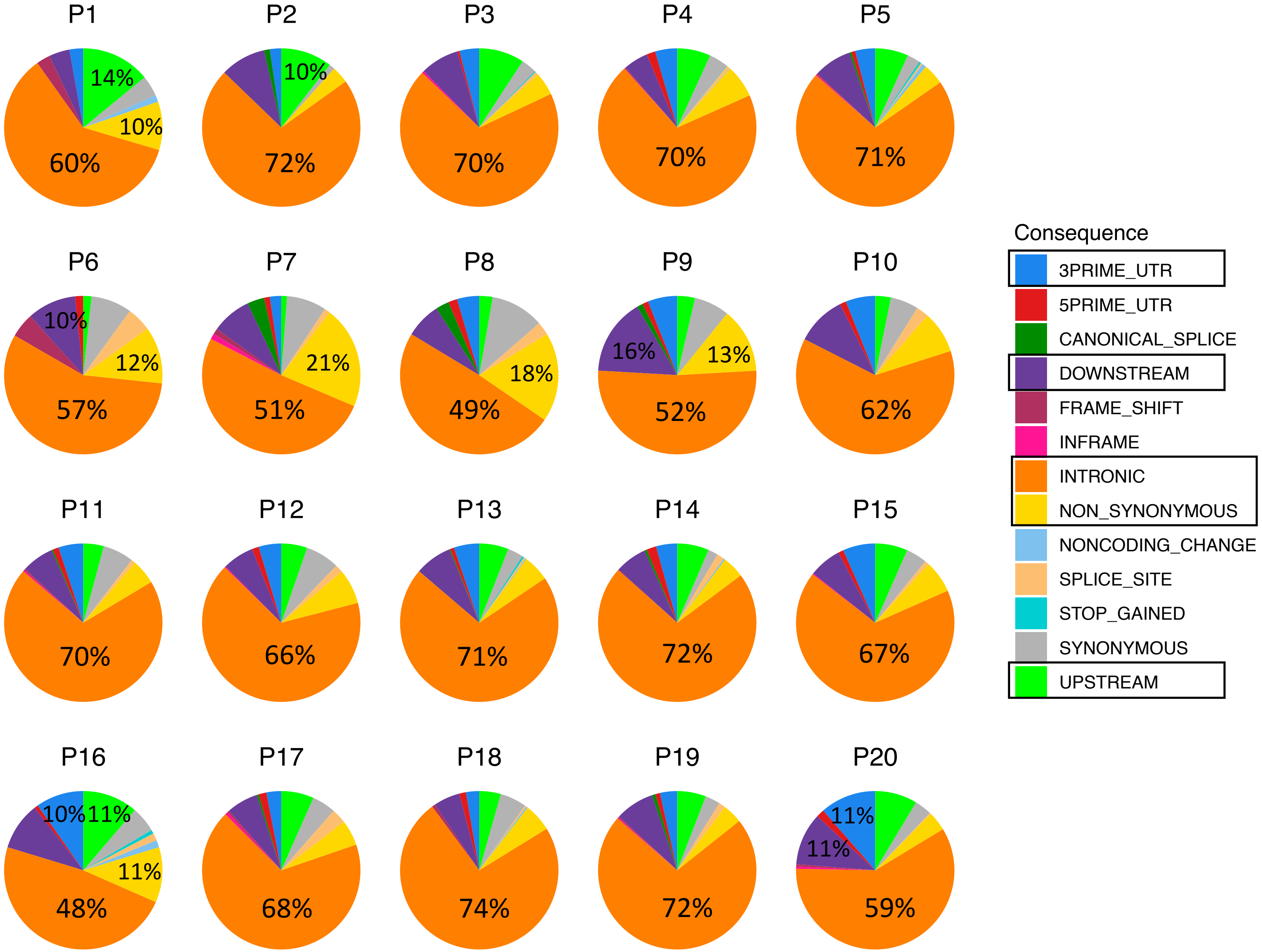
- Enrichment in MitoCarta genes



Genes bearing top variants are more enriched in MitoCarta genes than genes bearing bottom variants

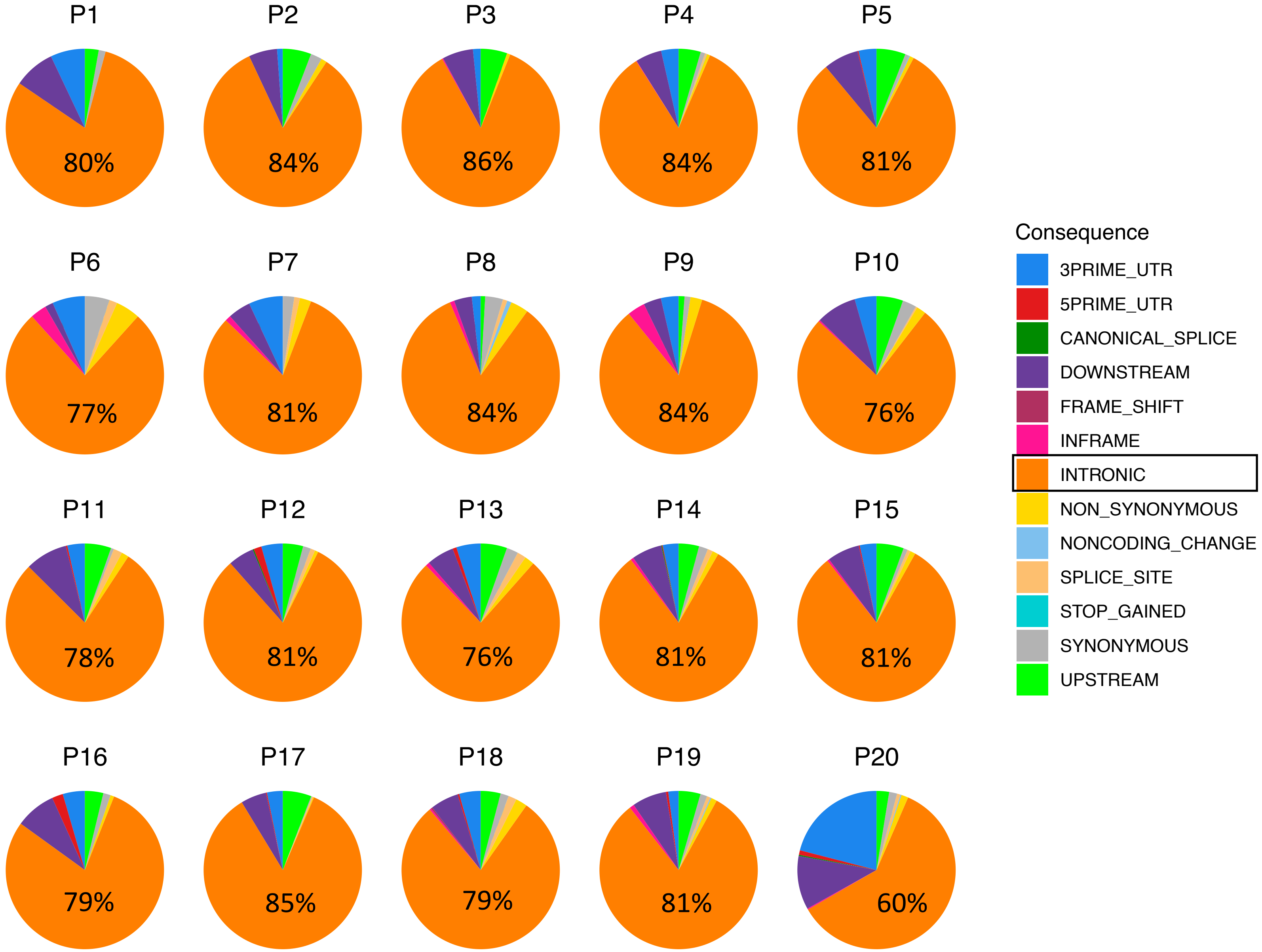
VIOLA results

- Consequences of **top** variants (upper quartile)



VIOLA results

- Consequences of **bottom** variants (lower quartile)



VIOLA find 2 potential candidates for 2 patients of the cohort

- **Case 1**:

Clinic

- Male baby
- Died shortly after birth
- Dilated cardiomyopathy with elevated lactates

Variant

- Intronic SNV in the C1QBP gene
- Heterozygous and rare (not listed in databases)
- Only found for this patient
- Ranked 7th with the VCS

Gene

- C1QBP = Encodes a multifunctional protein found mainly in the mitochondrial matrix.
- Listed in MitoCarta and known to be involved in MD
- Similar symtoms for 2 other patients with a variant in C1QBP gene

VIOLA find 2 potential candidates for 2 patients of the cohort

- **Case 2**:

Clinic

- Male adult (24 years old)
- Cardiomyopathy
- Transplanted

Variant

- Intronic SNV in the LAMA4 gene
- Heterozygous and rare (frequency of 0.000014 in GnomAD)
- Only found for this patient
- Ranked 3th with the VCS

Gene

- LAMA4 = Encodes extracellular matrix glycoprotein
- Known to be involved in cardiomyopathy

Conclusion of part 2

Patient-specific tool, very convenient in a diagnostic context

Model based on the **integration** of **genomics**, **transcriptomics** and **phenomics** data

Devlopment of a new model to prioritize genetic variants

For 3 out of 4 patients, VIOLA **outperfoms** Exomiser by ranking the responsible variant in the top 20

VIOLA found 2 potential **candidate** variants for 2 patients in the cohort

Take Home Messages

ABEILLE

Identification of Aberrant Gene Expression from transcriptomics data for small cohorts

VIOLA

Prioritization of genetic variants potentially responsible for MD using latent space

Personalized medicine

The diagnosis of MD is complex, important to move as far as possible towards a personalized medicine approach

Acknowledgements

Thesis supervisors :

Dr. Silvia Bottini

Dr. Sylvie Bannwarth

Genetic team (CHU de Nice):
Dr. Samira Ait-El-Mkadem Saadi

M2P2 team (ISA, Sophia Antipolis)

Master students:
Youssef Boulaimen
Gwendal Le Bideau
Jean Elisée Yao
Jasmine Kaur

UNIVERSITÉ CÔTE D'AZUR
FRANCE 2030 Initiative d'Excellence
IRCAN
chu de Nice
anr
INRAE