Isocomp: Comparing Iso-Seq Isoform Profiles in Mendelian Disease Diagnosis for Trio Sequencing

Yutong Qiu, Bida Gu, Chia Sin Liew, Muhamad Sohail Raza, Chase Mateusiak, Rupesh Kesharwani, Evan Biederstedt

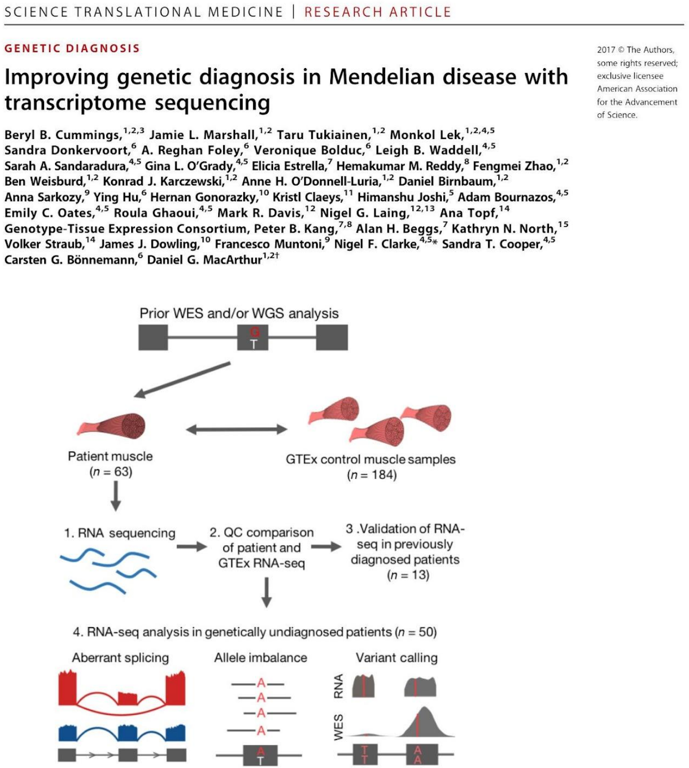
12 October 2022

Outline

- Background & Motivation
- Isocomp: comparing isoforms in trio sequencing
- Algorithm
- Results
- Future Directions
- Thank you

Background

Utilization of RNA transcriptomics in a clinical setting



SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

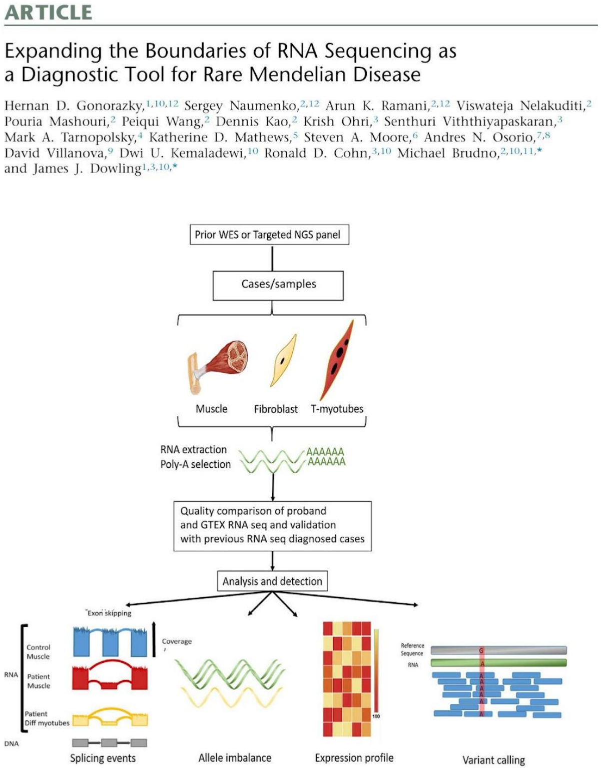
GENETIC DIAGNOSIS

2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science.

Improving genetic diagnosis in Mendelian disease with transcriptome sequencing

Beryl B. Cummings,[1,2,3] Jamie L. Marshall,[1,2] Taru Tukiainen,[1,2] Monkol Lek,[1,2,4,5] Sandra Donkervoort,[6] A. Reghan Foley,[6] Veronique Bolduc,[6] Leigh B. Waddell,[4,5] Sarah A. Sandaradura,[4,5] Gina L. O'Grady,[4,5] Elicia Estrella,[7] Hemakumar M. Reddy,[8] Fengmei Zhao,[1,2] Ben Weisburd,[1,2] Konrad J. Karczewski,[1,2] Anne H. O'Donnell-Luria,[1,2] Daniel Birnbaum,[1,2] Anna Sarkozy,[9] Ying Hu,[6] Hernan Gonorazky,[10] Kristl Claeys,[11] Himanshu Joshi,[5] Adam Bournazos,[4,5] Emily C. Oates,[4,5] Roula Ghaoui,[4,5] Mark R. Davis,[12] Nigel G. Laing,[12,13] Ana Topf,[14] Genotype-Tissue Expression Consortium, Peter B. Kang,[7,8] Alan H. Beggs,[7] Kathryn N. North,[15] Volker Straub,[14] James J. Dowling,[10] Francesco Muntoni,[9] Nigel F. Clarke,[4,5]* Sandra T. Cooper,[4,5] Carsten G. Bönnemann,[6] Daniel G. MacArthur[1,2]†

Prior WES and/or WGS analysis
G
T
Patient muscle
(n = 63)
GTEx control muscle samples
(n = 184)
1. RNA sequencing
2. QC comparison of patient and GTEx RNA-seq
3 .Validation of RNA-seq in previously diagnosed patients
(n = 13)
4. RNA-seq analysis in genetically undiagnosed patients (n = 50)
Aberrant splicing
Allele imbalance
Variant calling
A
A
A
A
A
T
RNA
WES
T
T
A
A



ARTICLE

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

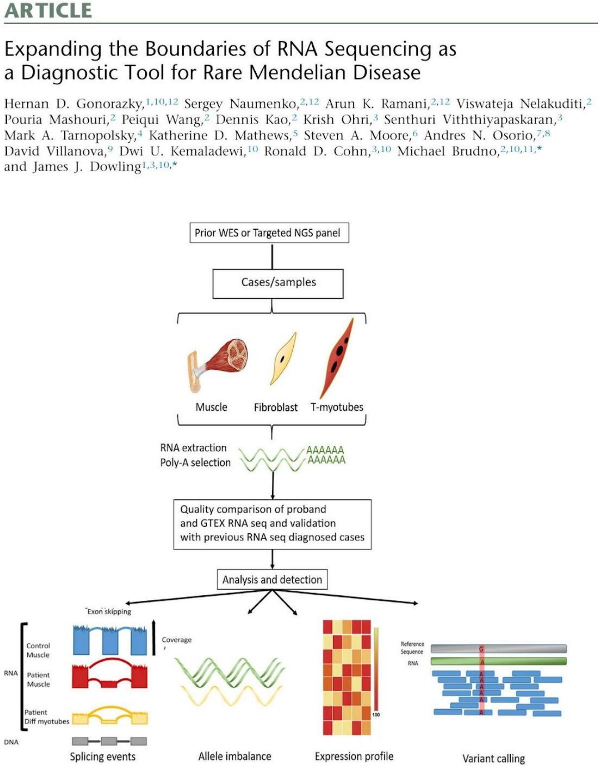
Hernan D. Gonorazky,[1,10,12] Sergey Naumenko,[2,12] Arun K. Ramani,[2,12] Viswateja Nelakuditi,[2] Pouria Mashouri,[2] Peiqui Wang,[2] Dennis Kao,[2] Krish Ohri,[3] Senthuri Viththiyapaskaran,[3] Mark A. Tarnopolsky,[4] Katherine D. Mathews,[5] Steven A. Moore,[6] Andres N. Osorio,[7,8] David Villanova,[9]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

Analysis outputs

- Splicing
- Allelic imbalance
- Variant calling
- Expression

Background

Utilization of RNA transcriptomics in a clinical setting



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,*]

Problem

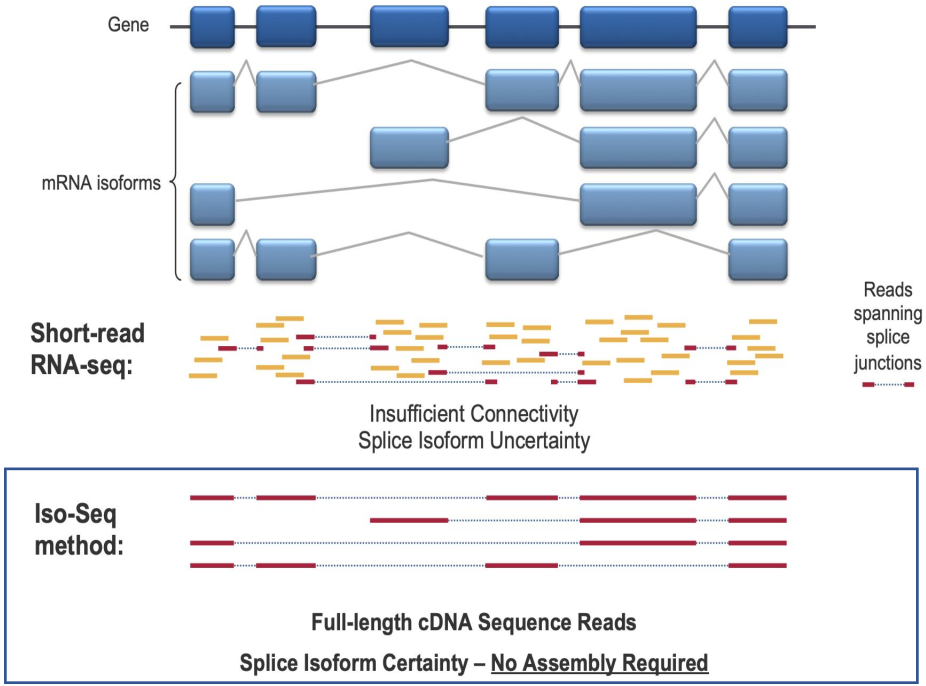
- Splicing
- Allelic imbalance
- Variant calling
- Expression

Short-reads not good enough for clinic: FPs, dodgy inferences

Even tissue-specific bulk RNA is not appropriate here! That’s why we use scRNAseq

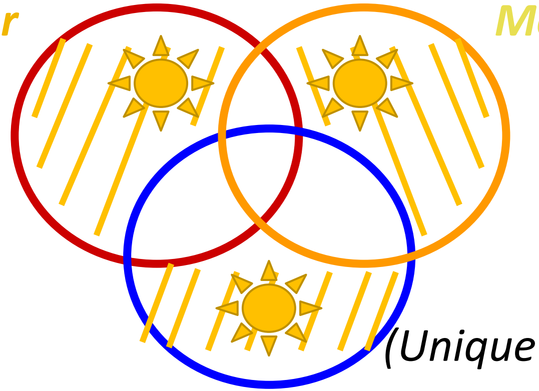
Motivation – Long Read RNA-seq

- **High-Quality Long Read RNA-seq Data: IsoSeq3**


Gene
mRNA isoforms
Short-read RNA-seq:
Reads spanning splice junctions
Insufficient Connectivity
Splice Isoform Uncertainty
Iso-Seq method:
Full-length cDNA Sequence Reads
Splice Isoform Certainty – No Assembly Required

- **Trio sequencing**

  Given trio-samples of isoforms, identify the unique casual isoform implicated in disease phenotype



Father

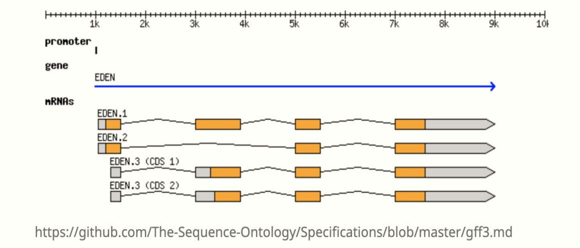
Mother

(Unique Isoforms)

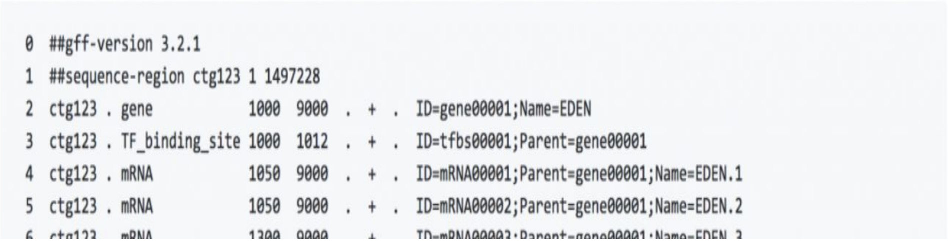
Child with disease phenotype

Existing Tools

- **gffcompare – overlap the exon coordinates. Inclusion/exclusion of the exons in isoforms**



https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

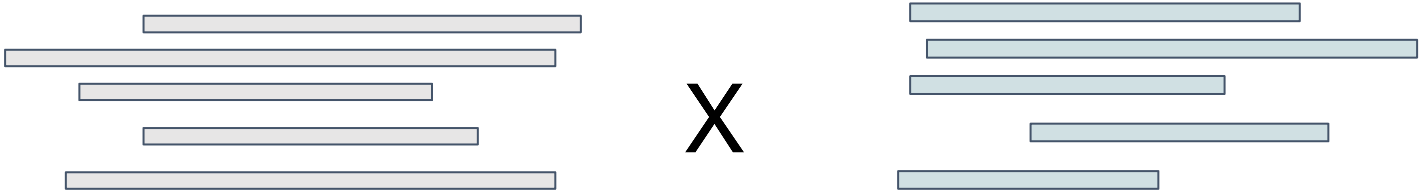


```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
```

Cons: Not designed for this problem! It misses biologically relevant sequence variation, questionable how well it works with gene fusions as well

- **All-against-All Alignments – exact sequence alignment**

Cons: Algorithmically cannot scale, uninterpretable for large-scale variants



Inputs & Data QC

Inputs

- HG002
- HG004
- HG005

(GIAB samples sequenced at BCM)

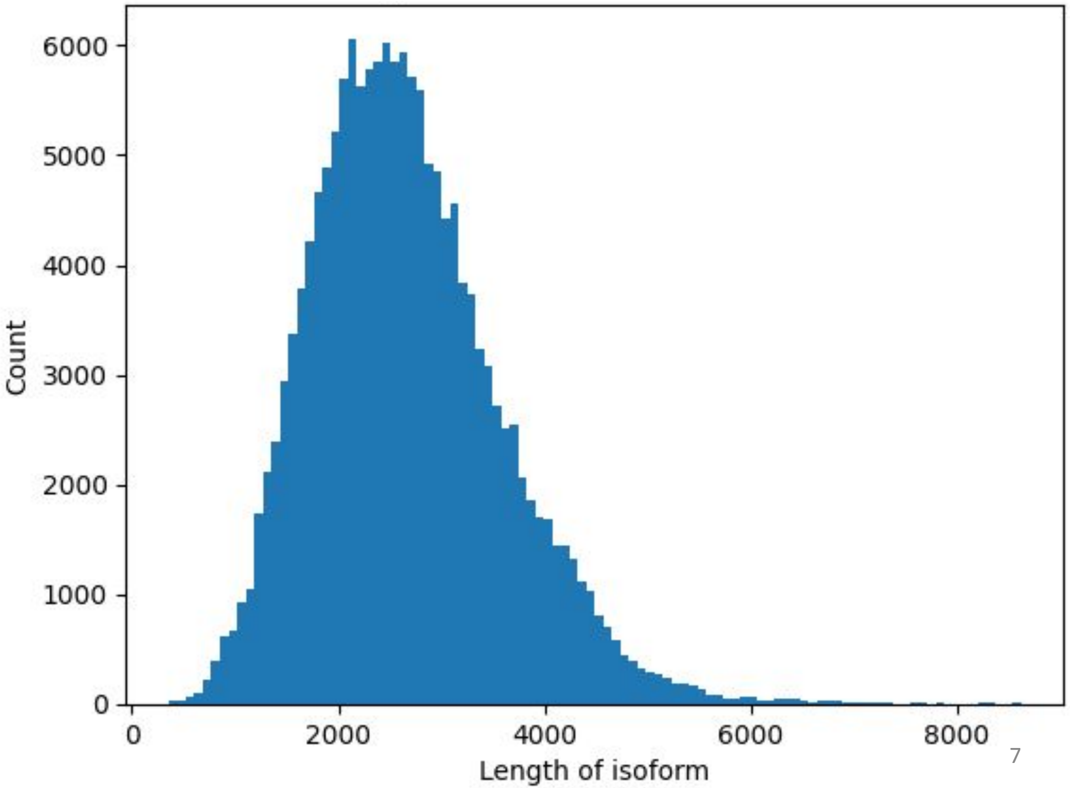
Average length

2661 bp

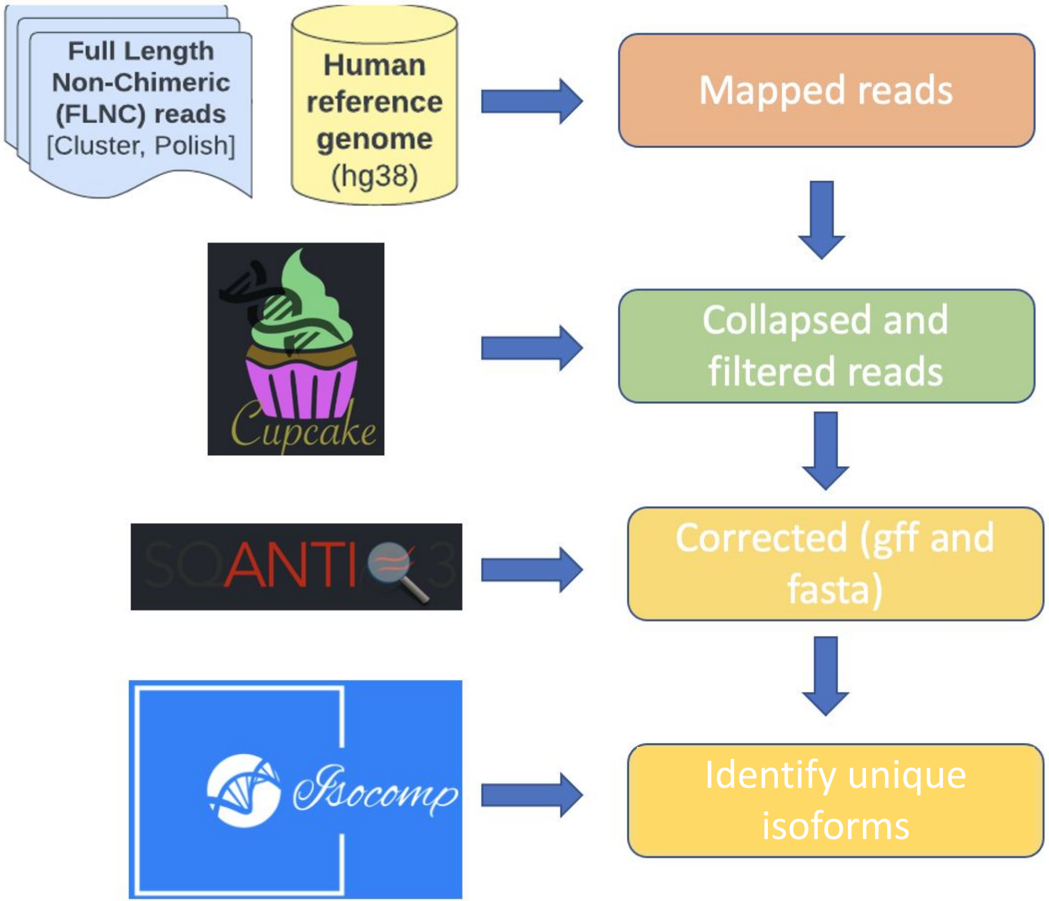
Standard Deviation

951 bp

Isoform length distribution in HG002 + HG004 + HG005



Workflow Overview



Full Length Non-Chimeric (FLNC) reads [Cluster, Polish]

Human reference genome (hg38)

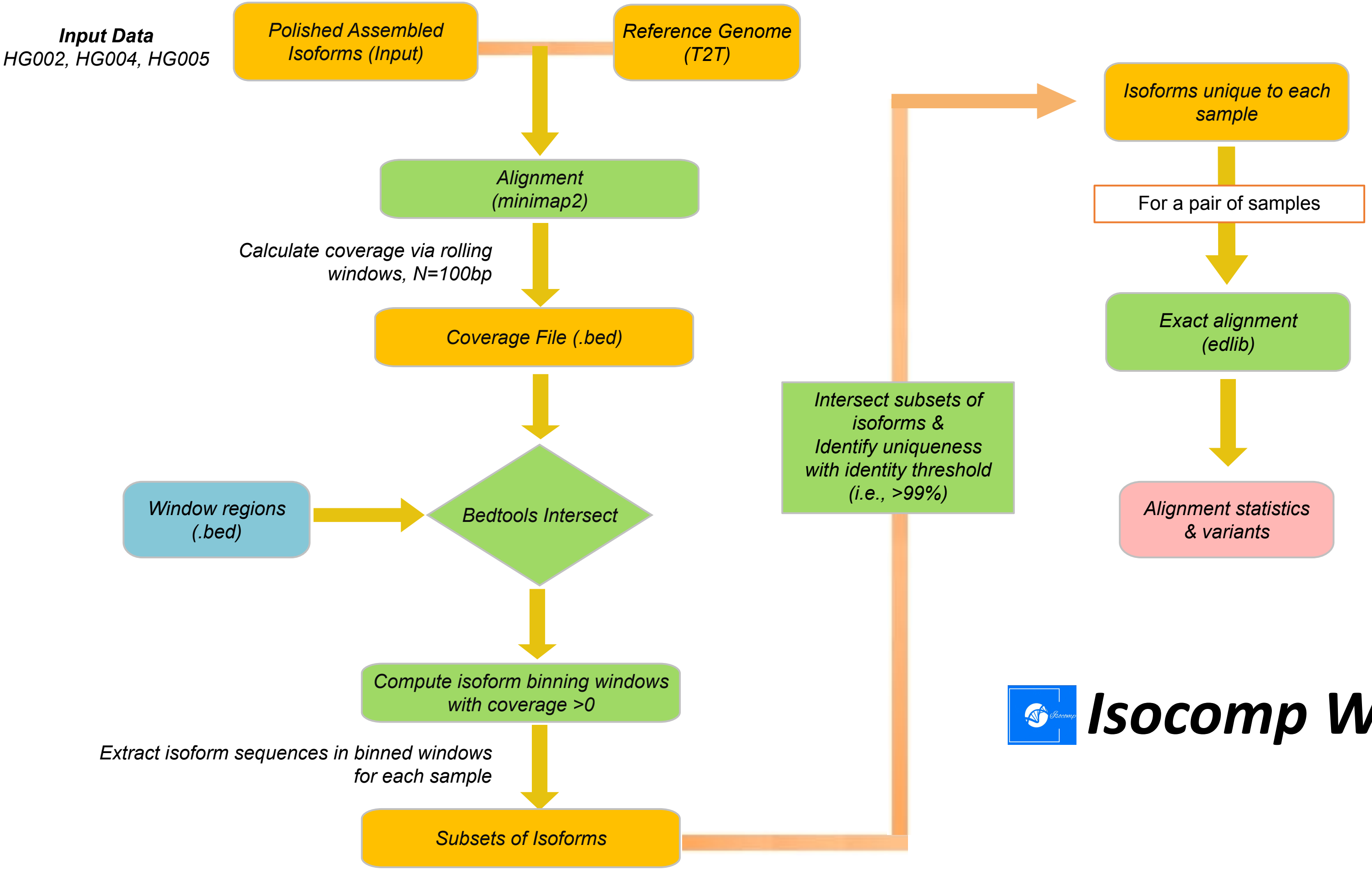
Mapped reads

Collapsed and filtered reads

Corrected (gff and fasta)

Identify unique isoforms

Isocomp Workflow



Input Data
HG002, HG004, HG005

Polished Assembled Isoforms (Input)

Reference Genome (T2T)

Alignment (minimap2)

Calculate coverage via rolling windows, N=100bp

Coverage File (.bed)

Window regions (.bed)

Bedtools Intersect

Compute isoform binning windows with coverage >0

Extract isoform sequences in binned windows for each sample

Subsets of Isoforms

Intersect subsets of isoforms & Identify uniqueness with identity threshold (i.e., >99%)

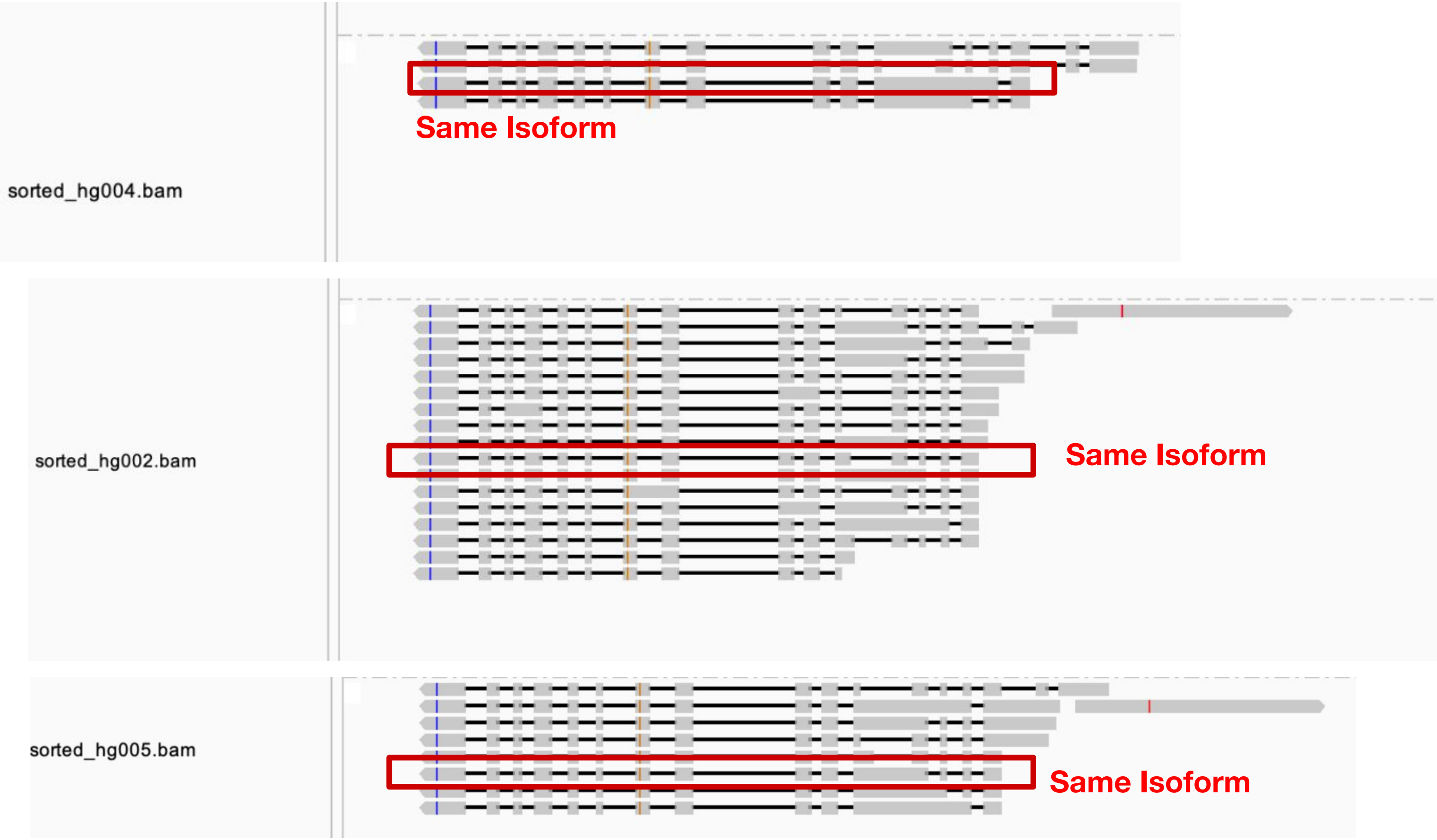
Isoforms unique to each sample

For a pair of samples

Exact alignment (edlib)

Alignment statistics & variants

Examples of different isoform composition in HG002, HG004 & HG005

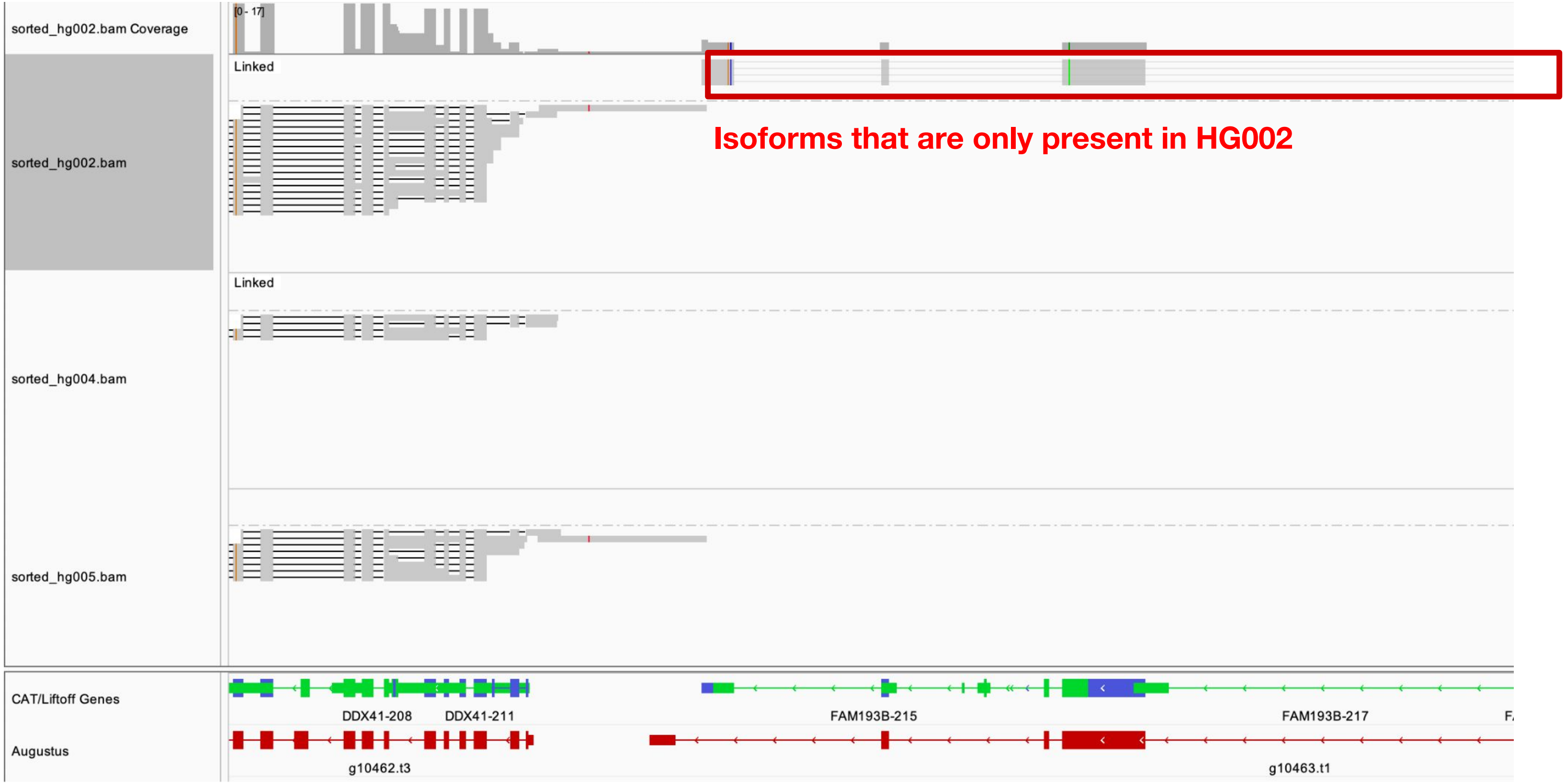


Same Isoform

Same Isoform

Same Isoform

Examples of different isoform composition in HG002, HG004 & HG005

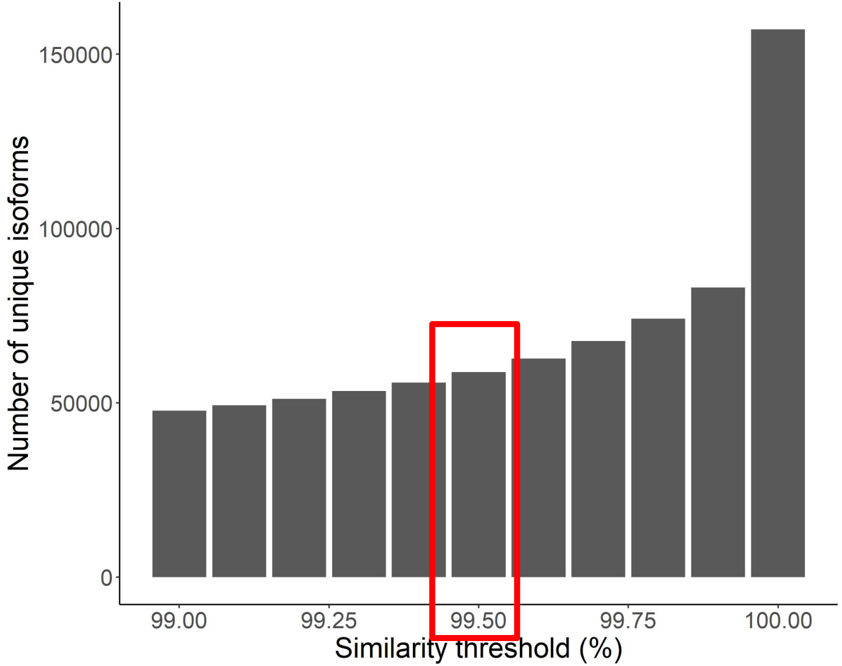


Isoforms that are only present in HG002

Number of isoforms that are unique to at least one sample

Two isoforms are different if the percent matched bases between them is < 99.5% (~15 mismatches).

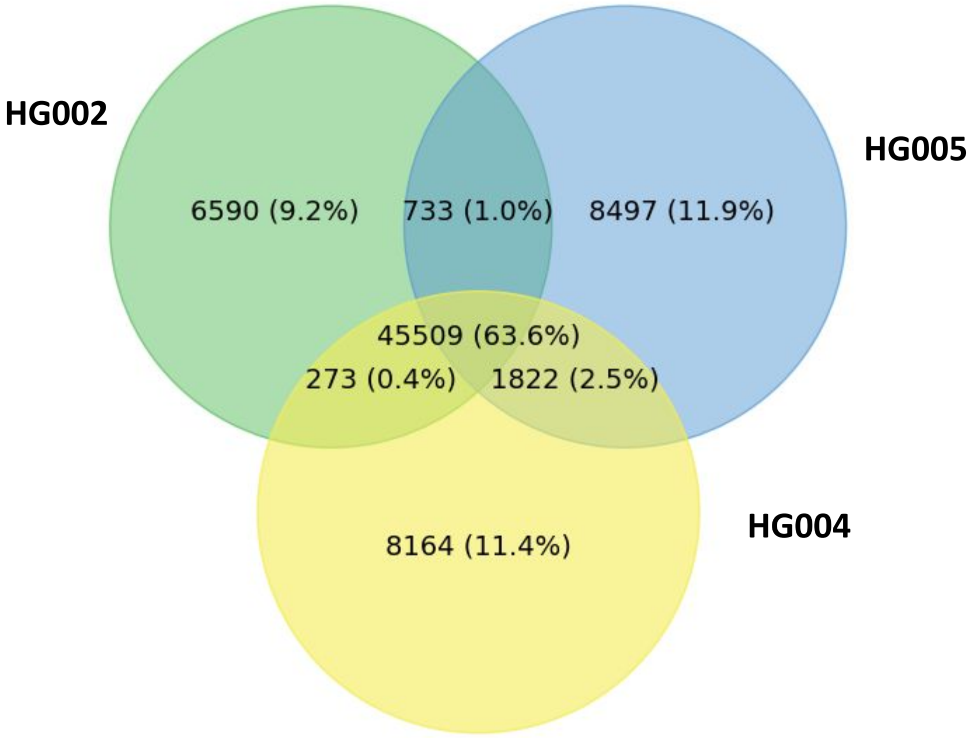
$$\textbf{Similarity} := \frac{\text{matched bases}}{\text{all bases in REF isoform}}$$



Total number of isoforms
HG002 + HG004 + HG005
389644

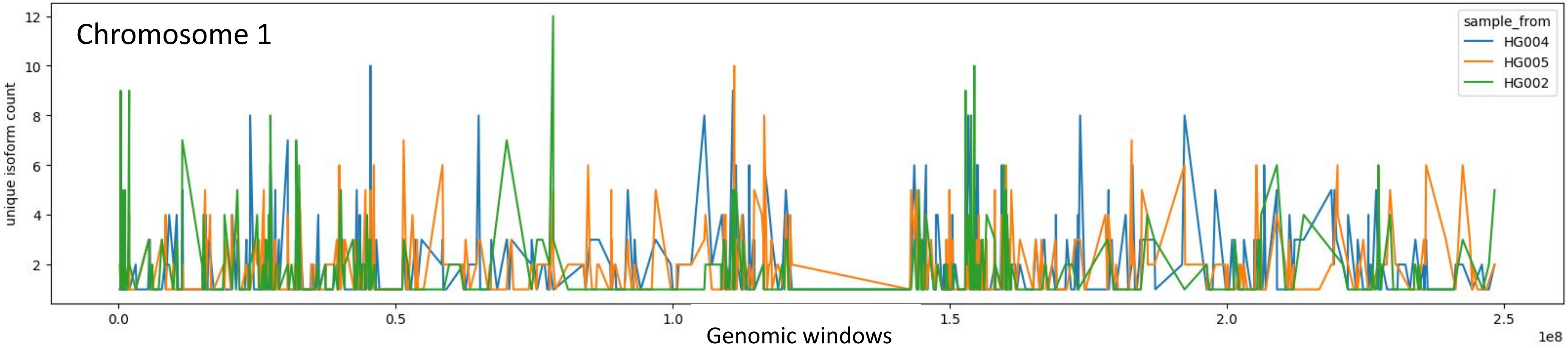
Applying the similarity threshold **reduces false positives** of uniqueness detection, which increases tolerance for differences in lengths of the isoforms

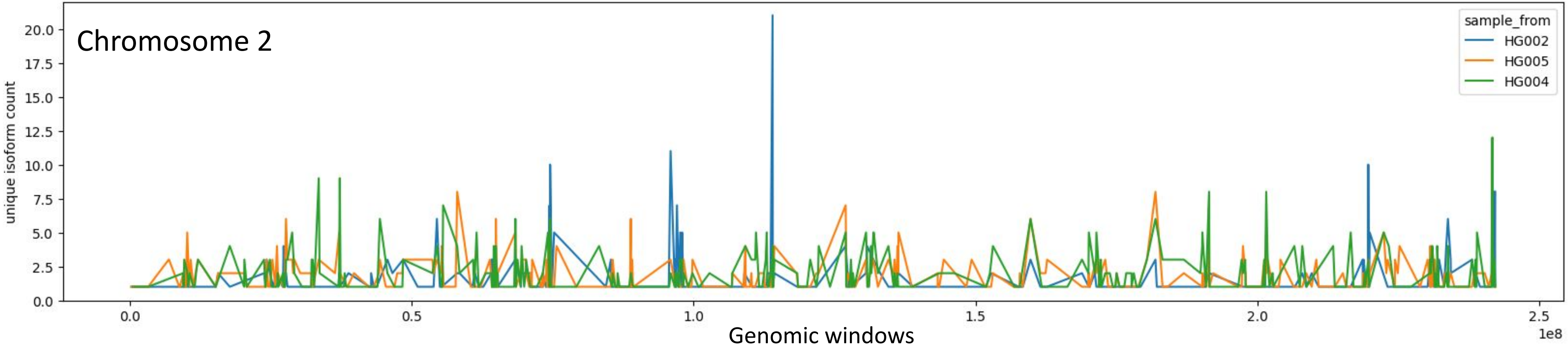
Number of isoforms shared by/unique to samples



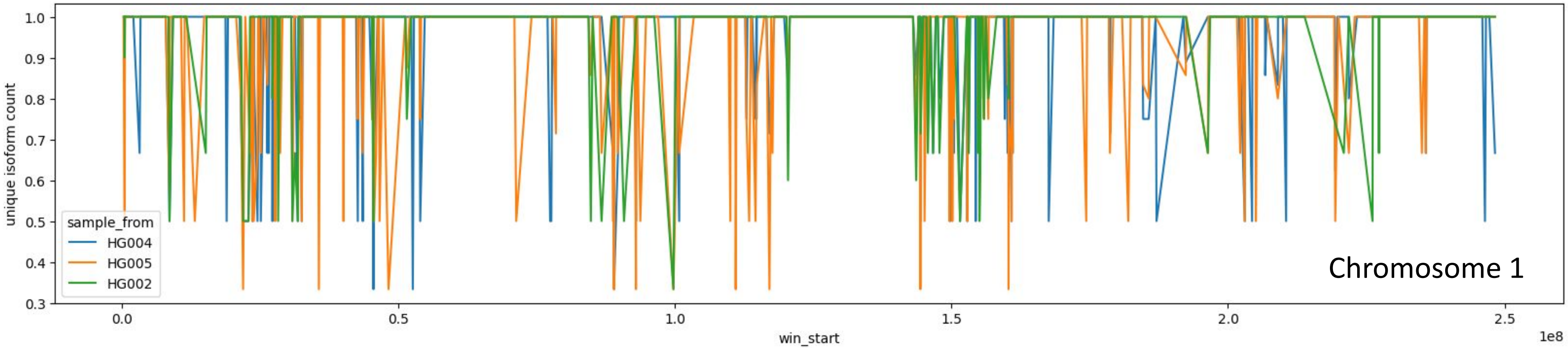
- HG005 and HG004 are more similar than HG002 in terms of isoform composition
- Majority of the isoforms are shared among three samples

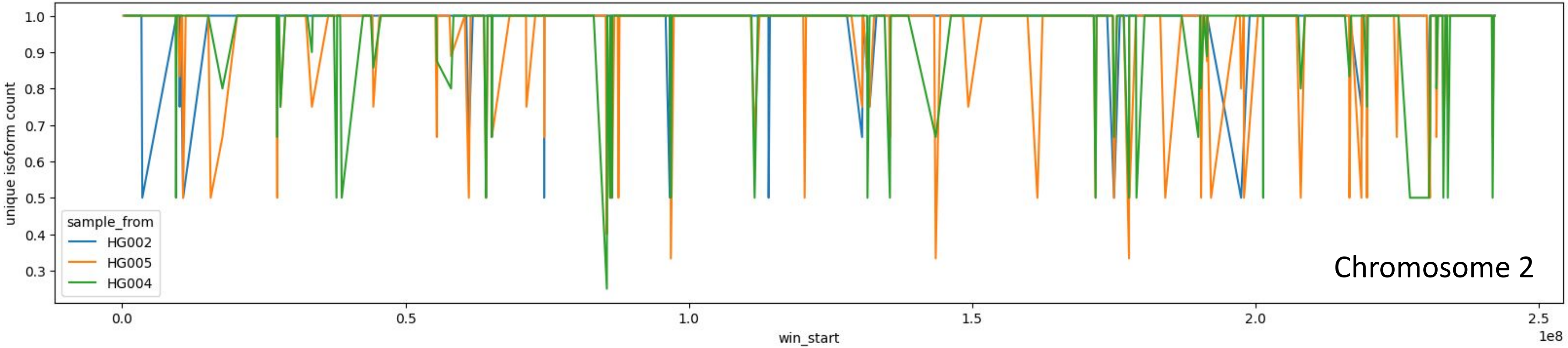
Total number of isoforms unique to ONE sample





Normalized number of isoforms unique to ONE sample





Next steps

- **Algorithmic strategy for gene fusions**

  After the alignment to T2T and windowing, we should really remove potential gene fusion isoforms, as they align to multiple regions on the genome

- **Categorize mismatches**

  At the moment, we are using the metric of “Percentage matched bases”. But mismatches at the ends of isoforms may not reflect isoform sequence variation…

- **Biological use cases**

  Allow to quickly query gene of interest?

- **Wrap up package**

  Python library with C++ speed-up; could port to R package as well

Thank you!

Baylor College of Medicine

HGSC HUMAN GENOME SEQUENCING CENTER

DNAnexus

PacBio

Oxford NANOPORE Technologies

THE STRUCTURAL VARIANT CRYING CLUB PRESENTS...

PAN-STRUCTURAL VARIATION HACKATHON

...in the Cloud!

OCTOBER 9-12, 2022

HYBRID SV PAN HACKATHON

TEXAS MEDICAL CENTER, HOUSTON

TAME FEROCIOUS STRUCTURAL VARIATIONS!

SEE AWESOME GRAPH GENOMES!

TACKLE HIDEOUS MUTATION ANNOTATIONS!

AND ADVANCE THE FRONTIERS OF PERSONALIZED MEDICINE!

PRIZES FOR TOP PARTICIPANTS!

SIGN UP AT WWW.HGSC.BCM.EDU/HACKATHON

DNAnexus

Baylor College of Medicine

PACBIO NANOPORE

RICE UNIVERSITY

HGSC HUMAN GENOME SEQUENCING CENTER

- **Baylor College of Medicine**
- **HGSC**
- **Rice University**
- **DNANexus**
- **PacBio**
- **Oxford Nanopore**

- Ben Busby & Fritz Sedlazeck
- Richard Gibbs
- Todd Treangen
- Everyone at the hackathon, local & remote!