A simulation framework to benchmark long-read RNA quantification tools

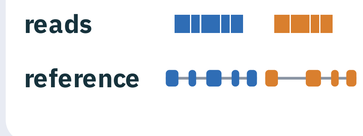
Hiep Dang[1,2], Andrew Lonsdale[1,2], Alicia Oshlack[1,2]

[1] Peter MacCallum Cancer Centre, Melbourne · [2] Sir Peter MacCallum Department of Oncology, The University of Melbourne

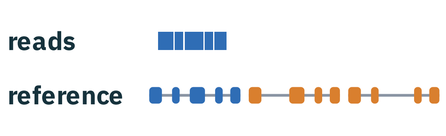
The problem

Tools count reads **against a reference**, which can differ from the reads:

reads and reference match



reference has extra isoforms not in reads



reads have isoforms not in reference



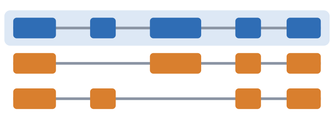
Real data has no known truth. **Which tool can we trust**, and in what context?

Simulation framework

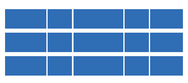
Separate reads & reference

Simulated reads: true count known.

genes & isoforms
canonical + ≤20 more



100 reads per transcript
simulated, perfect, full-length

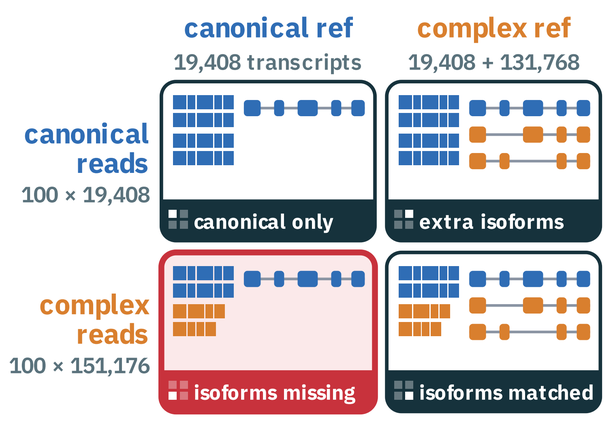


reference
isoforms a read can go to



Vary reads × reference

canonical: canonical isoform only
complex: + isoforms of the same gene



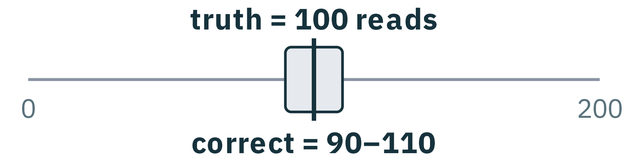
Each scenario is simulated separately

Tool groups

Tool	Read handling	Algorithm
Transcriptome-based		
Probabilistic assignment		
oarfish	read coverage	EM
lr-kallisto	k-mer match	pseudo + EM
TranSigner	alignment score	EM
Filtered assignment		
NanoCount	3′-end filter	EM
FLAIR	indel filter	best hit
Genome-based		
Intron-chain matching		
IsoQuant	intron-chain match	unique count
bambu	read classes	EM
Annotation-extending quantification		
StringTie2	splice graph	network flow
ESPRESSO	reads → new isoforms	EM

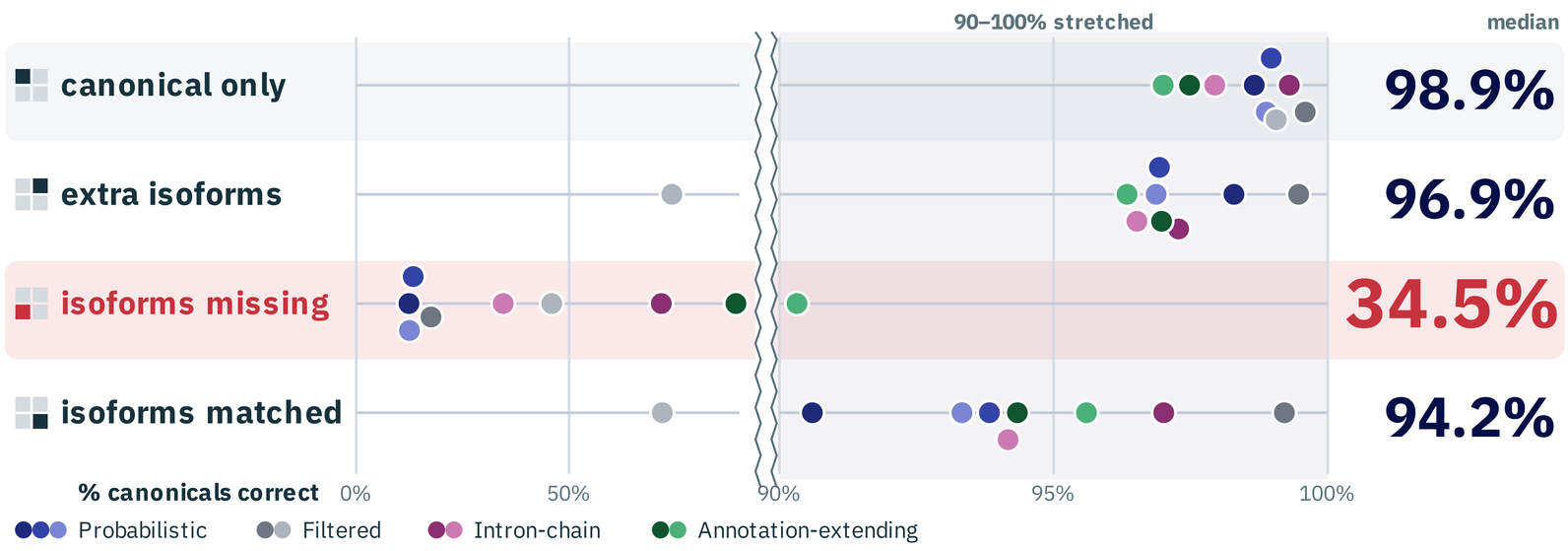
Evaluation metric

→ 9 tools × 4 runs



Score = % canonicals correct
one dot per tool and run in 1

1 Tool estimates are never completely correct, especially if the reference is missing isoforms

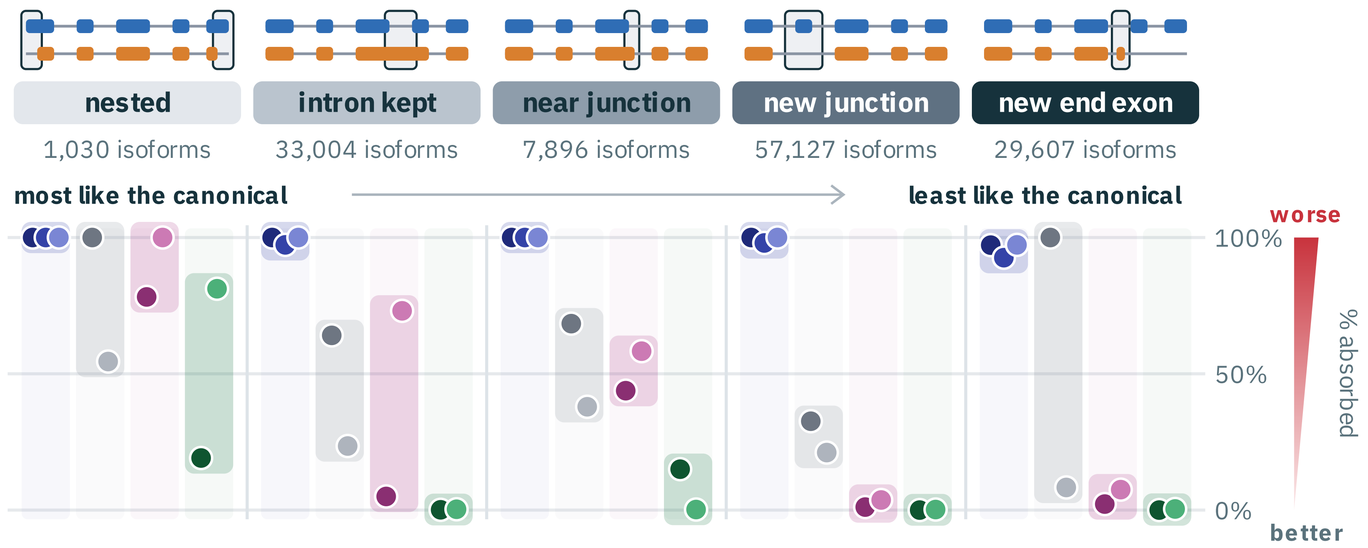


2 The type of excluded isoform matters

isoforms missing

absorbed =
unmatched reads
↓
on the canonical

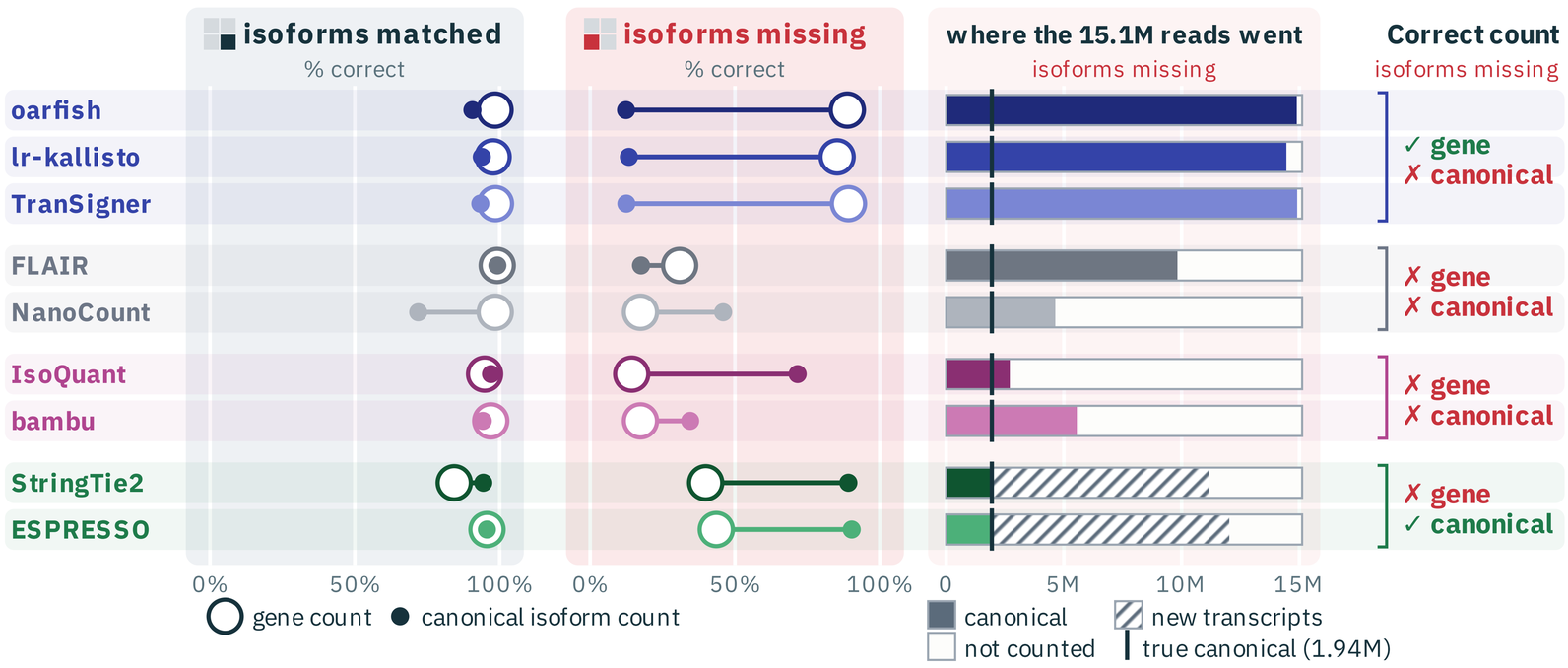
Probabilistic
Filtered
Intron-chain
Annotation-extending



Unmatched reads are absorbed into the canonical count: nested reads by most tools, all reads by probabilistic tools.

Check your gene

3 Are gene counts correct or isoform counts correct?



TAKE HOME

1. Quantification tools are **never completely correct**, even with perfect data. They struggle **most** when isoforms are **missing from the reference**.
2. The more a missing isoform **looks like the canonical**, the more of its reads **all tools** assign to the canonical isoform.
3. **Probabilistic** tools keep the **gene count** correct by allocating all the reads; **annotation-extending** tools keep the **transcript counts** right.

Next step: add sequencing errors

Now: perfect reads



Next: reads with errors

