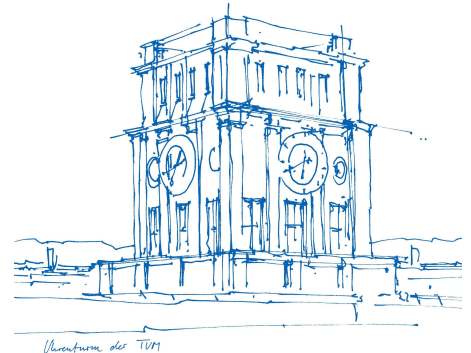


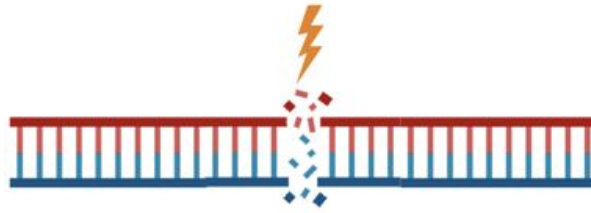
When the outlier is the signal: RNA-seq based diagnostics of rare disorders

Vicente A. Yépez, PhD
Scientific researcher
Gagneur lab
Technical University of Munich

BioInfoDiag
May 14th, 2024

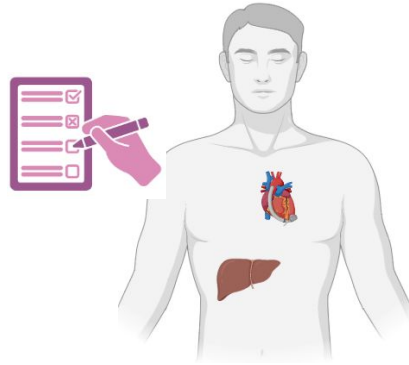


Rare genetic diseases
*life-threatening, chronically
debilitating conditions
predominantly caused by
variants in a single gene*

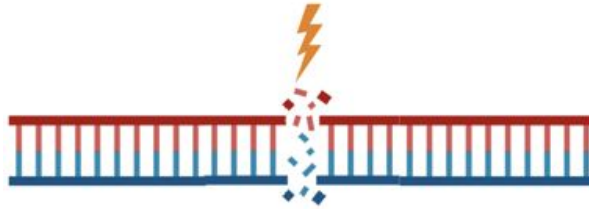


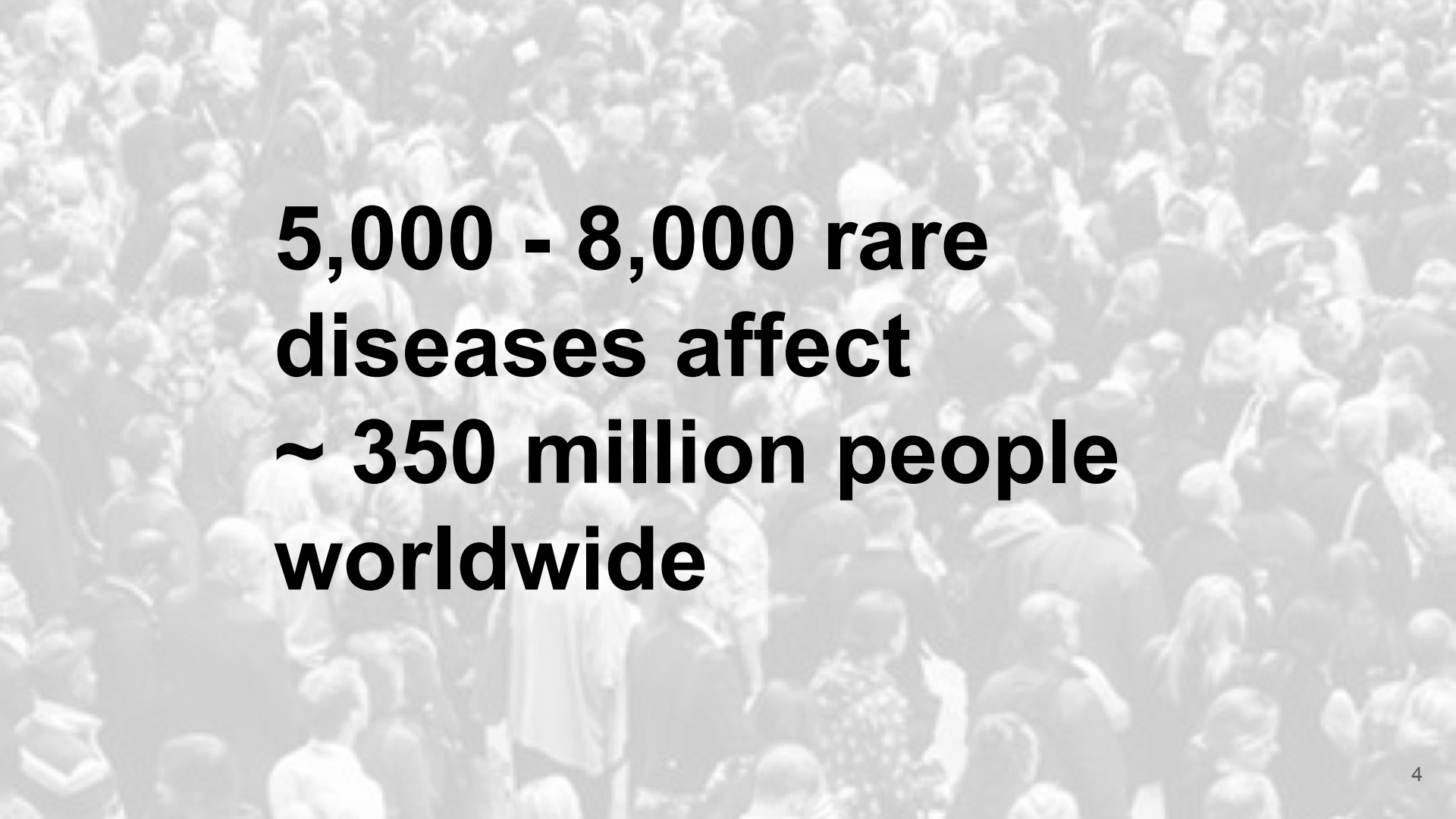
Rare: prevalence < 1 in 2,000

Step 1: Clinical evaluation



Step 2: DNA sequencing





**5,000 - 8,000 rare
diseases affect
~ 350 million people
worldwide**



**The current diagnosis
rate is ~50%**

Amberger et al, Nuc Ac Res, 2019
EURORDIS, Rare Diseases
Boycott and Ardigo, Nat Rev, 2018

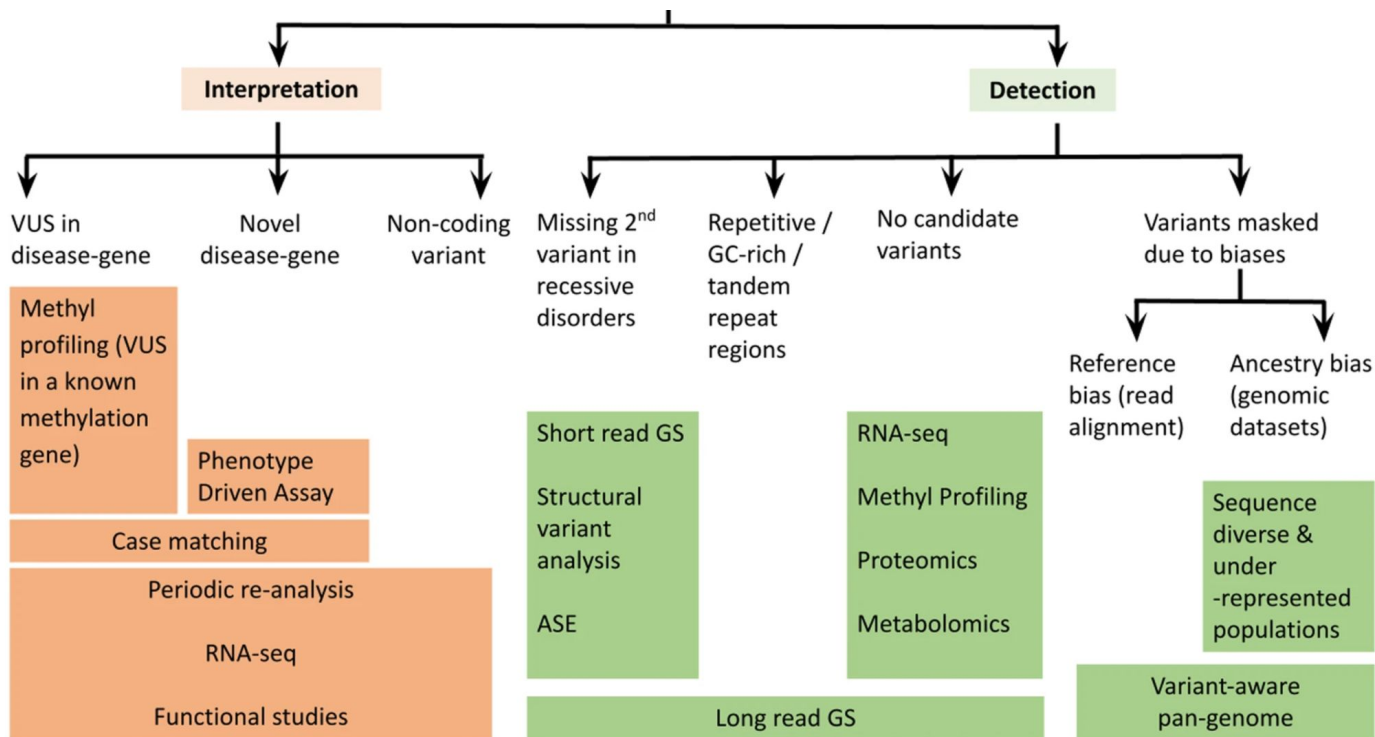
Challenges to diagnose rare disorders

From the medical side:

- Due to rarity of the disease, clinicians are often faced to new cases
- Overlapping phenotypes between disorders
- They're usually progressive, sometimes lethal -> time to diagnose matters
- Incomplete penetrance: same variant, different severity
- At least 2 cases are needed to prove pathogenicity of a new disease gene

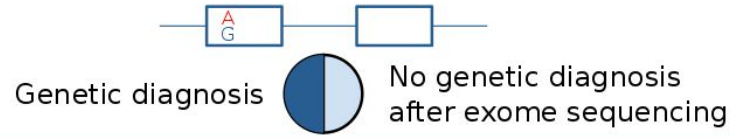
Challenges to diagnose rare disorders

From the genetic side:



Gene expression to increase diagnostics

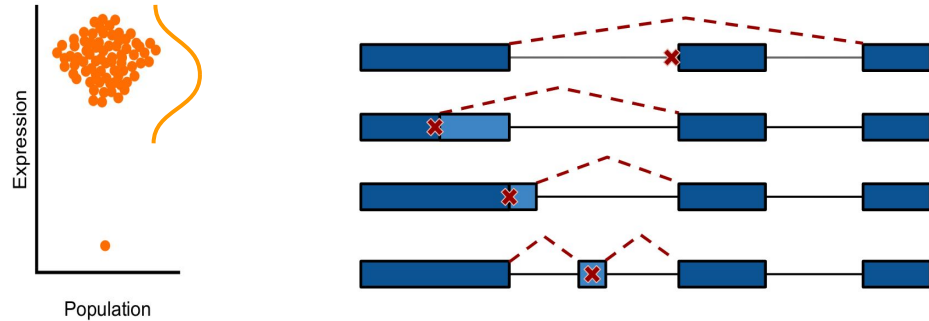
**~15% diagnostic
increase over WES**



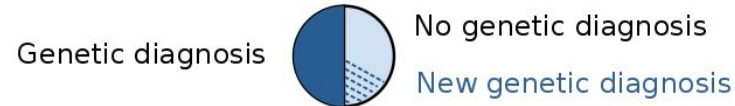
1. Sequence RNA (from clinically-accessible tissues)



2. Aberrant (not differential!) expression & splicing detection



3. Integrate with DNA and clinical data



Cummings *et al.*

- N=50, DR=35%
- Muscle
- Aberrant splicing
- Neuromuscular disorders

Brechtmann *et al.*

- OUTRIDER



Gonorazky *et al.*

- N=25, DR=36%
- RNA-seq variant calling
- Cell transdifferentiation

Murdock *et al.*

- N=115, DR=17%
- Transcriptome-directed analysis faster than candidate-variant
- Blood & fibroblasts

2017

2018

2019

2020

Kremer, Bader *et al.*

- N=105, DR=10%
- Skin fibroblasts
- Aberrant expression
- Mono-allelic expression
- Mitochondrial disorders



Frésard *et al.*

- N=94, DR=7.5%
- Blood
- Integration with external controls
- Systematic phenotypic integration
- Different pathophysiology



Timeline of RNA-seq studies and methods
1 study ⇔ 1 center

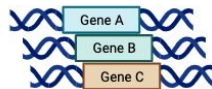


Mertes *et al.*

- FRASER

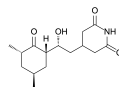
Yépez, Gusic *et al.*

- N=303, DR=16%
- Clinical implementation
- Expressed genes comparison
- Establishment of DROP



Dekker *et al.*

- N=67, DR=13%
- Neurodev. disorders
- Cycloheximide to inhibit NMD



Lunke *et al.*

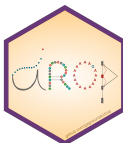
- N=115, DR=3%
- Trio RNA-seq
- Impact in care management

2021

2022

2023

2024



Yépez *et al.*

- Detection of RNA Outliers Pipeline, DROP



Lee, Kwong *et al.*

- N=48 fetuses, DR=8%
- Prenatal approach using amniotic fluid

Deshwar, Yuki, Hou *et al.*

- N=39 families, DR=13%
- Trio RNA-seq



Li *et al.*

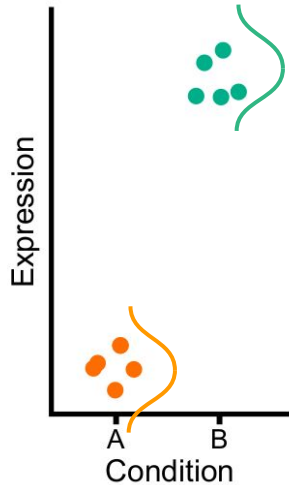
- N=71, DR=25%
- Fibroblast-to-neuron cell transdifferentiation
- Evaluate time & cost



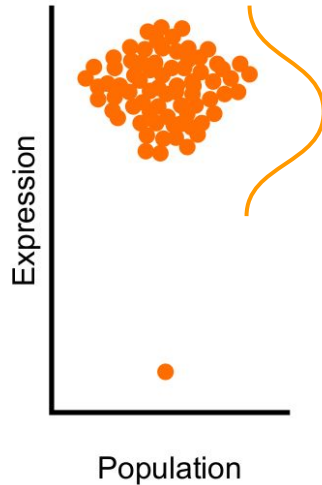
**In rare disease diagnostics,
the outlier is the signal**

Aberrant (not differential!) expression detection

Differential expression analysis



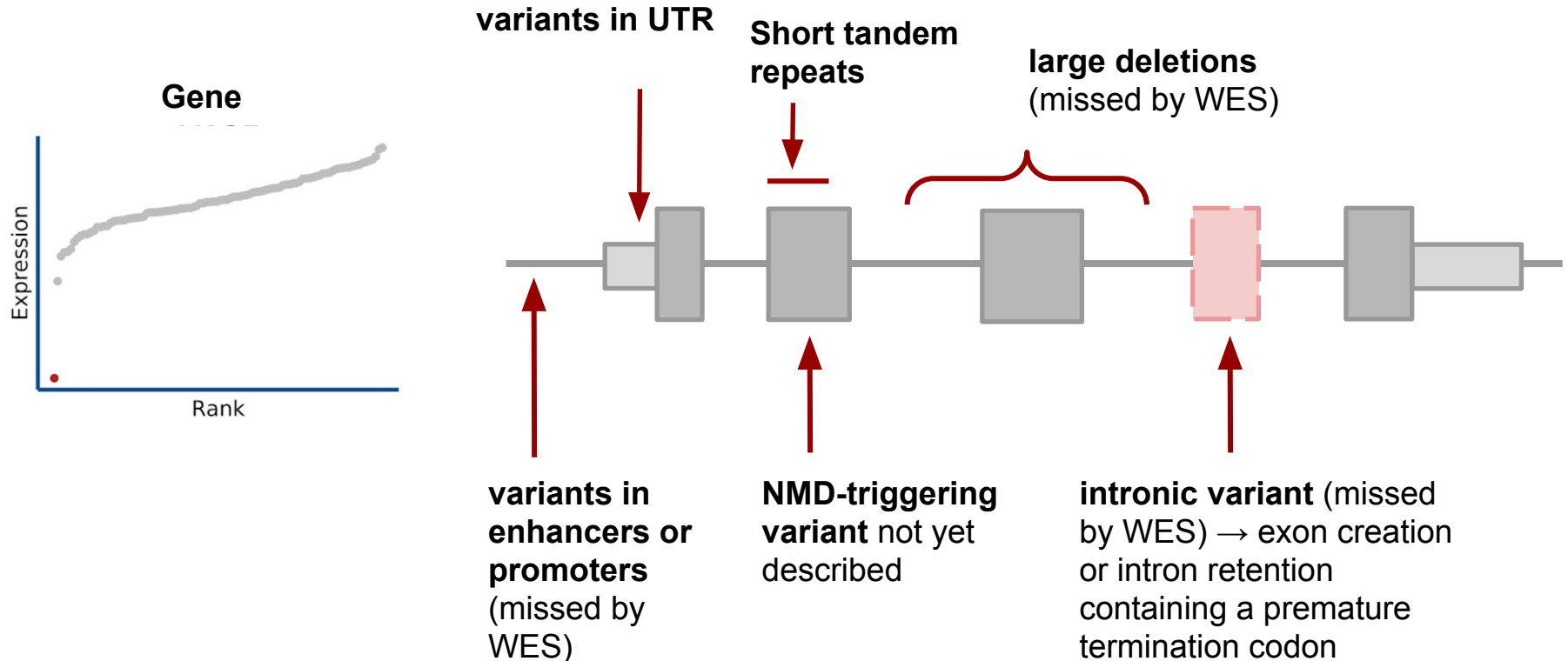
Outlier detection



- Not interested in differential expression between 2 groups
 - Condition A vs Condition B
- Find the 'outlier' - the gene whose impaired function could explain the disease
 - The population can be composed of all affected samples
 - Controls can be included to increase sample size

DESeq2/edgeR

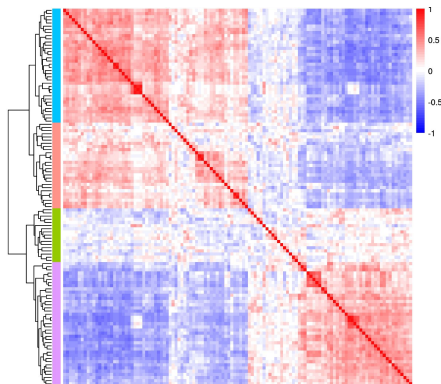
Detecting aberrant gene expression can lead to finding or validating disease causal variants



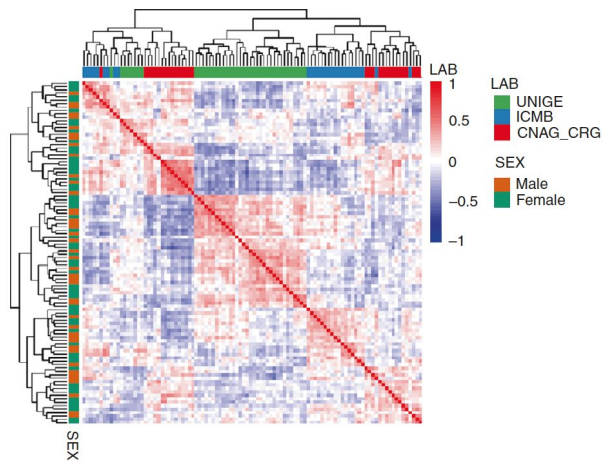
Gene expression data exhibits strong covariation

Sample-sample correlation heatmaps

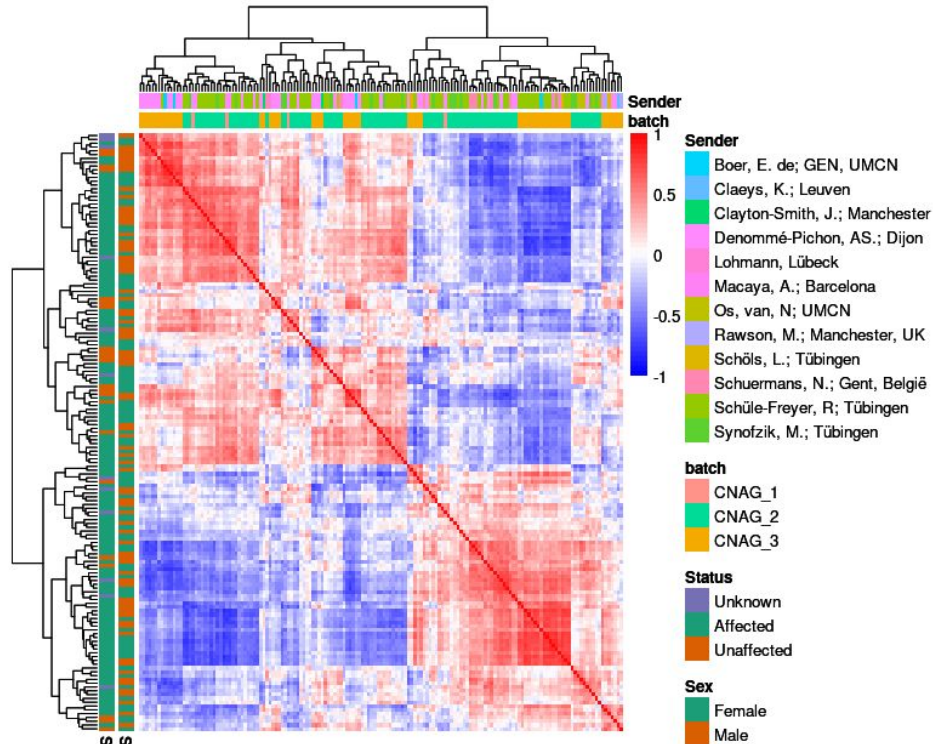
Mitochondrial disease dataset N = 119
Kremer, Bader et al., Nat Commun, 2017



GEUVADIS dataset N=100
Lappalainen et al., Nature, 2013

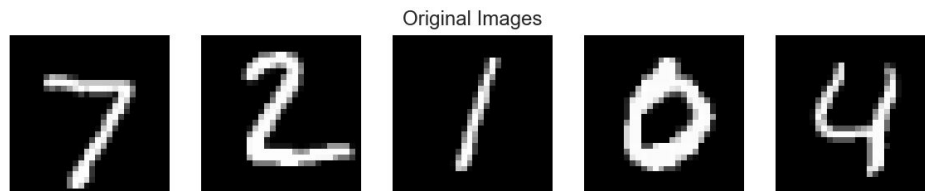


Solve-RD's whole blood cohort N = 253

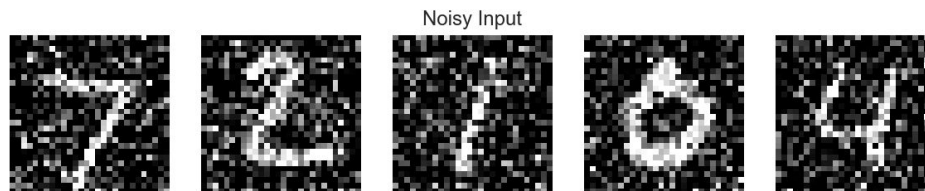


Fitting denoisers with denoising autoencoders

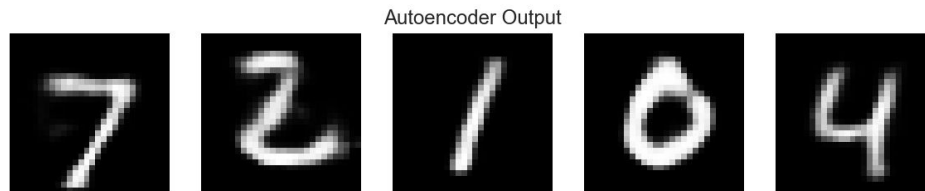
\mathbf{X}



$\mathbf{X}^{\text{corrupt.}}$



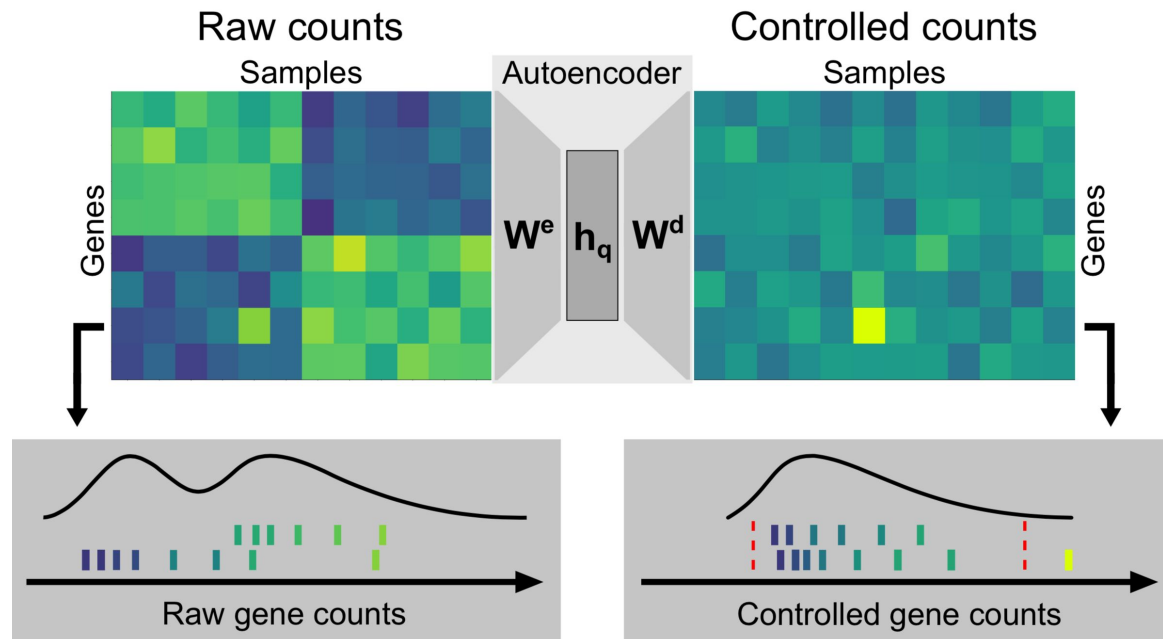
$f(\mathbf{X}^{\text{corrupt.}}, \theta)$



$$\min_{\theta} \|\mathbf{X} - f(\mathbf{X}^{\text{corrupt.}}, \theta)\|^2$$

OUTRIDER: denoising autoencoder for RNA-seq data

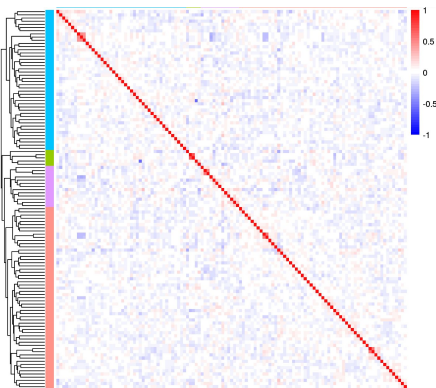
- Negative Binomial loss
- Number of latent factors (q) set to maximise precision-recall of artificially injected outliers
- P-value per sample - gene combination



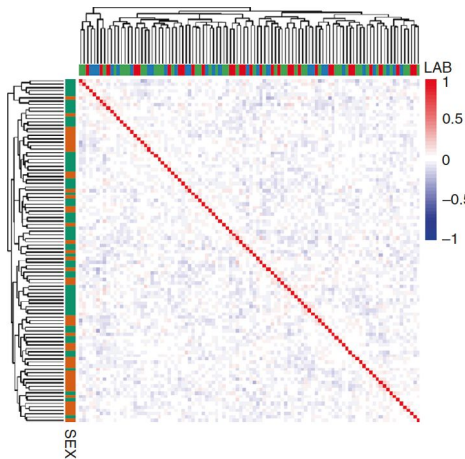
OUTRIDER successfully removes sample covariation

Sample-sample correlation heatmaps after correction

Mitochondrial disease dataset
N = 119
Kremer, Bader et al., Nat Commun, 2017

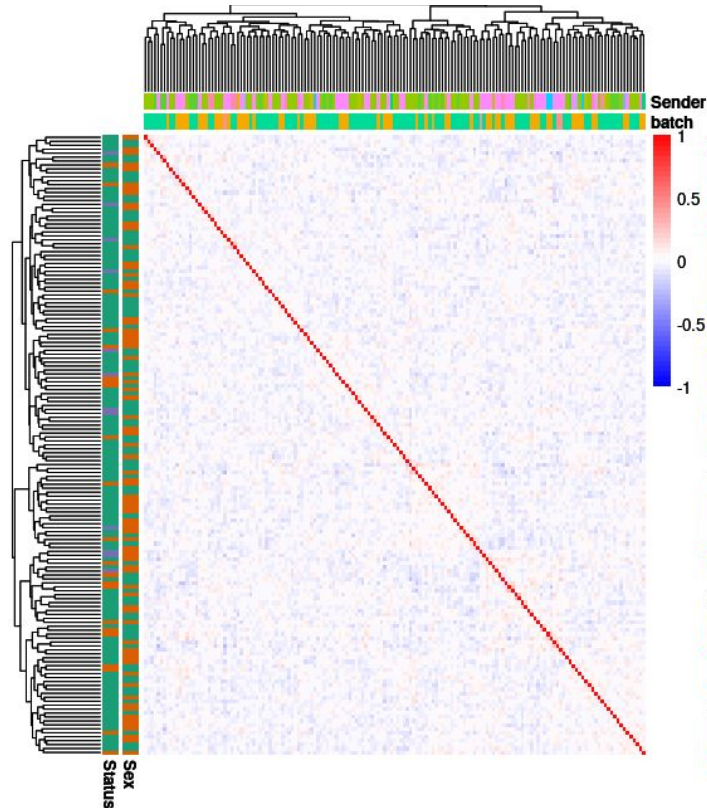


GEUVADIS dataset
N=100
Lappalainen et al., Nature, 2013

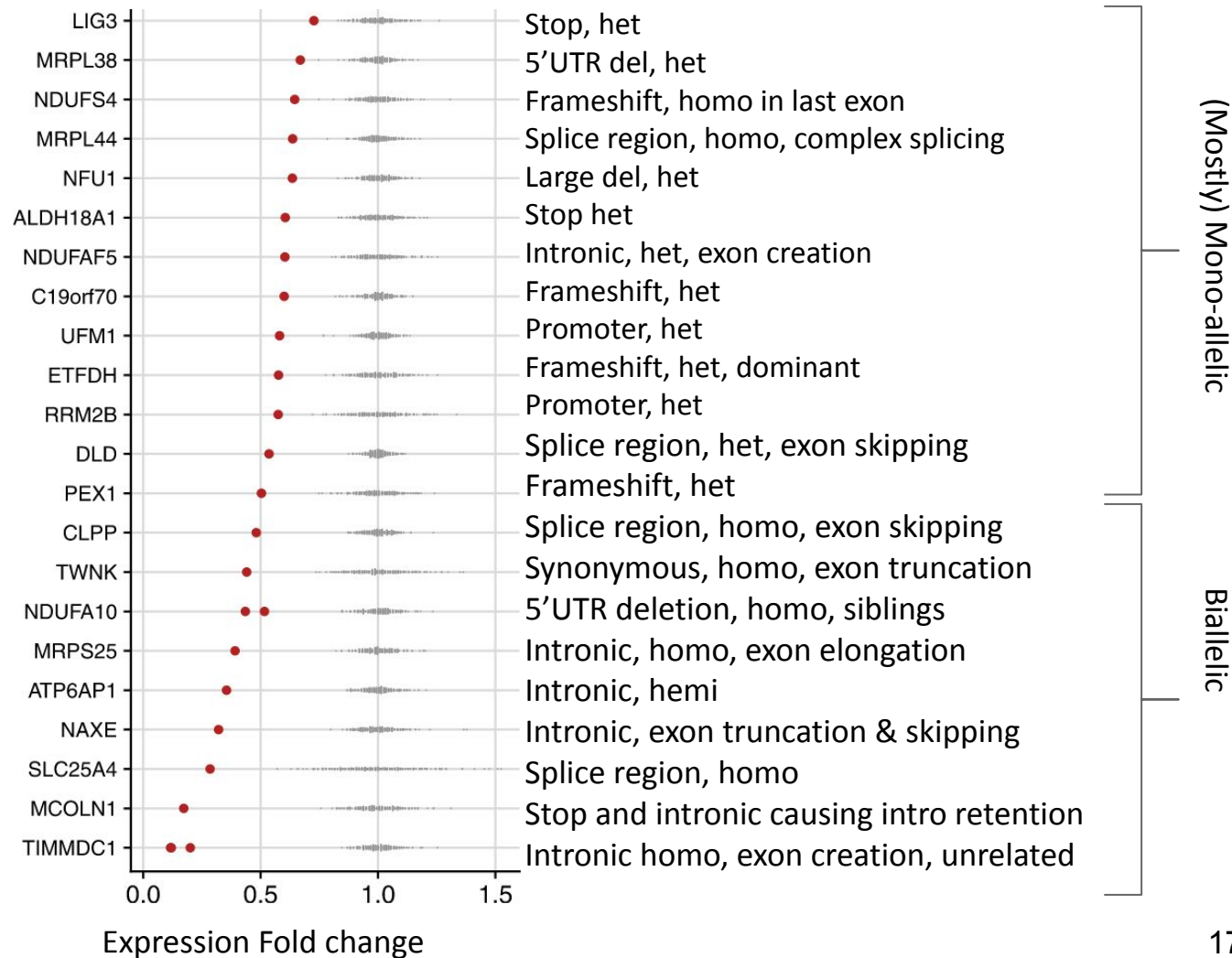


LAB
UNIGE
ICMB
CNAG_CRG
SEX
Male
Female

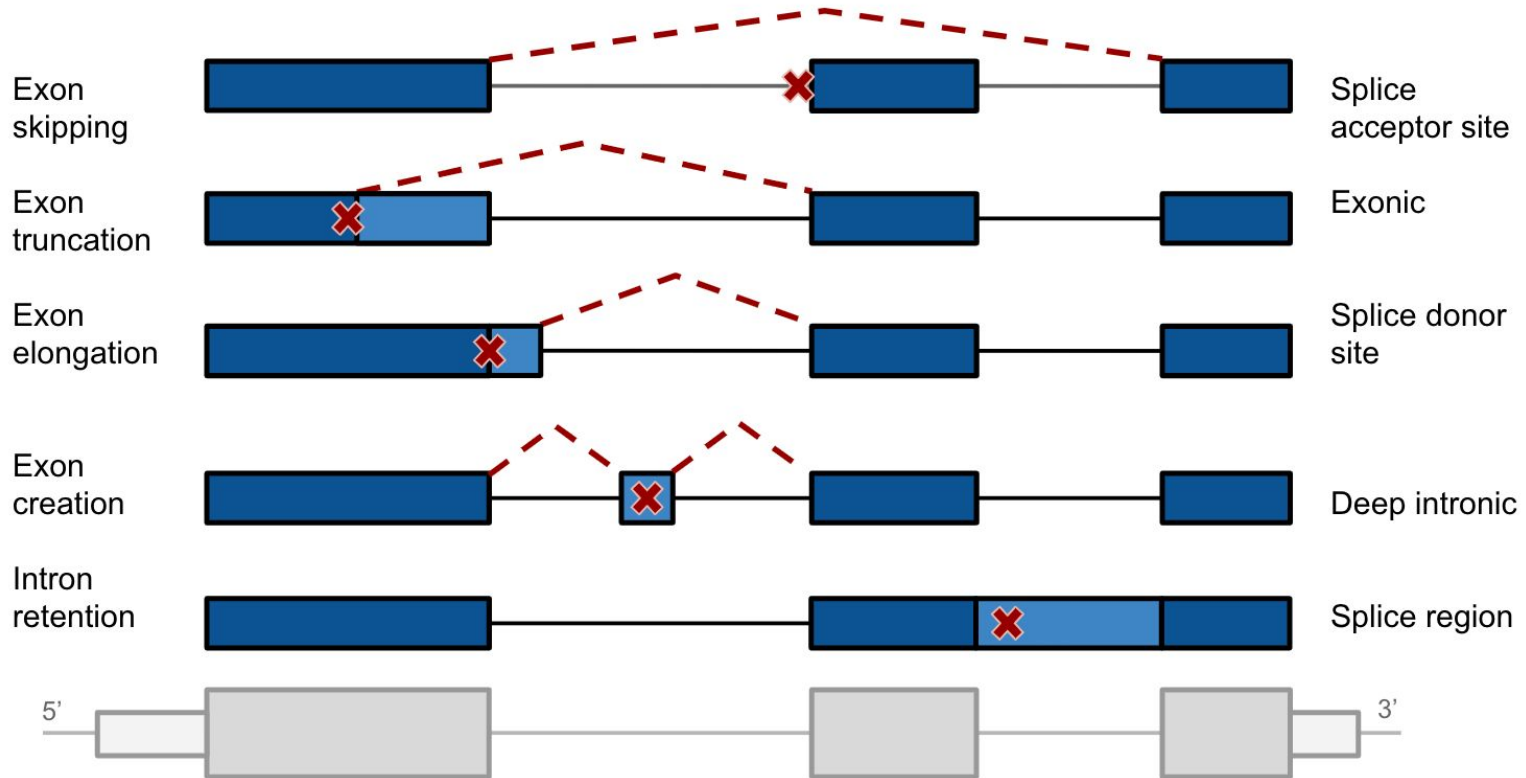
Solve-RD's whole blood cohort
N = 253



Reduction in expression suggests mono- or biallelic LoF variants

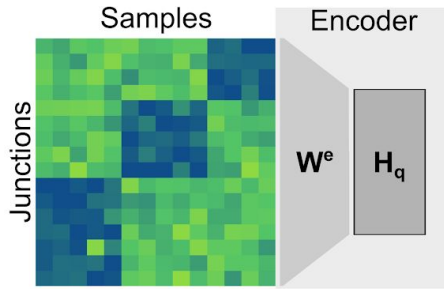


Detecting aberrant splicing, in its many forms, can also lead to finding or validating disease causal variants

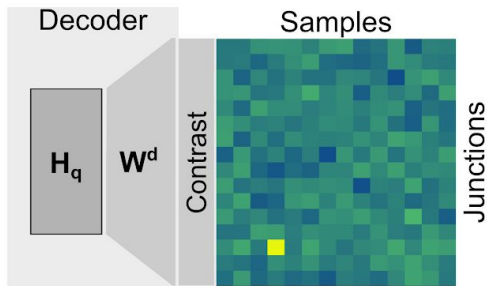


FRASER to detect aberrant splicing

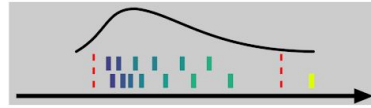
1. Latent space fitting



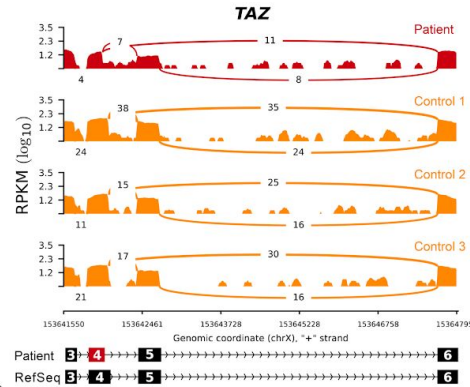
2. Distribution fitting



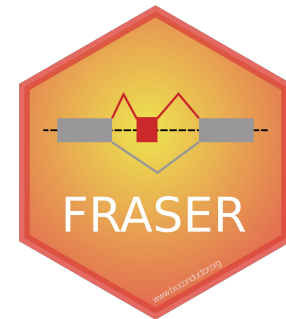
Significance based outlier detection



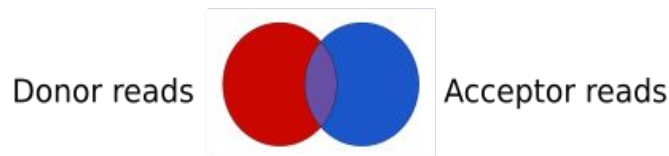
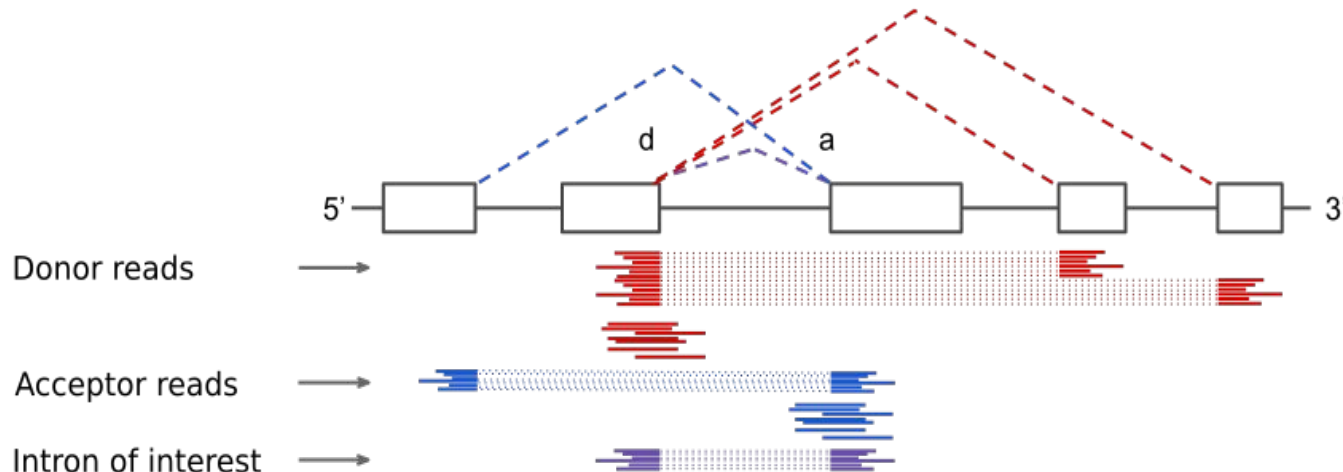
Visualization



- Splice-site centric metrics (ψ_5 , ψ_3 , SE_5 and SE_3)
- Similar as OUTRIDER but with Beta Binomial loss on each metric
- P -value per sample - junction - metric combination



Novel intron-centric metric to quantify splicing: Intron Jaccard Index



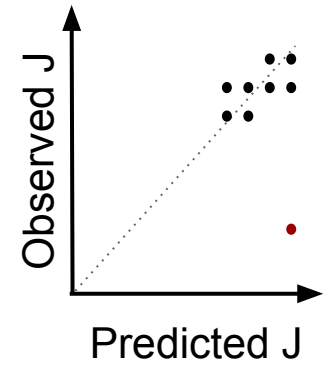
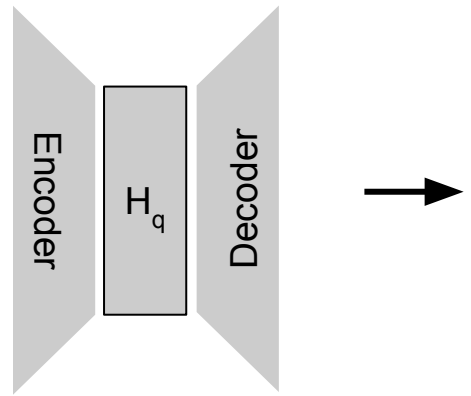
$$\text{Intron Jaccard Index: } J = \frac{|D \cap A|}{|D \cup A|}$$

- More robust than splice-site centric metrics (ψ_5 , ψ_3 , SE_5 and SE_3)
- ~120K introns / cohort are tested also using Beta Binomial

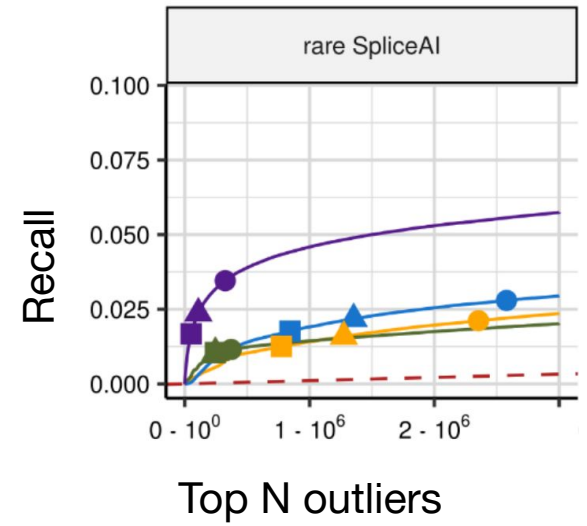
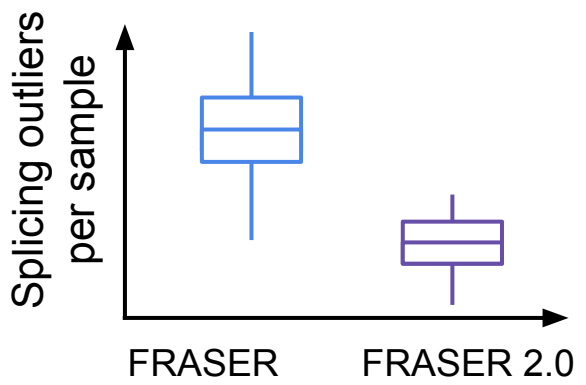
Scheller, et al., AJHG, 2023

FRASER 2.0 improves over FRASER and other methods

Intron Jaccard Index $J = \frac{|D \cap A|}{|D \cup A|}$



~8 times fewer splicing outlier calls (on GTEx and 2 rare disease datasets)



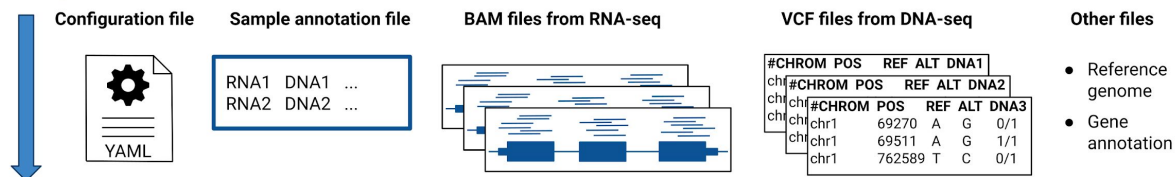
- LeafcutterMD
- SPOT
- FRASER
- FRASER 2.0



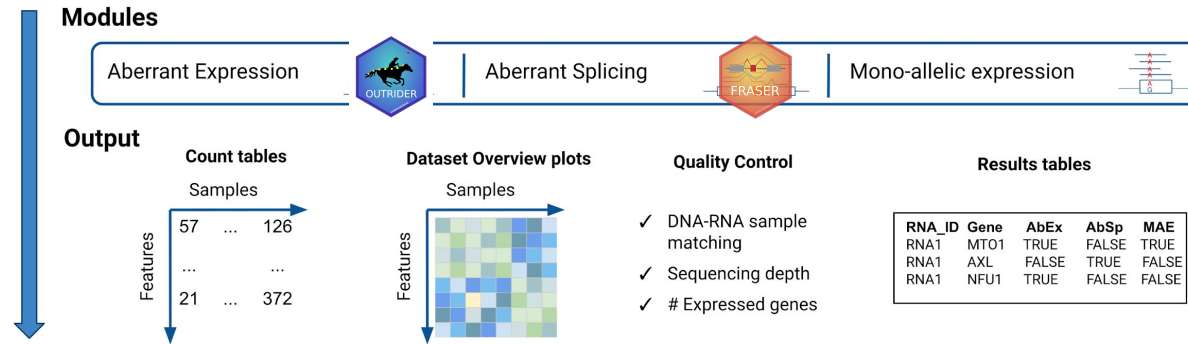
increased proportion of outliers with rare splice-disrupting candidate variant

Detection of RNA outliers pipeline - DROP

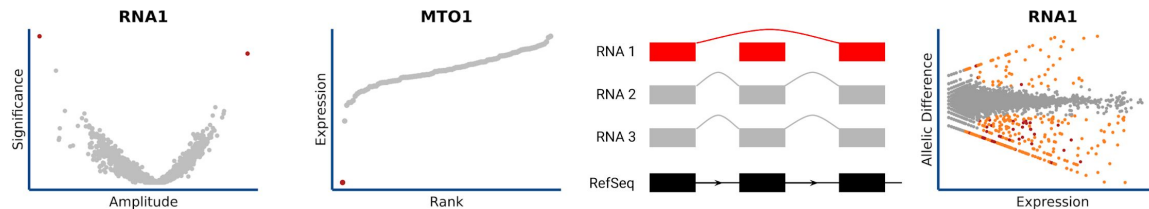
1. Input



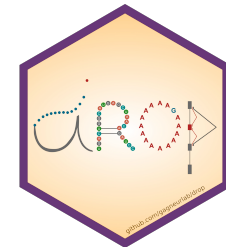
2. Detection of RNA outliers pipeline (DROP)



3. Analyze individual results



- Easy to install through GitHub including all dependencies
- After set-up, runs each module with 1 command
- Runs cohorts of few hundreds of samples in < 1 day
- Used by centers all over the world

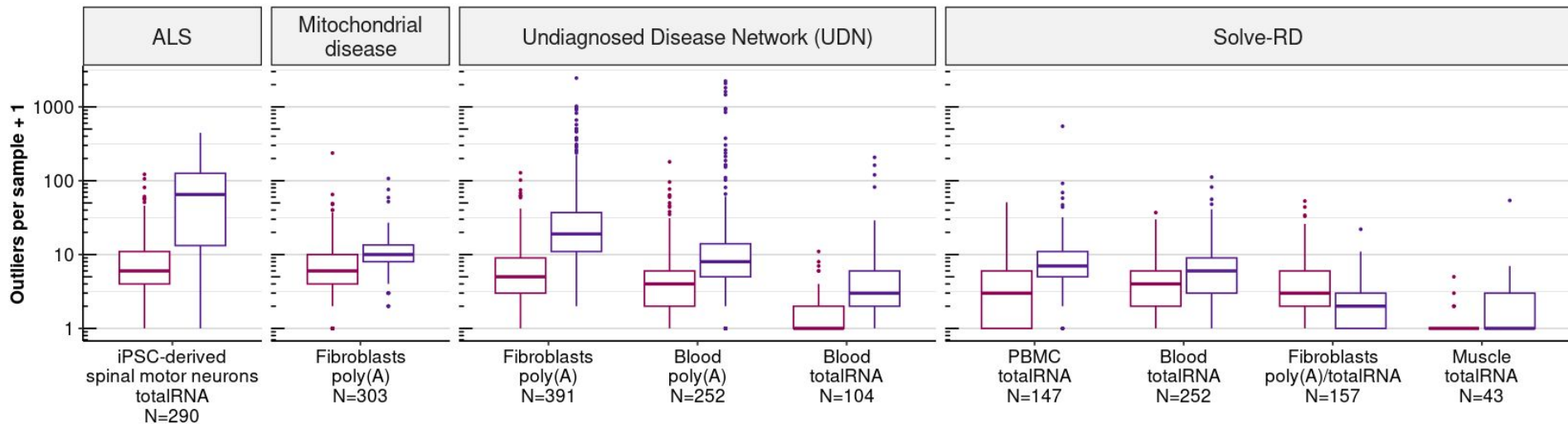


github.com/gagneurlab/drop

Yépez et al, Nat Protoc, 2021

Handful of outliers per sample across multiple rare disease cohorts

OUTRIDER FRASER 2.0



Median expression outliers: 3
Median splicing outliers: 8

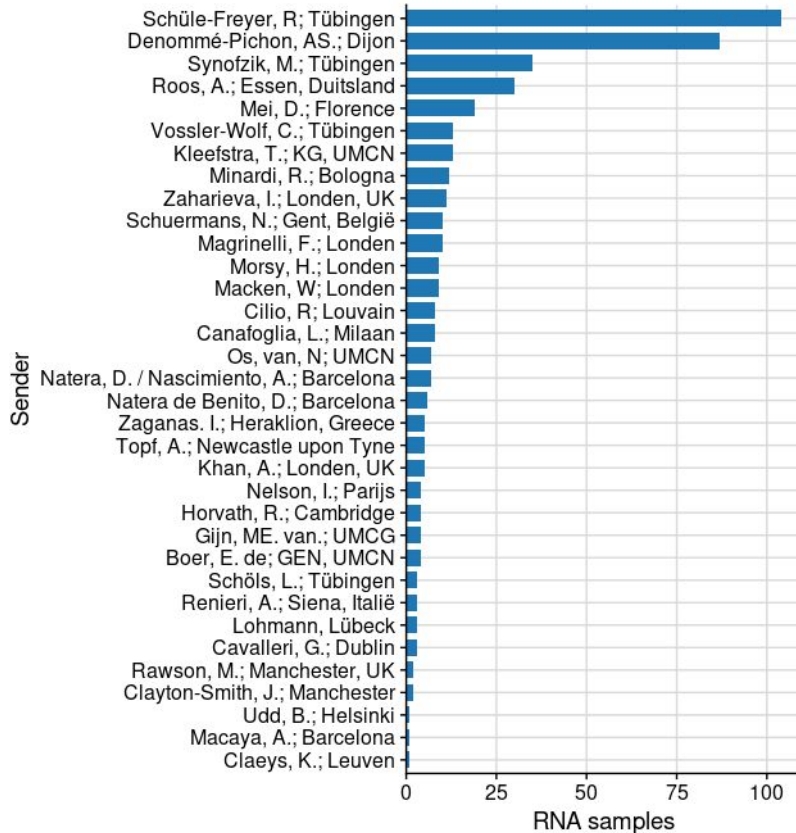


Need to upscale and standardize RNA-seq methods for big consortia



More to come!

Major challenge: cope with large number of senders



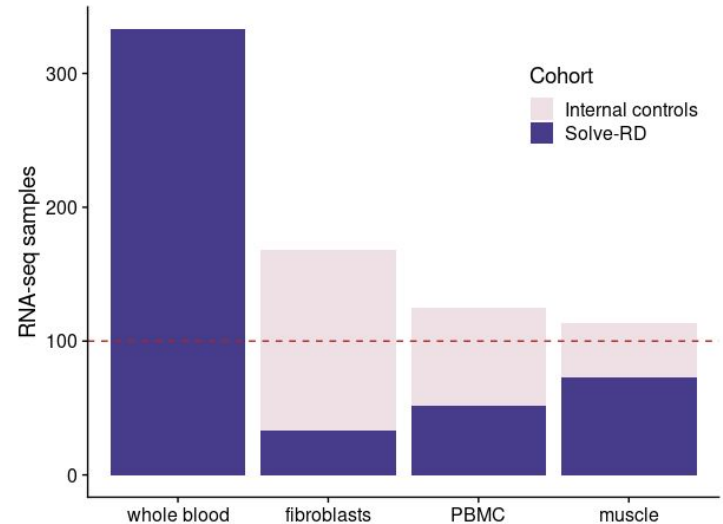
Extremely important to:

- Compile all results in one place (DNA, RNA & phenotype)
- Minimize manual inspection
- Instruct how to interpret results following similar criteria
- Standardize all steps

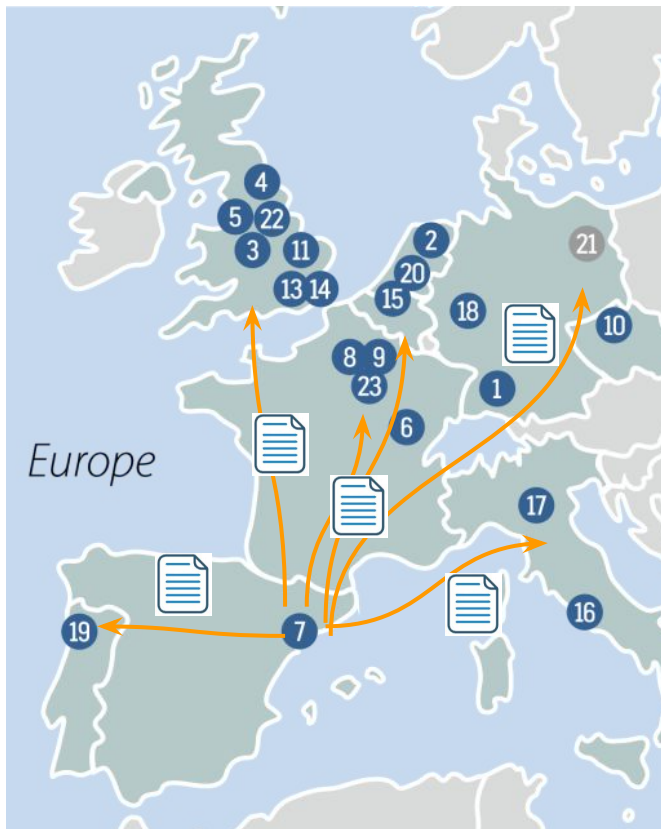
Centralized data preprocessing and results generation



- Same criteria for sample inclusion
- Standardize all steps from raw data to results
- Increased statistical power
- Minimize data transfer agreements



Instead of sending tables



⇒ organize *Solvathons*!

- 3-days on-site and online event
- Multidisciplinary: bioinformaticians, biologists, clinicians, geneticists, group leaders, postdocs, PhDs, ...
- Goals:
 - Instruct on how to analyze the results
 - Diagnose samples!
- Great starting point for follow-up analyses!



DNA 1st and RNA 1st as two avenues for diagnostics



Rare variants called from

- SNVs + indels: SR WES/WGS
- SVs: LR WGS + Optical Genome Mapping
- Short Tandem Repeats & de novo

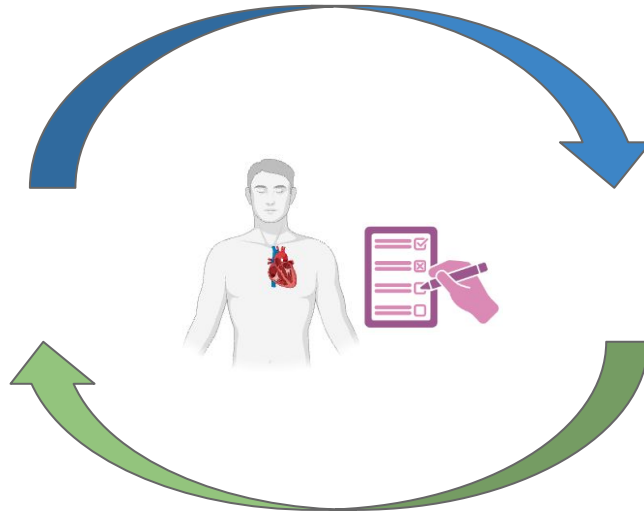


DNA 1st

Candidate VUS
~10s

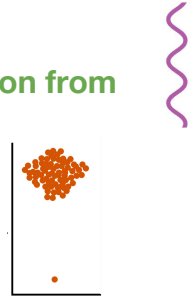
WGS
Novel disease genes
Relax filter cut-offs

evaluate



discover

Aberrant expression from
OUTRIDER
FRASER



Reduced multiple testing
burden

Transcriptome-wide testing ~12K
genes and 150K junctions
~10s outliers/sample

RNA 1st

DNA 1st and RNA 1st as two avenues for diagnostics



Rare variants called from
SNVs + indels: SR WES/WGS
SVs: LR WGS + Optical Genome Mapping
Short Tandem Repeats & de novo



DNA 1st

Candidate VUS

WGS
Novel disease genes
Relax filter cut-offs

To diagnose a case you need all:

- Variant ✓
- Phenotype ✓
- RNA defect ✓
- Segregation ✓

discover

Aberrant expression from
OUTRIDER
FRASER



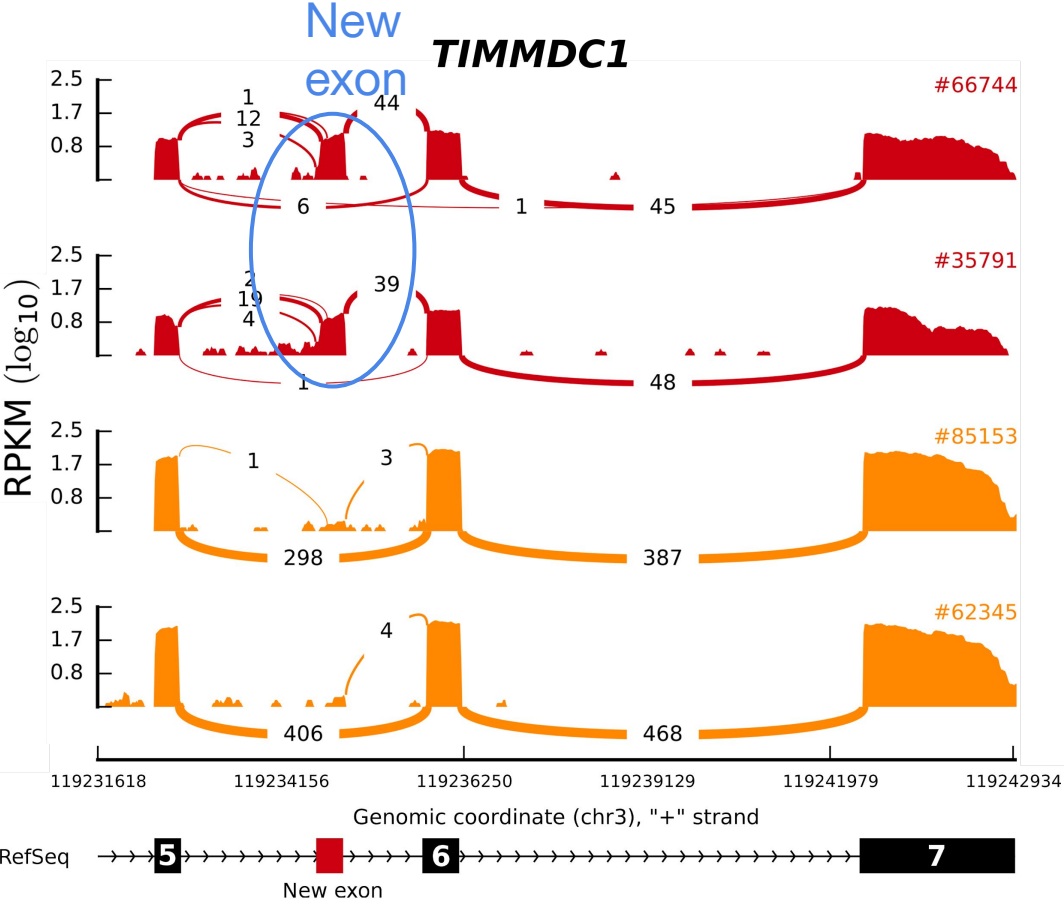
Reduced multiple testing burden

Transcriptome-wide testing

RNA 1st

(In the next slides I showed our current diagnostic rate in Solve-RD of 20 samples, including (unbalanced) translocations, insertions & deletions. As it is unpublished data, I unfortunately cannot share it. Keep updated for the preprint!)

RNA 1st: exon creation due to intronic variant



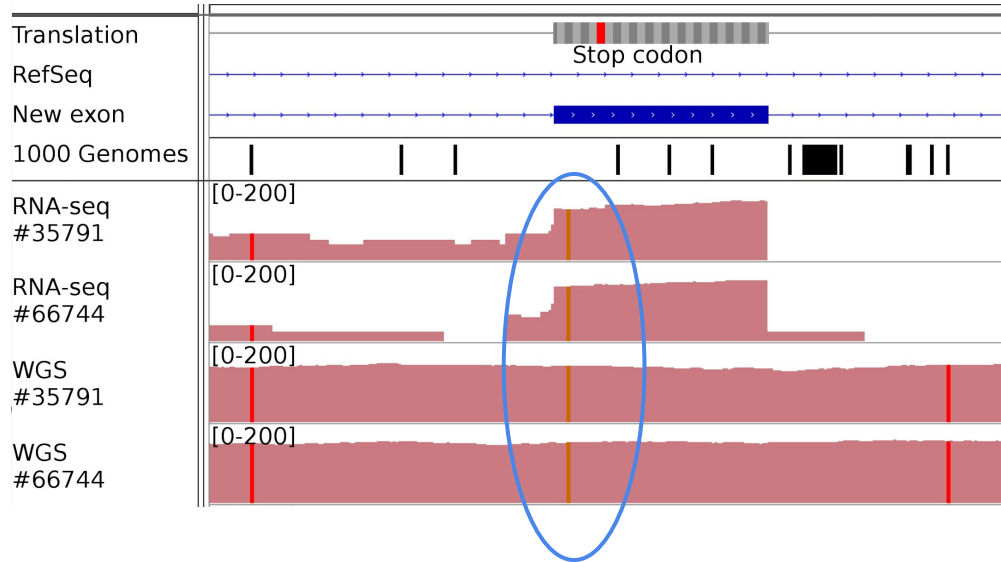
2 mitochondrial disease patients with TIMMDC1 defect

Other samples

TIMMDC1: Translocase Of Inner Mitochondrial Membrane Domain Containing 1

Kremer, Bader et al., Nat Commun, 2017

RNA 1st: exon creation due to intronic variant



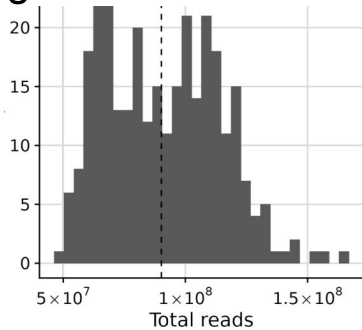
- Homozygous intronic variant missed by WES
- Gene was originally not a disease gene, but confirmed by proteomics (all complex I subunit down)
 - RNA defect ✓
 - Variant & segregation ✓
 - Phenotype ✓

Kumar et al, npj Genomic Medicine, 2022

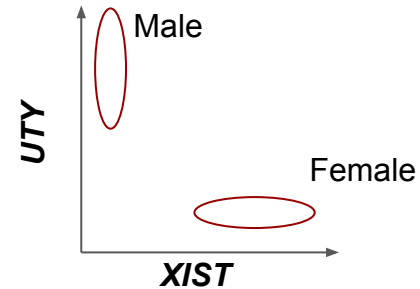
Oligonucleotide correction of an intronic *TIMMDC1* variant in cells of patients with severe neurodegenerative disorder

Extensive QC to verify that the RNA-seq samples:

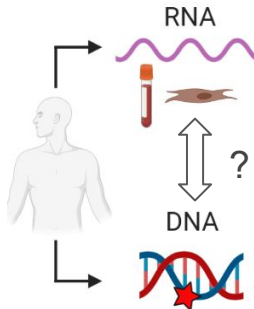
1. have comparable sequencing depth by comparing counts and size factors



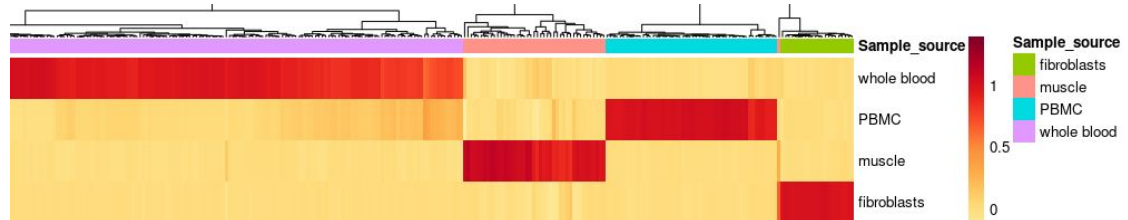
2. belong to the annotated sex by comparing the expression of XIST and UTY



3. match the annotated DNA by comparing the variants

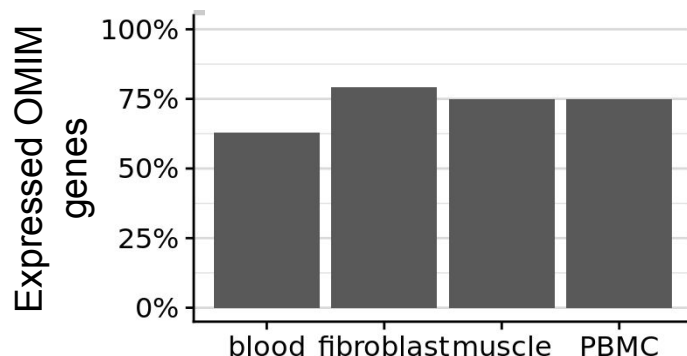


4. match the annotated tissue by mean value decomposition



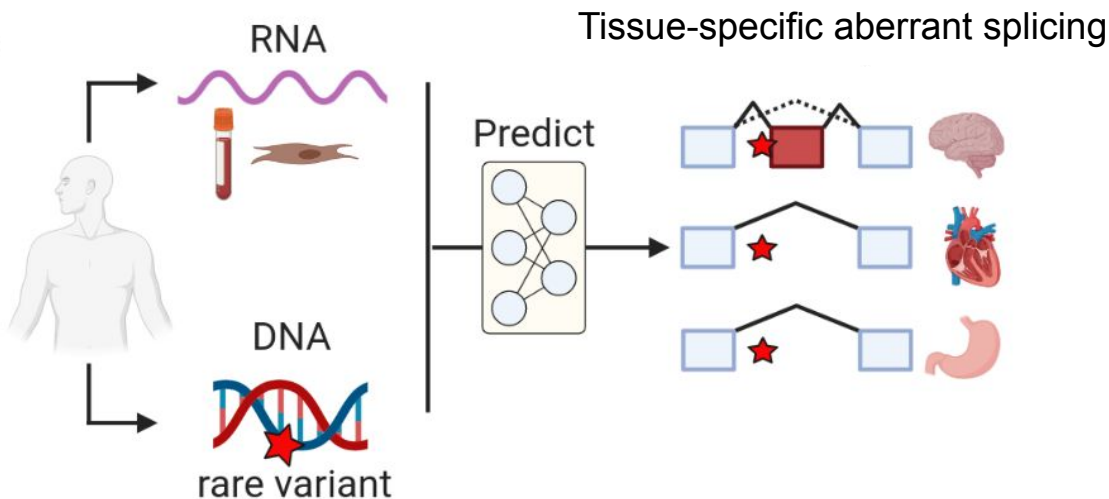
Limitations: RNA-seq is not always able to capture the effect of the variants

- Gene not expressed in probed tissue
 - Limited to accessible tissues - blood, skin, muscle
 - ~60-70% Mendelian disease genes expressed
- Variant does not affect transcript (e.g. missense or synonymous)
- Expression and splicing outliers are not highly reproduced across tissues
- Cohorts of at least 100 samples are needed to detect outliers



How can we overcome this?

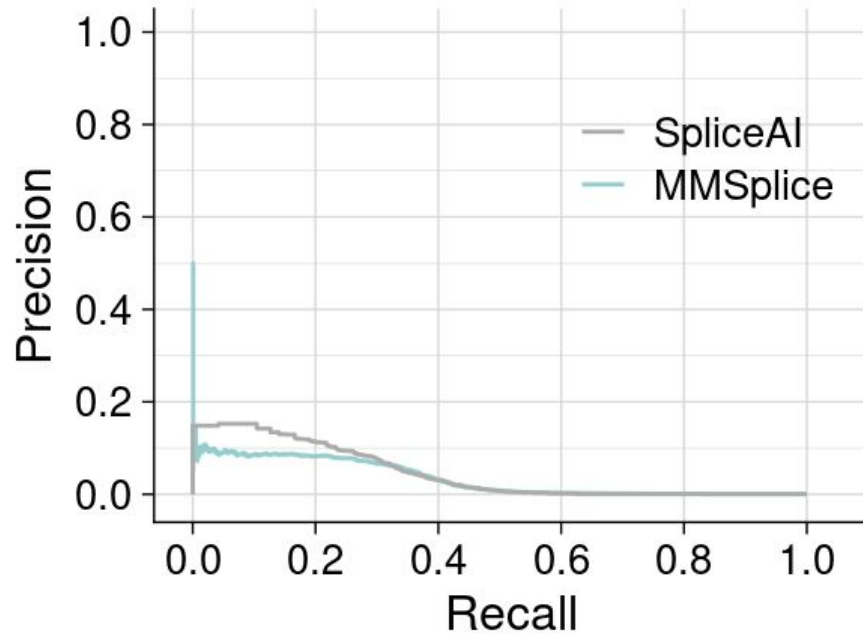
Aberrant splicing prediction in any human tissue



Having FRASER established, we generated the first benchmark dataset for tissue-specific aberrant splicing prediction using GTEx

- 16,213 post-mortem RNA-seq samples
- 946 individuals
- 49 tissues
- 8.8 million rare variants
- 21,000 aberrant splicing events

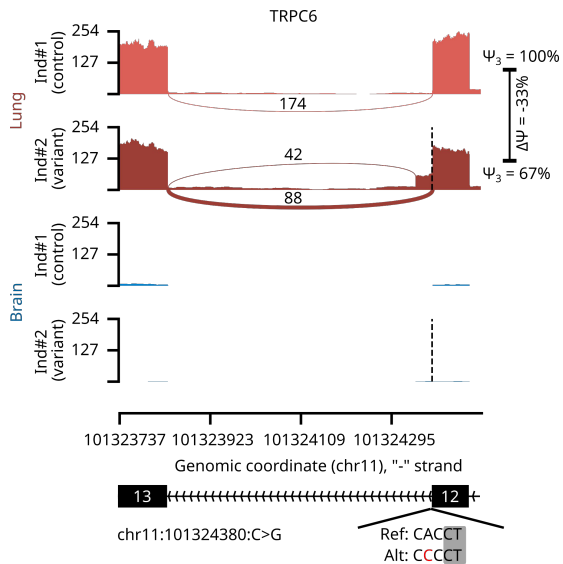
State-of-the-art sequence-based models poorly predict aberrant splicing



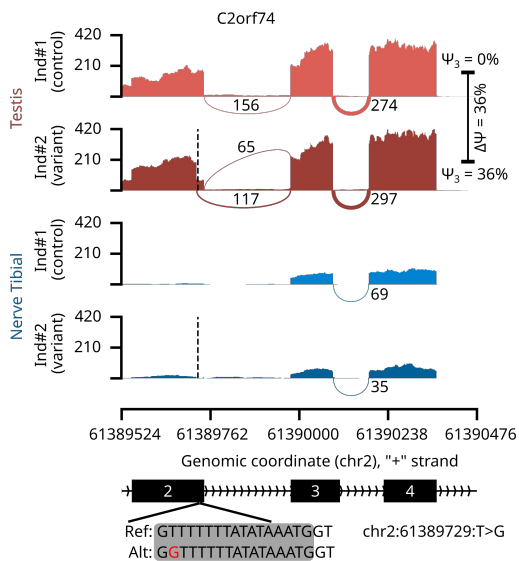
Ground truth: splicing outliers called using FRASER on all GTEx samples

Quantitative tissue-specific splice-site maps

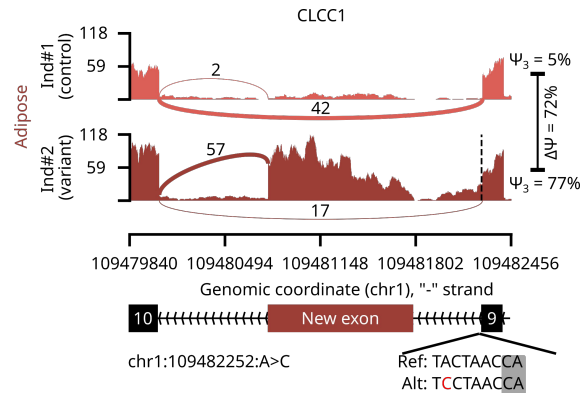
- unexpressed genes



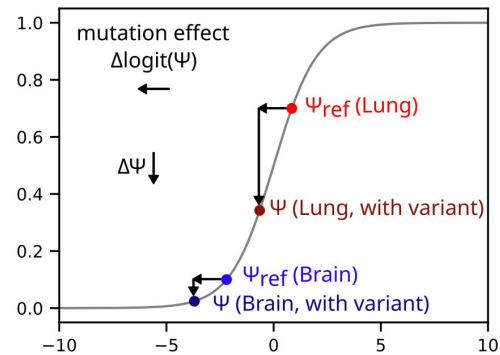
- unused splice sites



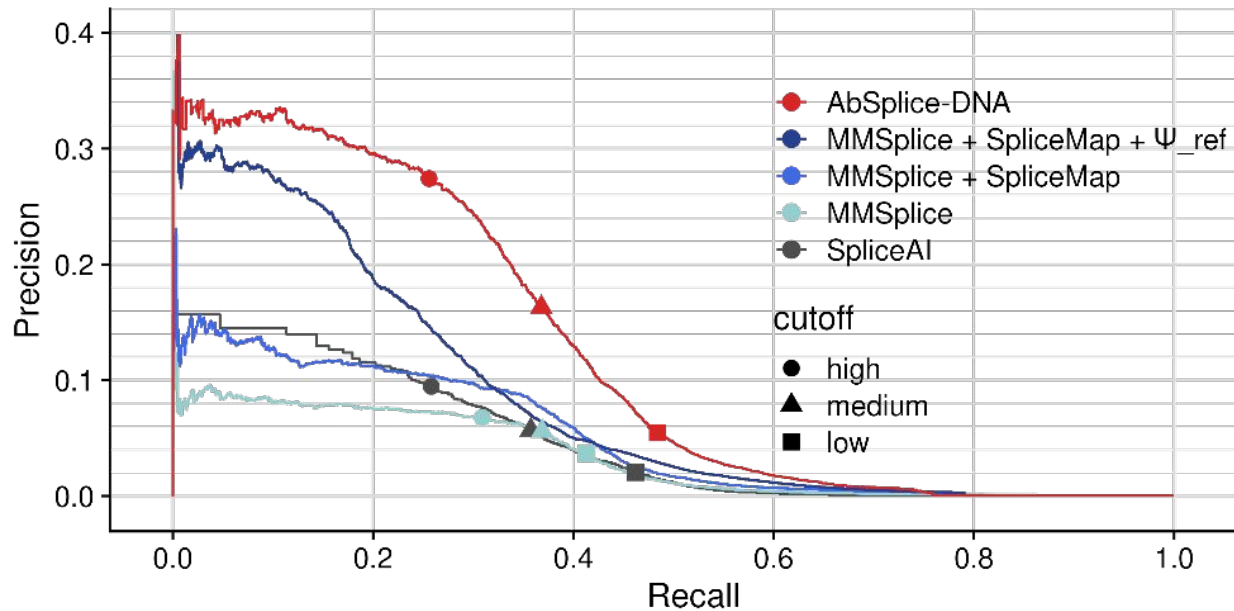
+ weak splice sites



Scaling law of splicing



AbSplice-DNA yields 3-fold improvement over state-of-the-art



Precomputed scores for all possible SNVs for all GTEx tissues

RNA-Seq for rare disease diagnostics - conclusion

- Added value of RNA-seq over Exome Seq / Genome Seq
- Specialized statistical methods and software to detect expression and splicing outliers
 - OUTRIDER
 - FRASER
 - DROP
- Outlook:
 - Proteomics - Kopajtich et al, medRxiv 2021
 - ATAC-seq - Celik et al, medRxiv 2023
 - Oncology - Cao et al, Genome Med 2024
 - Splicing outlier prediction from sequence - Wagner, Celik et al, Nat Genet, 2023
 - Expression outlier prediction from sequence - Hölzlwimmer et al, bioRxiv, 2023

Acknowledgements



Technical University of Munich

- **Julien Gagneur**, Christian Mertes, Ines Scheller, Felix Brechtmann, Nils Wagner, Nicholas Smith, M. Hasan Celik, Rebeka Luknárová, among others

Helmholtz Zentrum Munich

- Holger Prokisch, Mirjana Gusic, Laura Kremer, Robert Kopajtich, among others

Solve-RD European Commission Project:

- Anna Esteve, Kornelia Ellwanger, Leslie Matalonga, German Demidov, Joohyun Park, Josua Kegele, Davide Mei, Raffaella Minardi
- Other DATF and ERN members

All the patients included in the studies and their families

