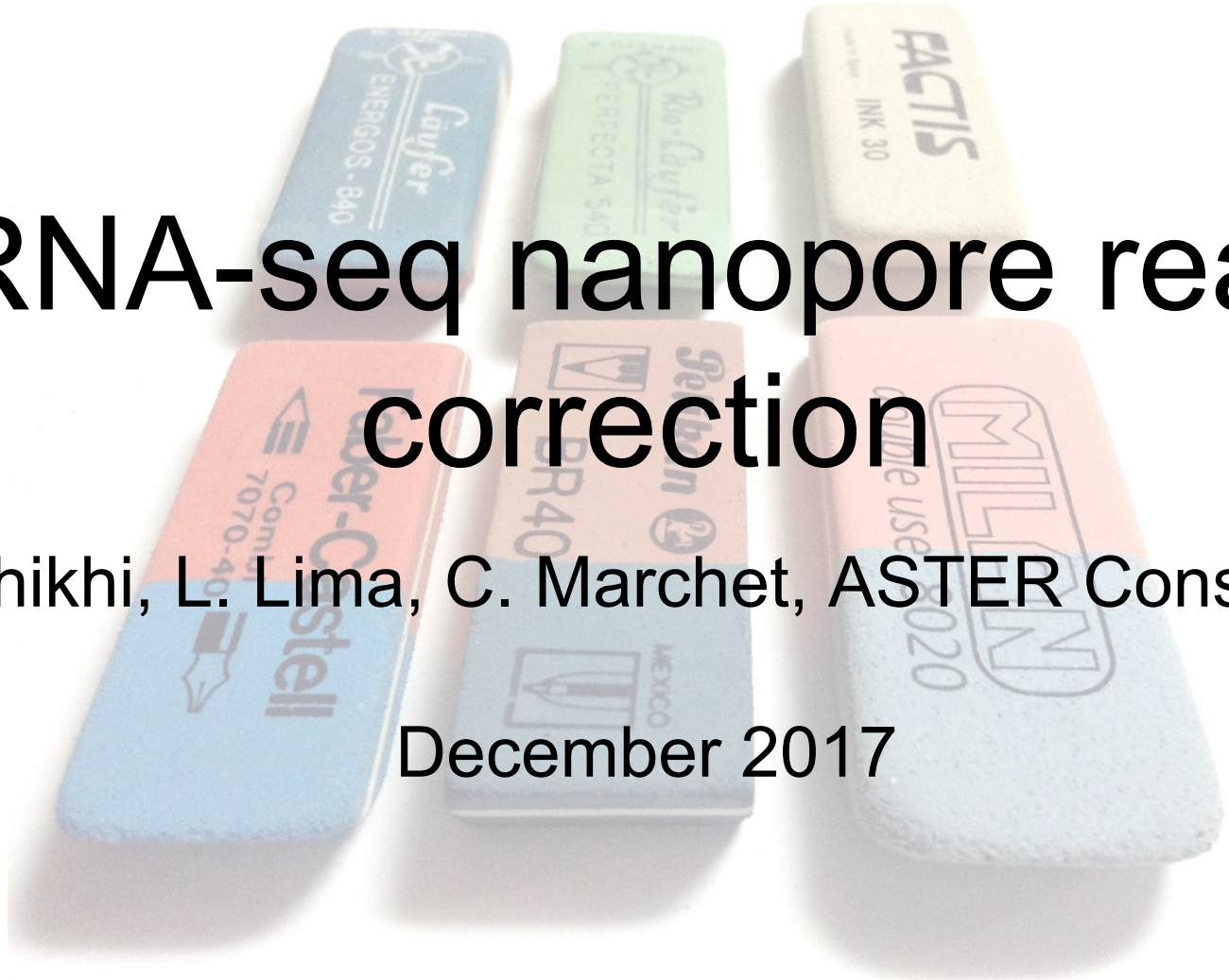


RNA-seq nanopore read correction

R. Chikhi, L. Lima, C. Marchet, ASTER Consortium

December 2017

Motivation

- Emerging cDNA and RNA nanopore data
- No dedicated error-correction tool yet

We evaluate existing DNA error-correction tools on RNA-seq data.

- Error rate? Lose coverage?
- Gene families collapsed? Isoform bias? (=overcorrection?)

Dataset

mouse brain cDNA

1D

sequenced @ Genoscope

filtered out mtRNA and rRNA

750k reads

Error-correction tools

Long+short (*hybrid*):

LoRDEC	**DNA PacBio/ONT**	**path in dBG**
PBcR	**mRNA/DNA PacBio/ONT**	**align short->long, consensus**
NaS	**DNA ONT**	**align short->long, read recruitment, assembly**
Proovread	**DNA PacBio**	**align short->long, consensus**
CoLorMap	**simulated**	**align short->long, read recruitment, assembly**

Long reads only (*non-hybrid or self*):

daccord	**DNA PacBio**	**path in dBG**
LoRMA	**DNA PacBio/ONT**	**path in dBG, multi-iterations**
MECAT	**DNA PacBio/ONT**	**k-mer based align all-pairs long, consensus**
Pbdagcon	**DNA PacBio**	**BLASR alignment, partial order graph**

Not tested:
Canu (option to correct ONT reads);
HG-Color;
HALC;
HECIL;
MIRCA;
Jabba;
Nanocorr (specific for ONT);
LSCPlus (specific for long reads RNA);

Qualitative observations (spoilers)

- Original data: 16.5% error rate
- Best correctors: 0.5% error rate
- Some reads are dropped
- Some tools split reads, some don’t
- Same with trimming
- Trend: fast = correct less, slow = correct more

Evaluation methodology

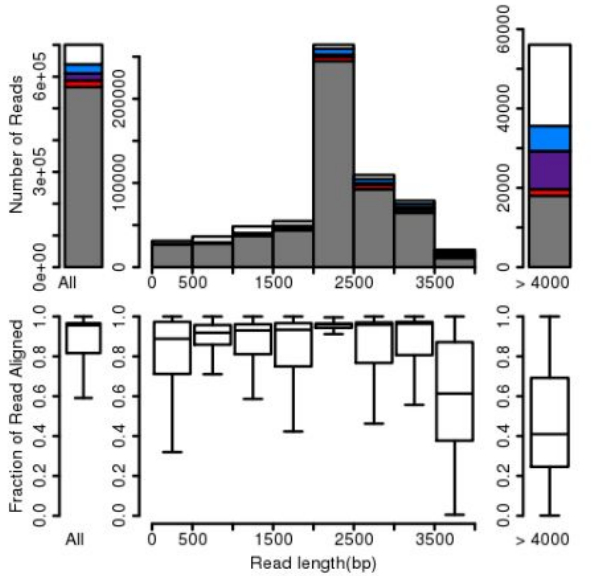
- AlignQC

Alignment analysis *91.2%* reads aligned *80.4%* bases aligned *(of aligned reads)*

Read Stats		
Total reads	700,452	
- Unaligned reads	61,526	8.8%
- Aligned reads	638,926	91.2%
--- Single-align reads	567,331	81.0%
--- Gapped-align reads	20,468	2.92%
--- Chimeric reads	51,127	7.30%
----- Trans-chimeric reads	28,615	4.09%
----- Self-chimeric reads	22,512	3.21%
Base Stats *(of aligned reads)*		
Total bases	1,551,053,577	
- Unaligned bases	304,649,514	19.6%
- Aligned bases	1,246,404,063	80.4%
--- Single-aligned bases	1,193,725,859	77.0%
--- Other-aligned bases	9,833,971	0.63%

Unaligned
Trans-chimeric alignment
Self-chimeric alignment
Gapped alignment
Single alignment

Summary [pdf]



More evaluation methodology

- Raw and corrected reads mapped to genome (GMAP) and transcriptome (BWA-MEM)

Custom plots and simulations to look at:

- Whether correction drops low-abundance isoforms
- Whether reads are corrected towards the major isoform

Performance

Tool	Hybrid error correctors				Self error correctors			
	LoRDEC	NaS	PBcR	Proovread	daccord	LoRMA	MECAT	pbdagcon
Time (wall-clock)	**2.4h**	~63.2h	116h	107.1h	**7.4h**	**3.4h**	**0.3h**	**6.2h**
Peak memory usage	**5.6Gb**	N/A	166.5Gb	53.6Gb	**27.2Gb**	79Gb	**9.9Gb**	**27.2Gb**

32 threads on Intel Core Processor (Broadwell) @ 1999 MHz

Number of error-corrected reads

Same #reads

LoRDEC
Proovread untrimmed
pbdagcon

Split and/or discard

All others

Number of error-corrected reads

Same #reads

LoRDEC
Proovread untrimmed
pbdagcon

Split and/or discard

All others

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC	NaS	PBcR	Proovread untrim.	Proovread trim.	daccord	daccord trimmed	LoRMA	MECAT	pbdagcon
# reads (millions)	0.74	0.74	0.61	1.32	0.74	0.62	0.67	0.83	1.54	0.49	0.77

Mapping error-corrected reads

Much improved mapping rate
from **83.5** %
to up to **99** %

Mapping error-corrected reads

Much improved mapping rate
from **83.5** %
to up to **99** %

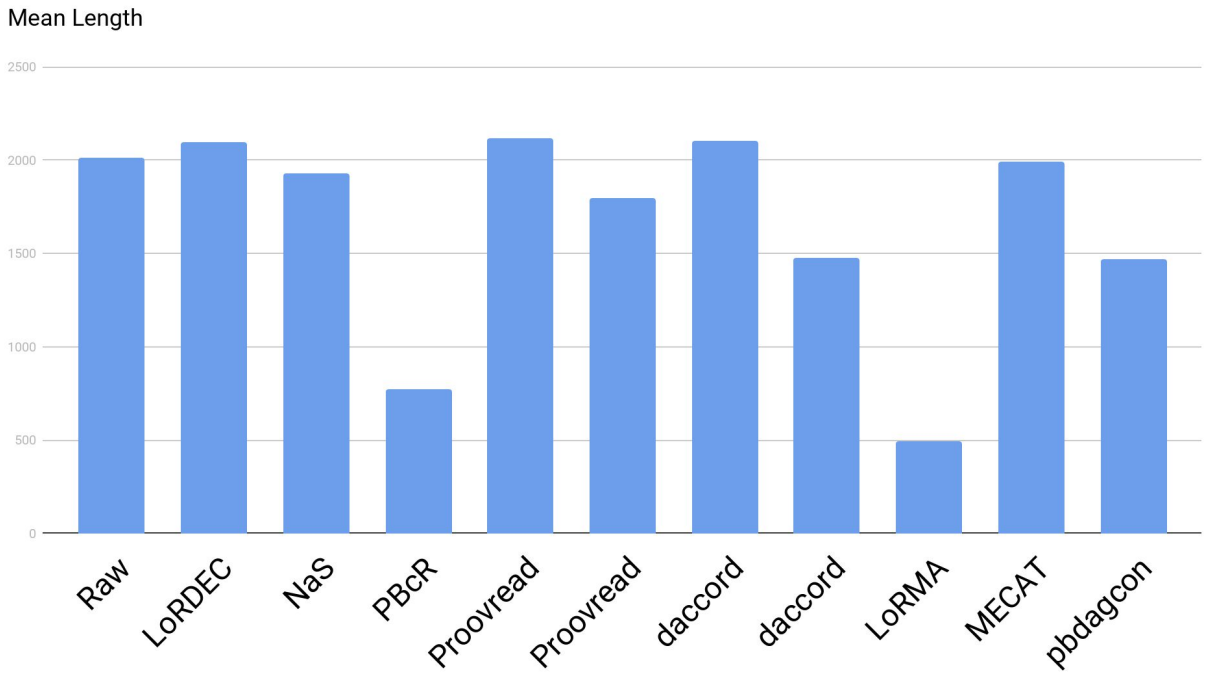
Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC	NaS	PBcR	Proovread untrim.	Proovread trim.	daccord	daccord trimmed	LoRMA	MECAT	pbdagcon
# reads	740 776	740 776	619 172	1 321 299	738 224	626 272	675 463	839 711	1 540 032	494 645	778 264
mapped reads %	83.5	85.5	98.7	99.2	85.5	98.9	92.5	94.0	99.4	99.4	98.2

Mapped bases in error-corrected reads

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC	NaS	PBcR	Proovread untrim.	Proovread trim.	daccord	daccord trimmed	LoRMA	MECAT	pbdagcon
# reads	740 776	740 776	619 172	1 321 299	738 224	626 272	675 463	839 711	1 540 032	494 645	778 264
mapped reads	83.5%	85.5%	98.7%	99.2%	85.5%	98.9%	92.5%	94.0%	99.4%	99.4%	98.2%
% mapped bases in mapped reads	89.0	90.6	97.5	99.2	92.4	99.5	92.5	94.7	99.1	96.9	97.0

Same trend as previous slide..

Mean length of error-corrected reads



Overall remarks on error-corrected reads

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC	NaS	PBcR*	Proovread untrim.	Proovread trim.	daccord	daccord trimmed	LoRMA*	MECAT	pbdagcon
# reads	740 776	740 776	619 172	1 321 299	738 224	626 272	675 463	839 711	1 540 032	494 645	778 264
mapped reads	83.5%	85.5%	98.7%	99.2%	85.5%	98.9%	92.5%	94.0%	99.4%	99.4%	98.2%
mean length	2010	2096	1930	775	2117	1796	2102	1475	496	1994	1472

Bottom line:

1. PBcR and LoRMA tend to split reads into short well-corrected subreads (long range connectivity is lost);

*

Overall error-corrected reads stats

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC	NaS	PBcR*	Proovread untrim.	Proovread trim.	daccord	daccord trimmed	LoRMA*	MECAT*	pbdagcon
# reads	740 776	740 776	619 172	1 321 299	738 224	626 272	675 463	839 711	1 540 032	494 645	778 264
mapped reads	83.5%	85.5%	98.7%	99.2%	85.5%	98.9%	92.5%	94.0%	99.4%	99.4%	98.2%
mean length	2010	2096	1930	775	2117	1796	2102	1475	496	1994	1472

Bottom line:

1. PBcR and LoRMA tend to split reads into short well-corrected subreads (long range connectivity is lost);
2. MECAT tends to eliminate many not well-corrected or short reads from the input;

Overall error-corrected reads stats

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC*	NaS+	PBcR*	Proovread untrim*	Proovread trim.+	daccord+	daccord trimmed+	LoRMA*	MECAT*	pbdagcon+
# reads	740 776	740 776	619 172	1 321 299	738 224	626 272	675 463	839 711	1 540 032	494 645	778 264
mapped reads	83.5%	85.5%	98.7%	99.2%	85.5%	98.9%	92.5%	94.0%	99.4%	99.4%	98.2%
mean length	2010	2096	1930	775	2117	1796	2102	1475	496	1994	1472

Bottom line:

1. PBcR and LoRMA tend to split reads into short well-corrected subreads (long range connectivity is lost);
2. MECAT tends to eliminate many not well-corrected or short reads from the input;
3. LoRDEC and Proovread untrimmed corrections are underwhelming;

\+ +

Correction accuracy

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC*+	NaS++	PBcR*+	Proovread untrim*+	Proovread trim.++	daccord+*	daccord trim++	LoRMA*+	MECAT*+	pbdagcon+*
% per-base error rate	13.6	4.1	0.4	0.6	2.6	0.2	5.5	4.2	2.8	4.5	5.8

Bottom line:

1. Hybrid error correctors have a natural advantage here (depth + low error rate from Illumina);
2. daccord and pbdagcon were underwhelming in this measure;

How homopolymers are corrected

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC*++	NaS+++	PBcR*++	Proovread untrim*++	Proovread trim.+++	daccord+**	daccord trim++*	LoRMA*+*	MECAT*+*	pbdagcon+**
% deletion homopolymers errors	2.9	0.7	<0.1	<0.1	0.4	<0.1	2.1	2	1.8	2	2.3
% insertion homopolymers errors	0.3	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1

Bottom line:

1. Hybrid error correctors have a natural advantage here (depth + Illumina has less homopolymer errors);
2. All self correctors were underwhelming in this measure;

How homopolymers are corrected

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC*++	NaS+++	PBcR*++	Proovread untrim*++	Proovread trim.+++	daccord+**	daccord trim++*	LoRMA*+*	MECAT*+*	pbdagcon+**
% deletion homopolymers errors	2.9	0.7	<0.1	<0.1	0.4	<0.1	2.1	2	1.8	2	2.3
% insertion homopolymers errors	0.3	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1

Trimming of badly corrected regions

Bottom line:

1. Hybrid error correctors have a natural advantage here (depth + Illumina has less homopolymer errors);
2. All self correctors were underwhelming in this measure (not their fault?);

Are gene families collapsed?

Tool	Raw	Hybrid error correctors					Self error correctors				
	Raw	LoRDEC*+++	NaS++++	PBcR*+++	Proovread untrim*+++	Proovread trim.++++	daccord+**+	daccord trim++*+	LoRMA*+**	MECAT*+**	pbdagcon+**+
number of genes	16.9k	16.9k	15k	15.4k	16.7k	14.5k	15.7k	14k	6.6k	10.3k	13.2k

Bottom-line

1. LoRMA and MECAT lose a lot of genes, likely not preserving gene families;

To trim or not to trim?

	Proovread	Proovread trim.	daccord	daccord trimmed
mapped reads	85.5%	98.9%	92.5%	94.0%
mapped bases[1]	92.4%	99.5%	92.5%	94.7%
per-base error rate[2]	2.6%	0.2%	5.5%	4.2%

Trimmed output of tools:

\+ more reads and bases are mapped, less errors;

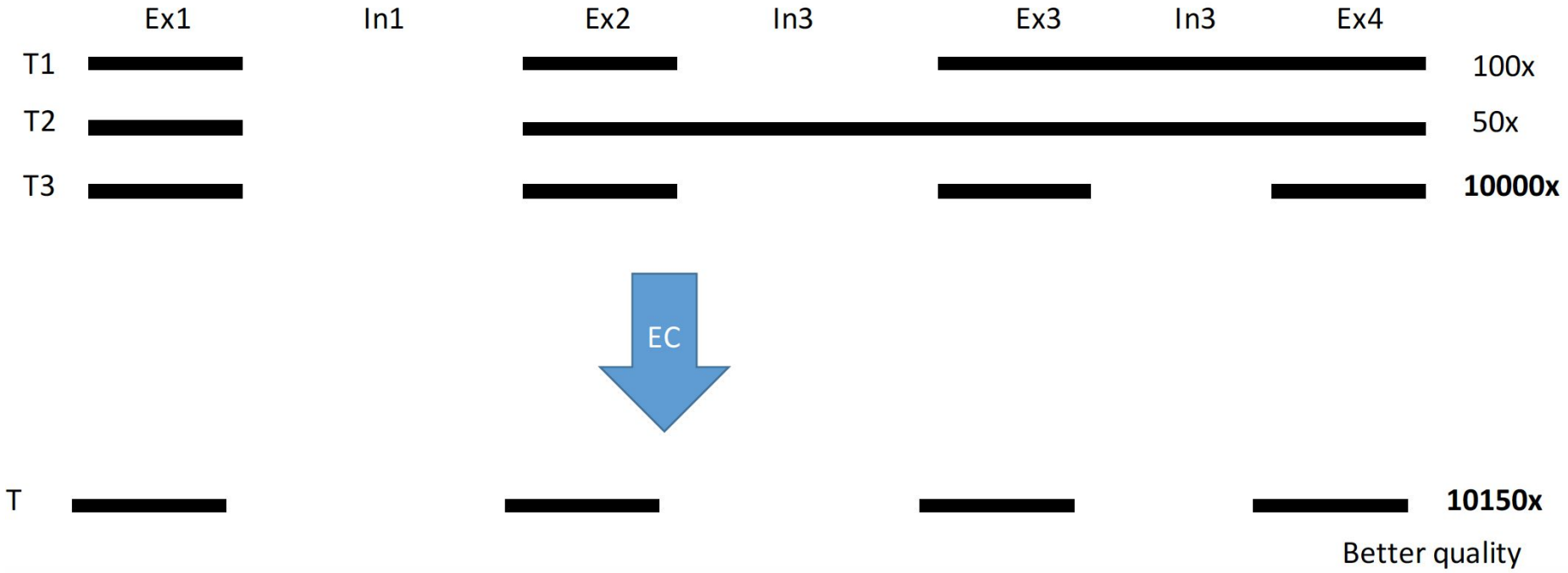
To trim or not to trim?

	Proovread	Proovread trim.	daccord	daccord trimmed
mean length	2117	1796	2102	1475
number of genes	16.7k	14.5k	15.7k	14k

Trimmed output of tools:

+ more reads and bases are mapped, less errors;
- reads are shorter, less genes are identified;

Is there a correction bias towards the major isoform?



Is there a correction bias towards the major isoform?

~~AlignQC~~

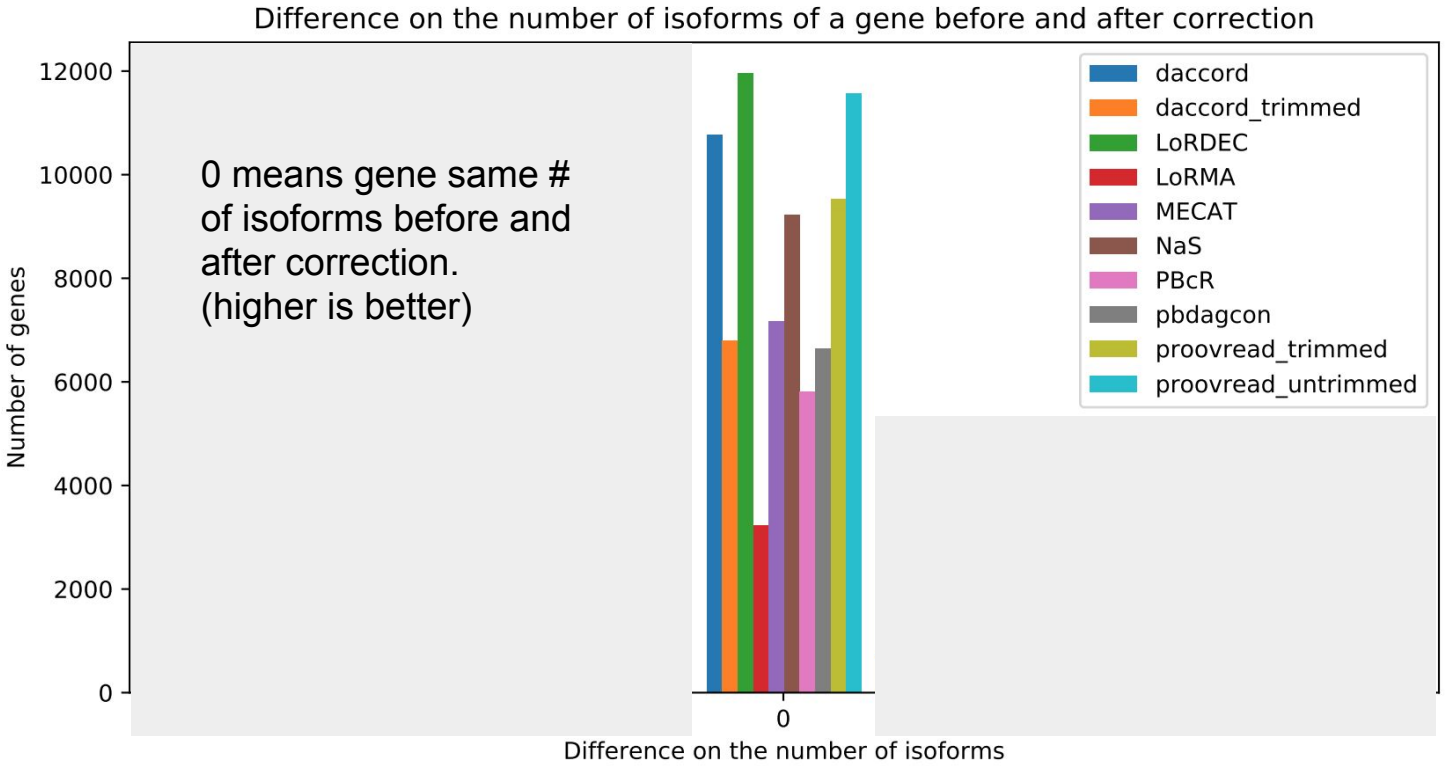
BWA-MEM on reference transcriptome

Filters: no secondary and >=80% QC

Genes before correction ∩ Genes after correction

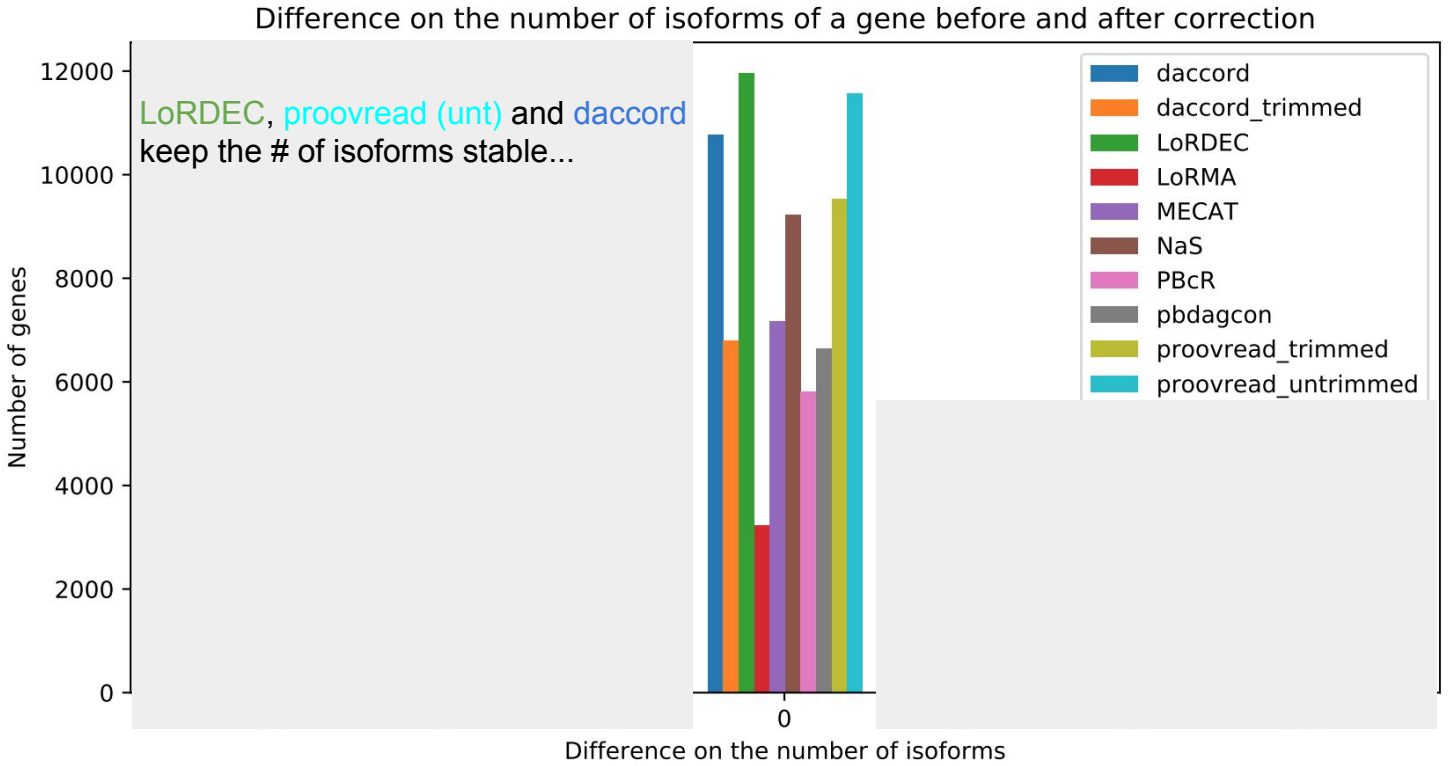
Is there a correction bias towards the major isoform?

Isoforms before and after correction



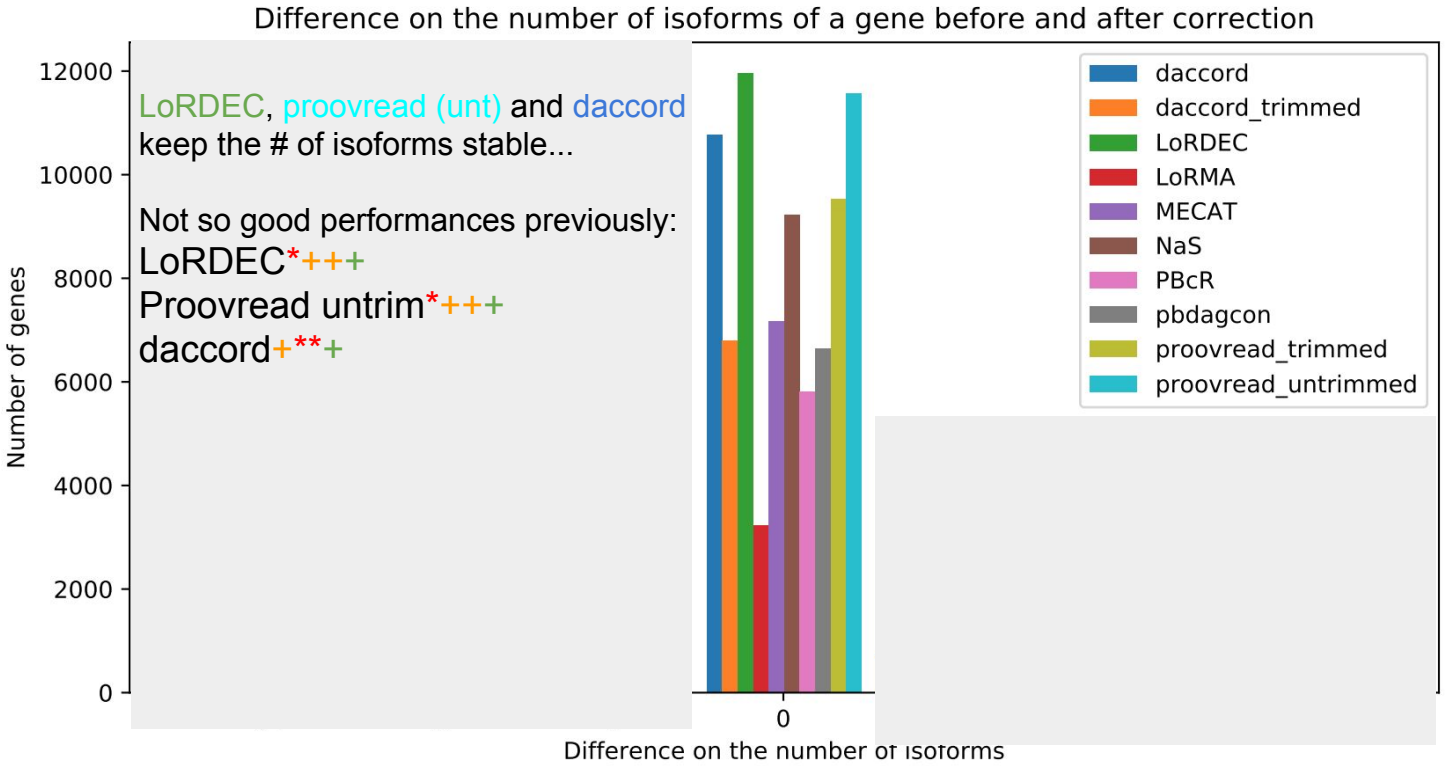
Is there a correction bias towards the major isoform?

Isoforms before and after correction



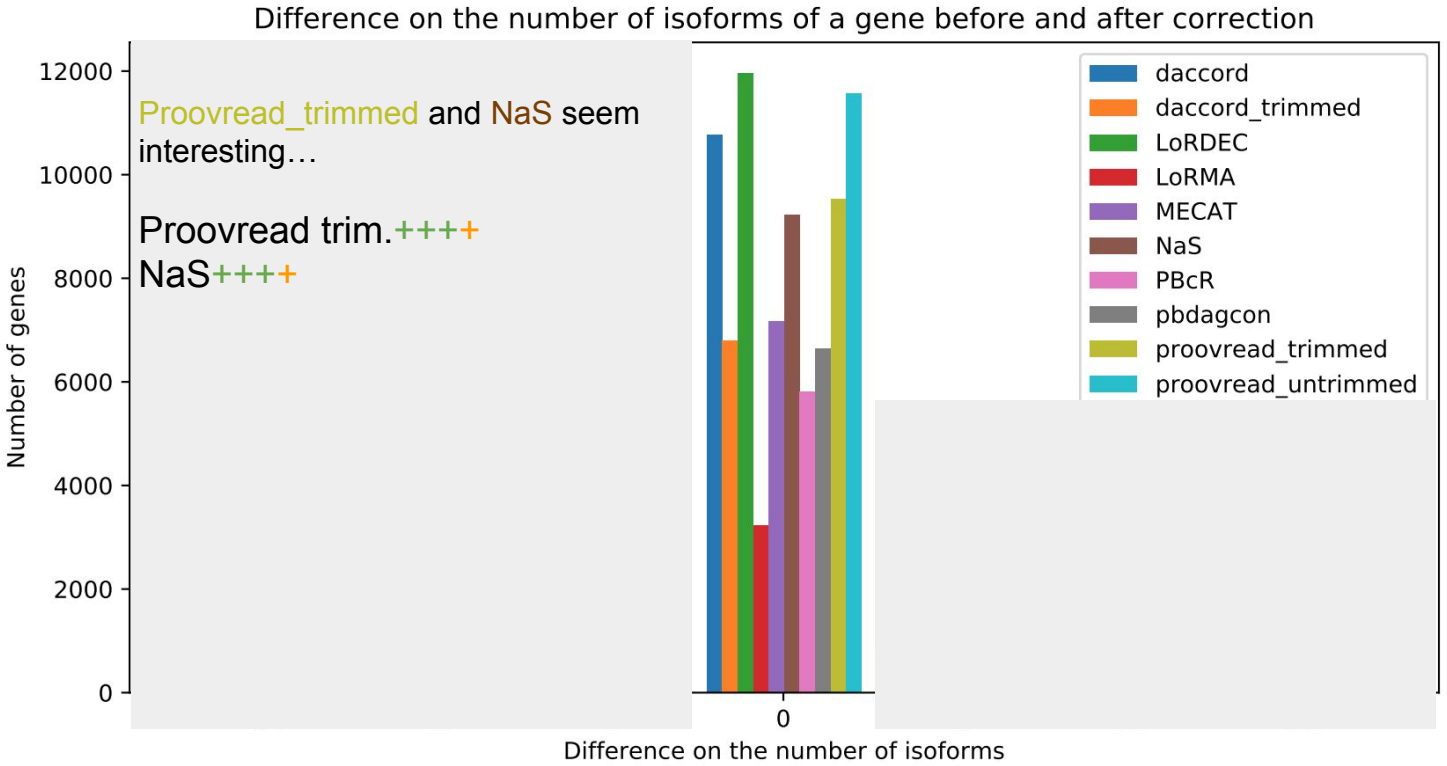
Is there a correction bias towards the major isoform?

Isoforms before and after correction



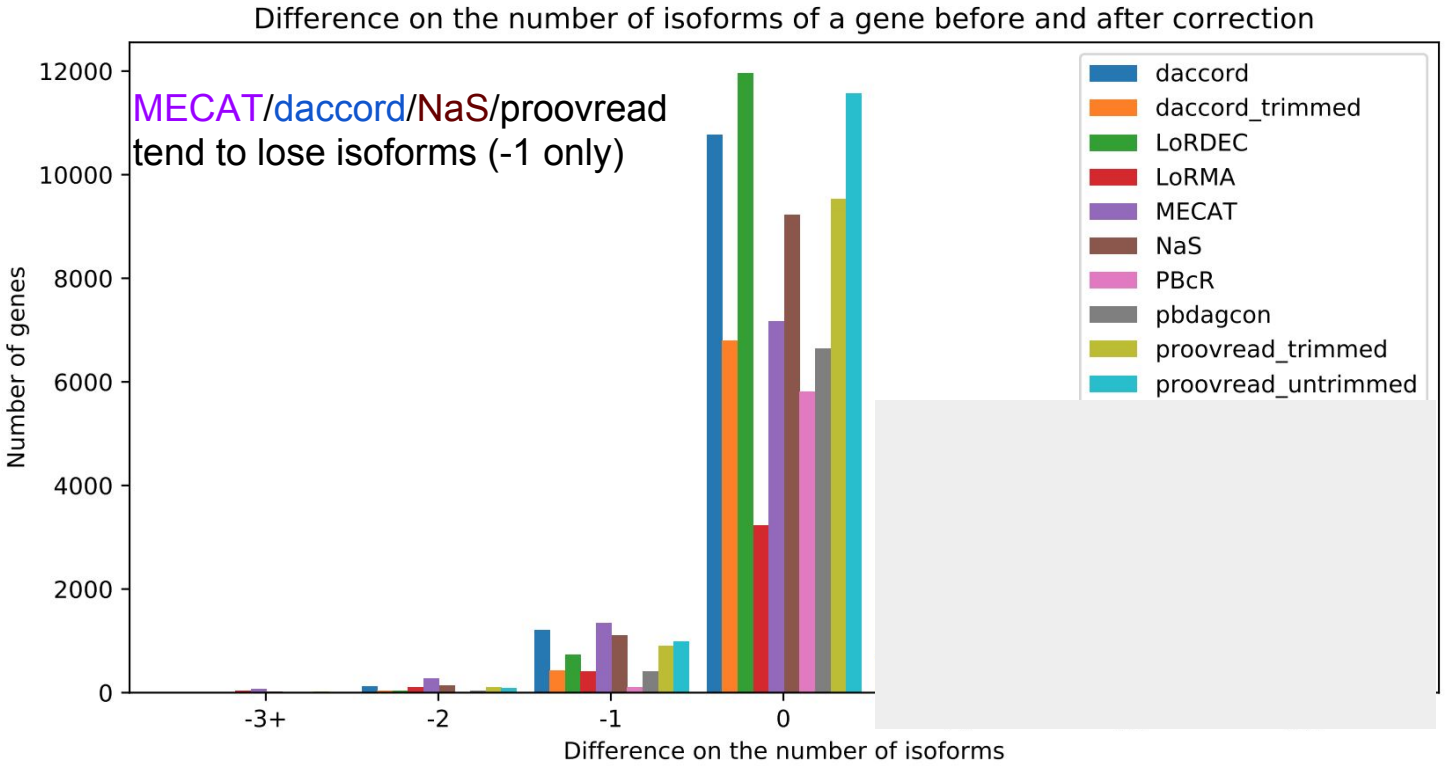
Is there a correction bias towards the major isoform?

Isoforms before and after correction



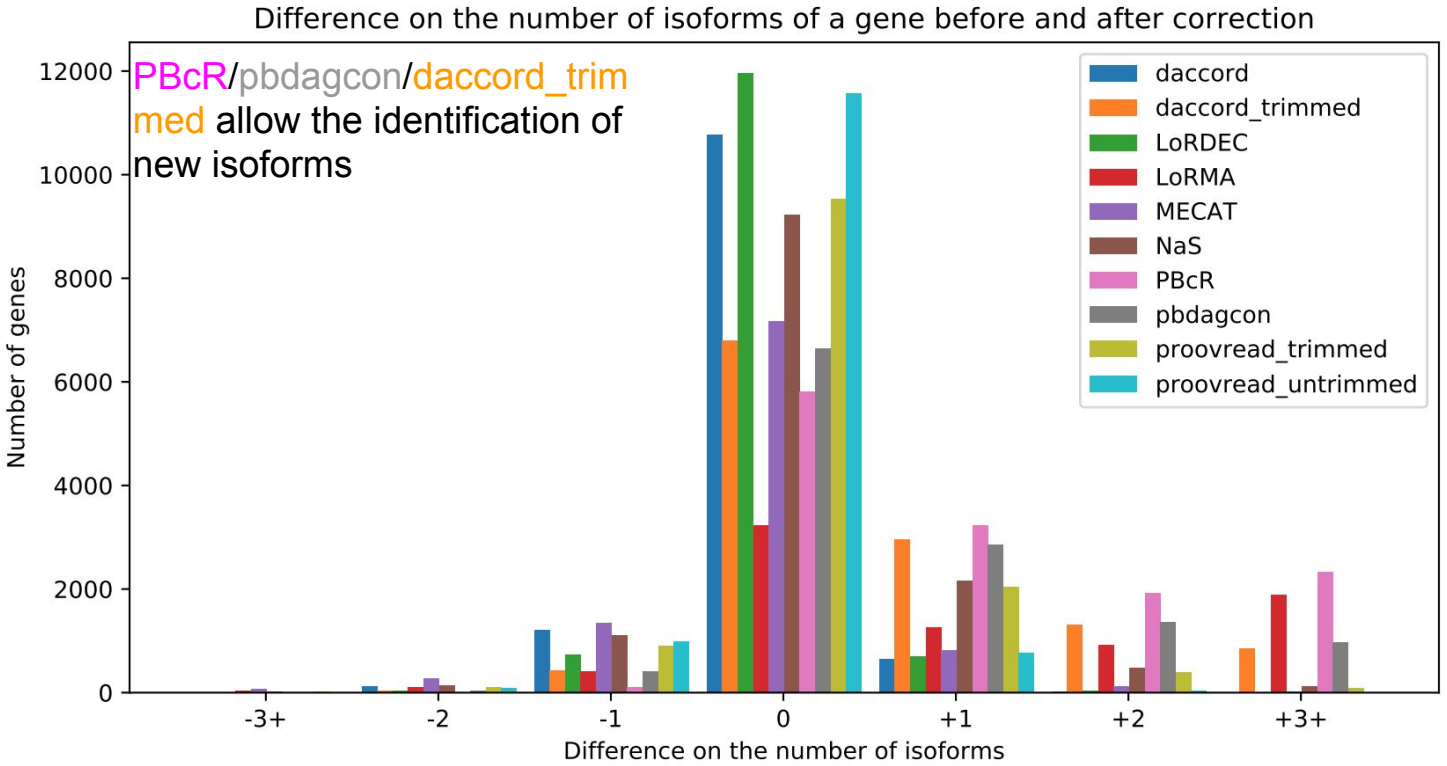
Is there a correction bias towards the major isoform?

Isoforms before and after correction



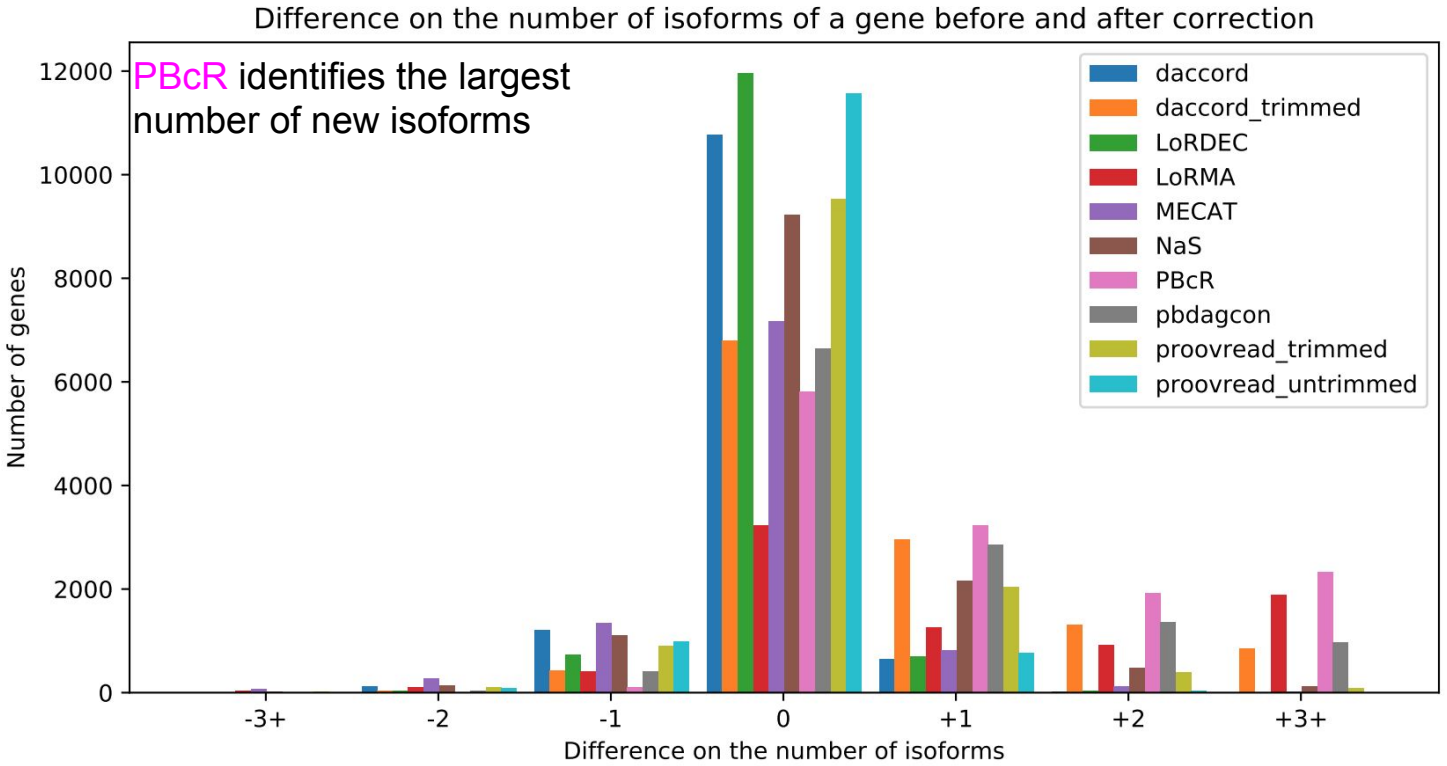
Is there a correction bias towards the major isoform?

Isoforms before and after correction



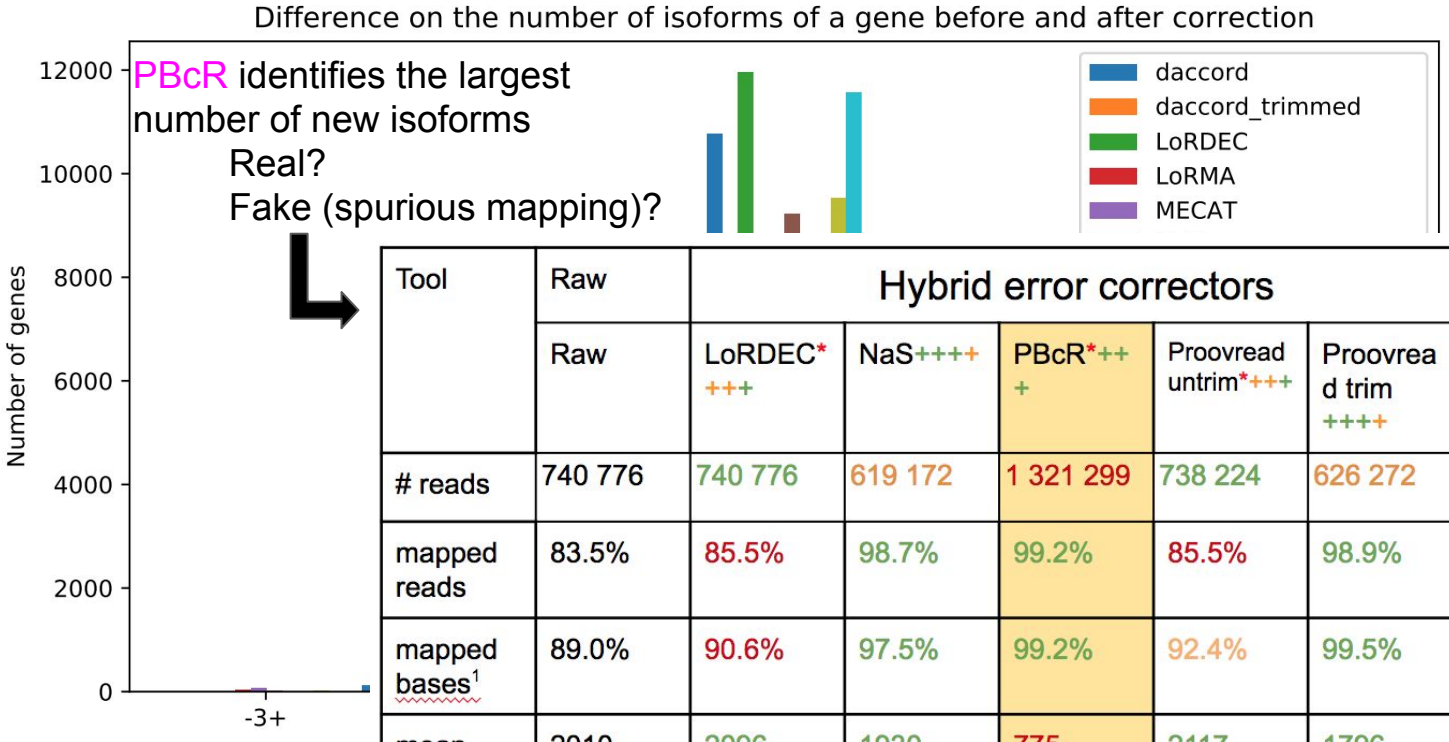
Is there a correction bias towards the major isoform?

Isoforms before and after correction



Is there a correction bias towards the major isoform?

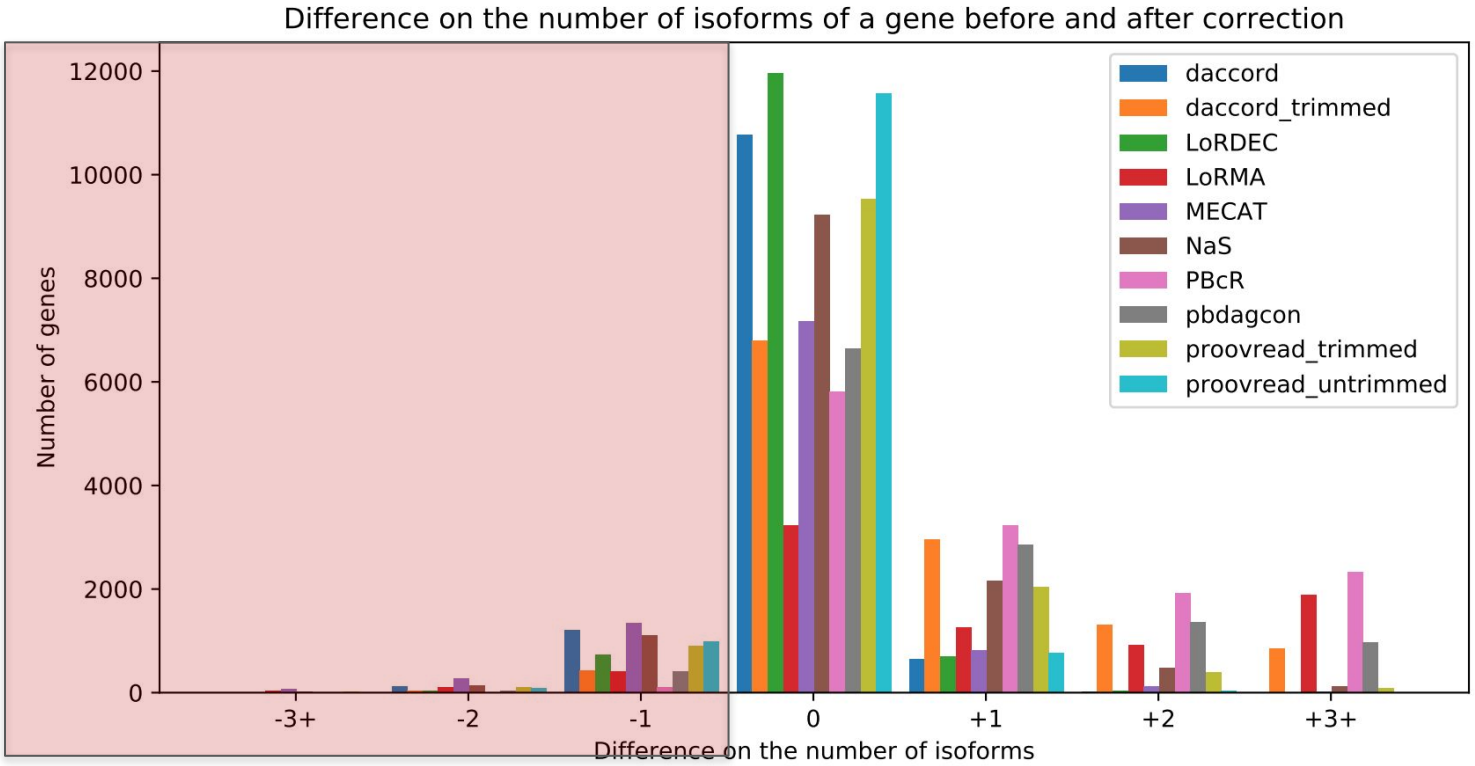
Isoforms before and after correction



Tool	Raw	Hybrid error correctors						S
	Raw	LoRDEC*+++	NaS++++	PBcR*+++	Proovread untrim*+++	Proovread trim++++	daccord+**+	dacco trim++
# reads	740 776	740 776	619 172	1 321 299	738 224	626 272	675 463	839 71
mapped reads	83.5%	85.5%	98.7%	99.2%	85.5%	98.9%	92.5%	94.0%
mapped bases[1]	89.0%	90.6%	97.5%	99.2%	92.4%	99.5%	92.5%	94.7%
mean length	2010	2096	1930	775	2117	1796	2102	1475

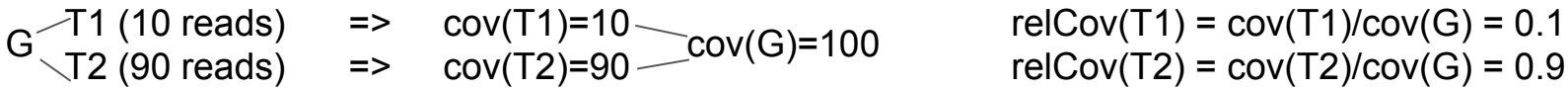
Is there a correction bias towards the major isoform?

Isoforms before and after correction



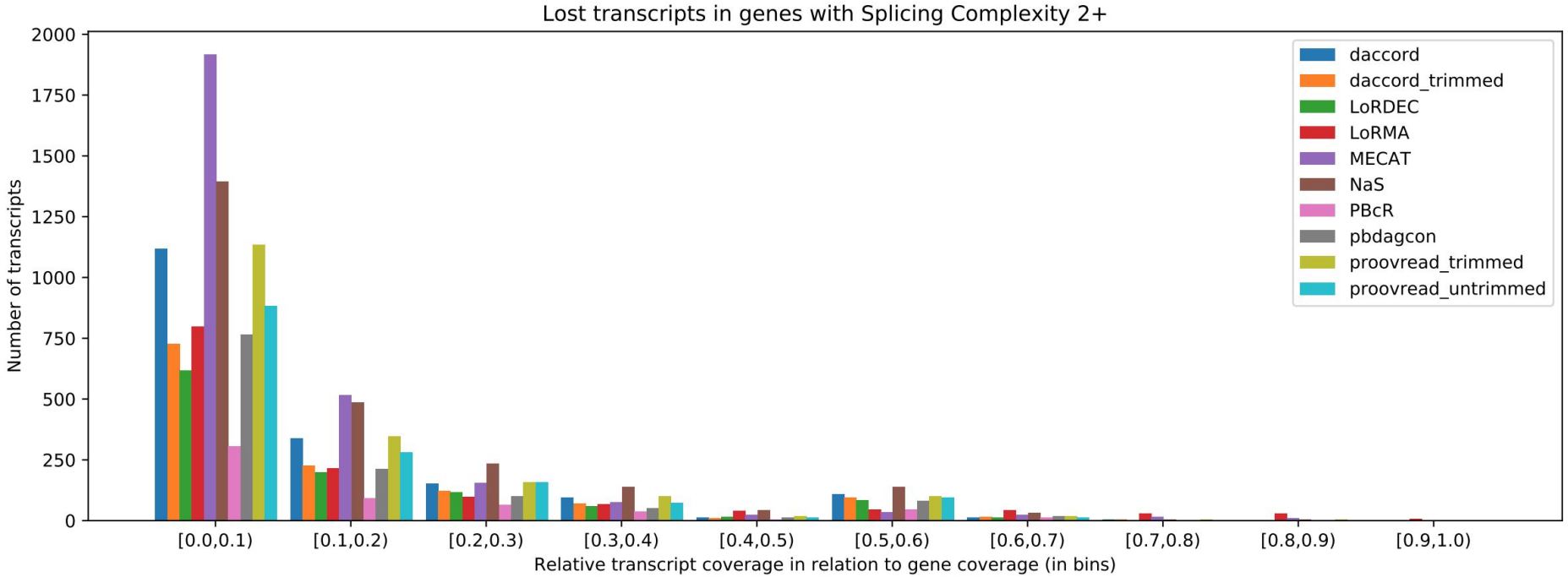
Is there a correction bias towards the major isoform?

Coverage of lost transcripts



Is there a correction bias towards the major isoform?

Coverage of lost transcripts

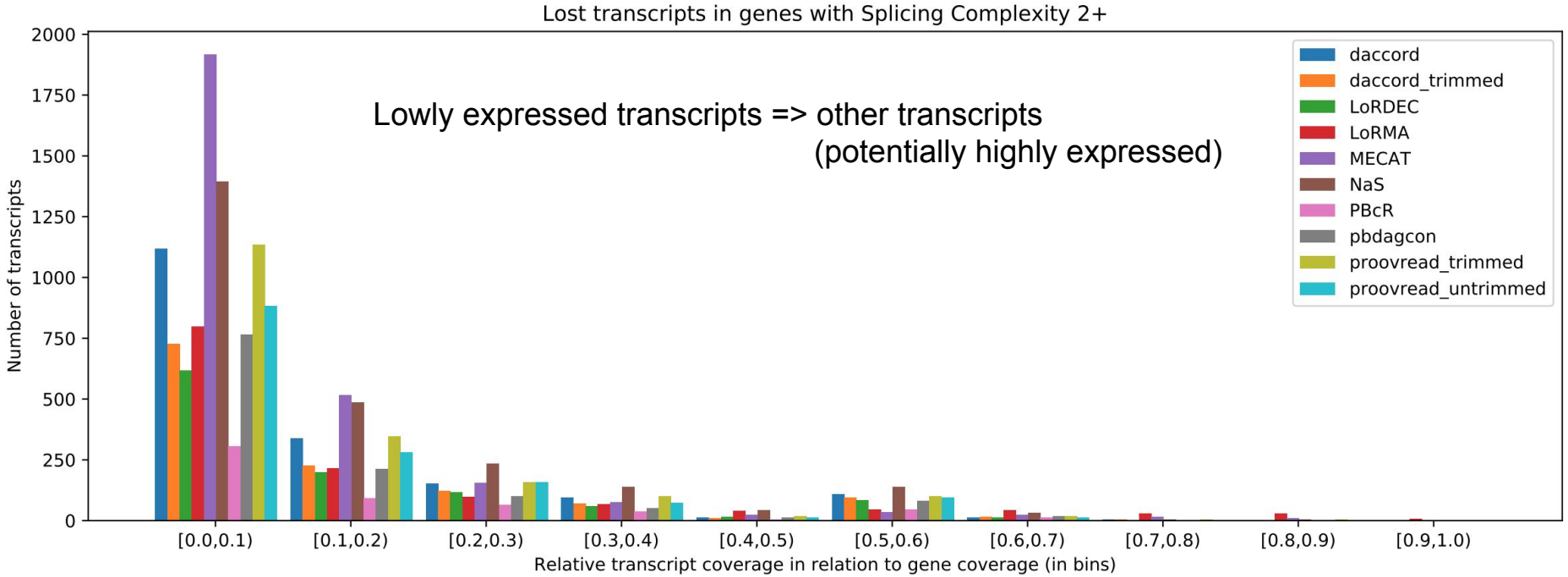


G — T1 (10 reads) => cov(T1)=10
G — T2 (90 reads) => cov(T2)=90
cov(T1)=10, cov(T2)=90 — cov(G)=100

relCov(T1) = cov(T1)/cov(G) = 0.1
relCov(T2) = cov(T2)/cov(G) = 0.9

Is there a correction bias towards the major isoform?

Coverage of lost transcripts

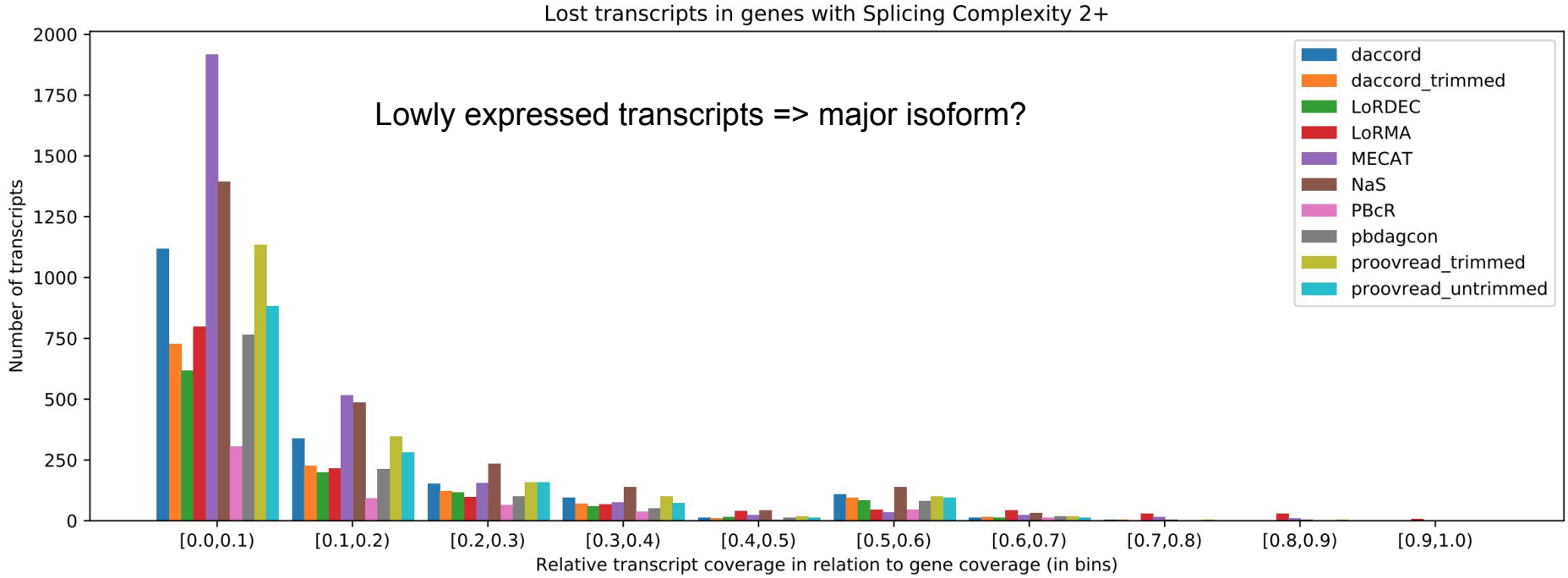


G — T1 (10 reads) => cov(T1)=10
G — T2 (90 reads) => cov(T2)=90
cov(T1)=10, cov(T2)=90 — cov(G)=100

relCov(T1) = cov(T1)/cov(G) = 0.1
relCov(T2) = cov(T2)/cov(G) = 0.9

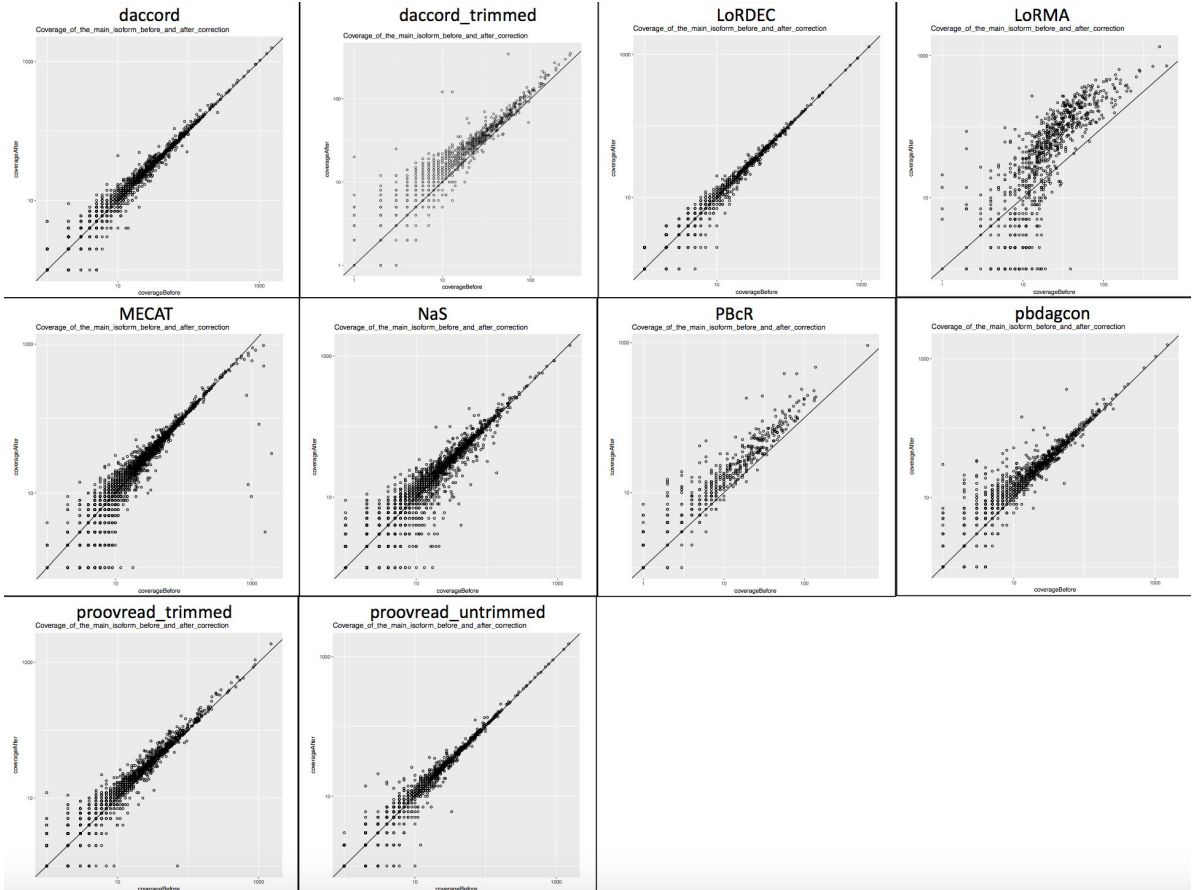
Is there a correction bias towards the major isoform?

Coverage of lost transcripts



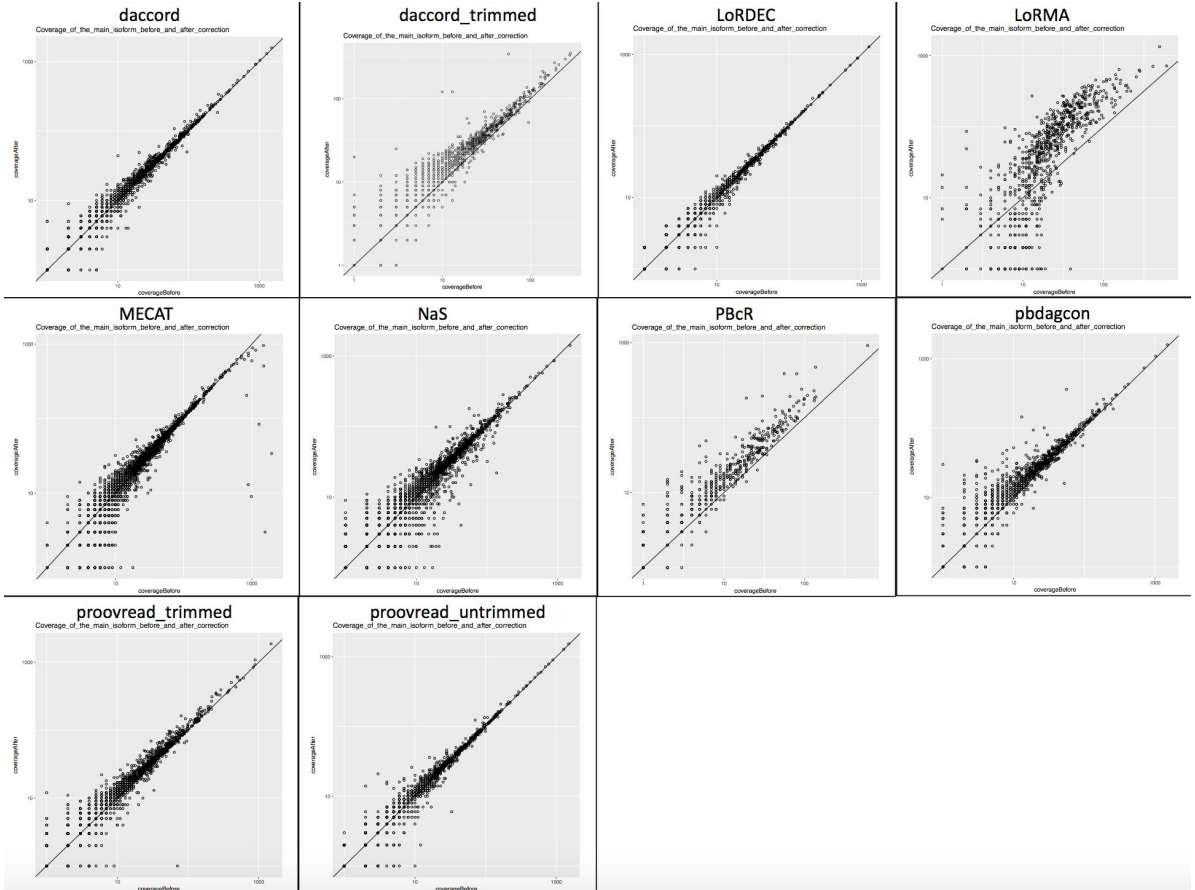
Is there a correction bias towards the major isoform?

Coverage of main isoform before (x) and after (y) correction



Is there a correction bias towards the major isoform?

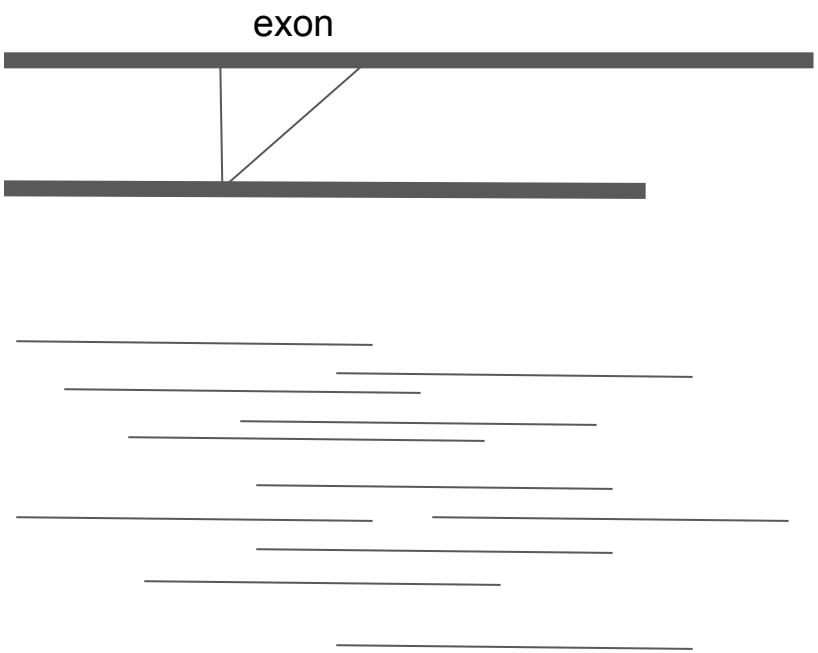
Coverage of main isoform before (x) and after (y) correction



LoRMA, PBcR, daccord_trimmed tend to overestimate main isoform expression:

- -Split reads?
- -Correction towards major isoform?

Simulation: when are reads corrected to major isoform?



2 transcripts

different abundances

Skipped exon

different sizes

Simulated reads

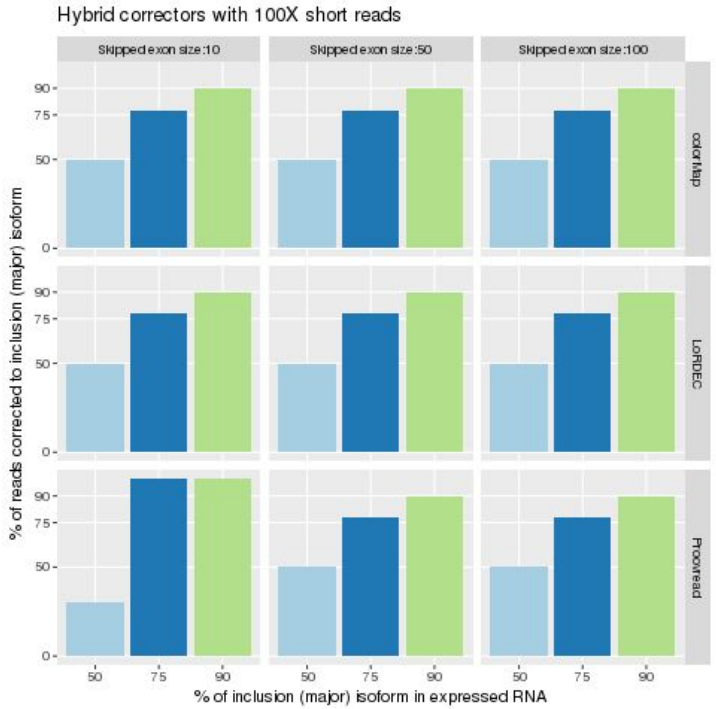
Simulation: when are reads corrected to major isoform?

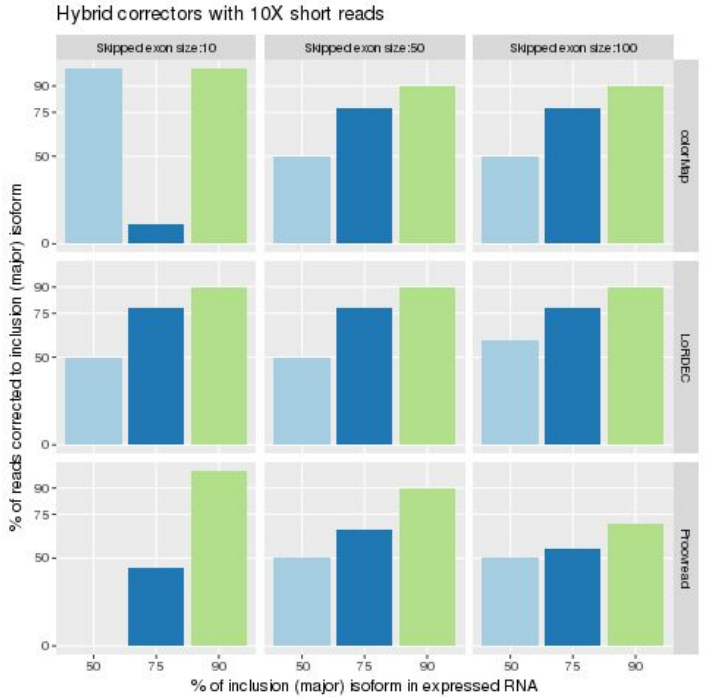
Ideal correction: Light blue should be 50%, dark blue should be 75%, green should be 90%

Colormap

LoRDEC

Proovread



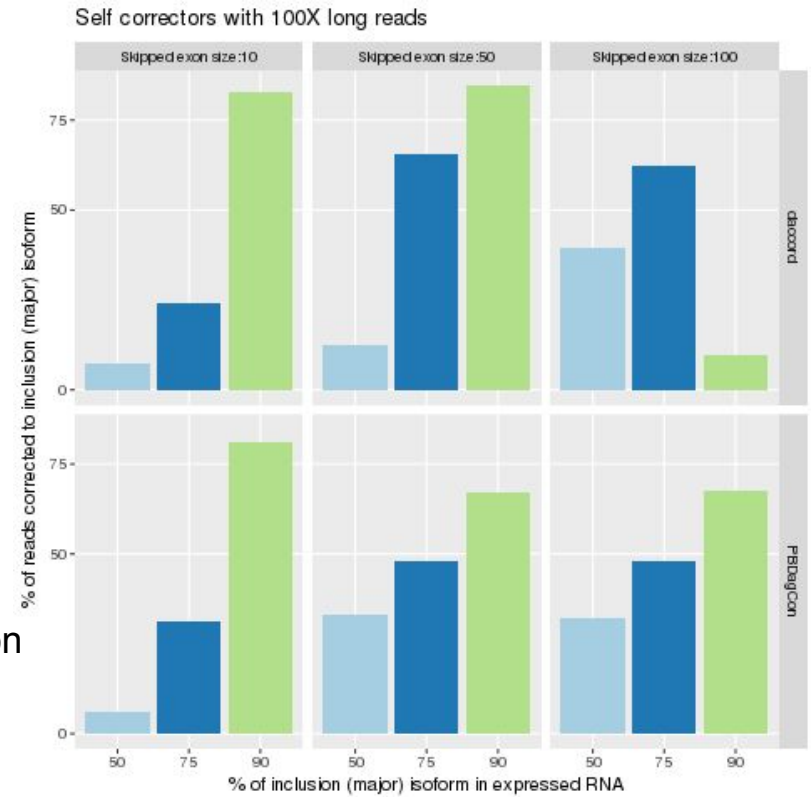


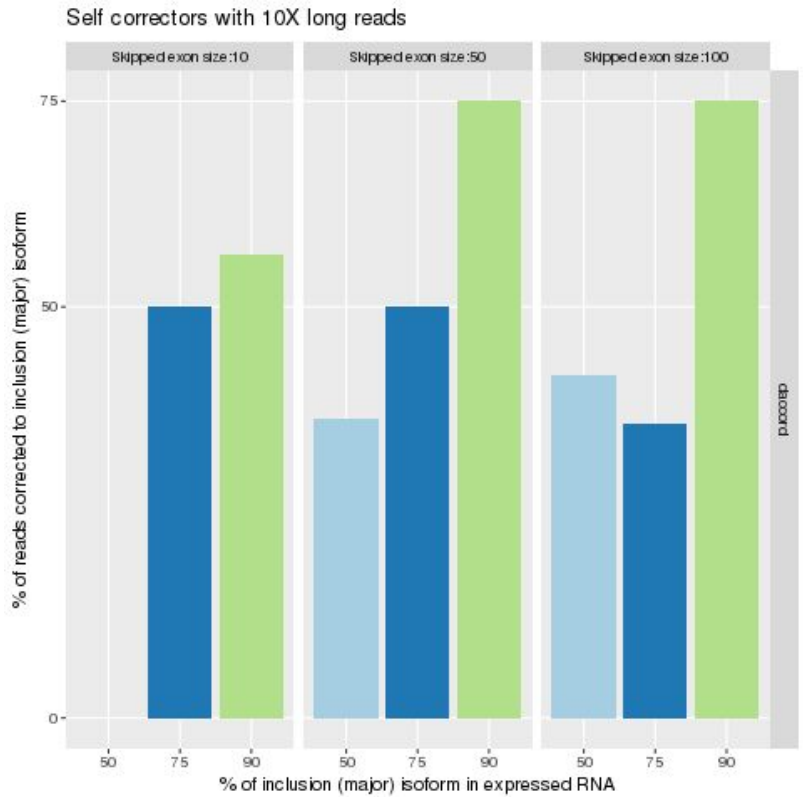
Bottom line: LoRDEC generally doesn't overcorrect, proovread and colormap do

Simulation: when are reads corrected to major isoform?

daccord

PBDagcon





daccord

Conclusion (1/3)

Performance:

LoRDEC, daccord, LoRMA, MECAT, pbdagcon

Error rate:

PBcR, NaS, proovread.
Rest: 2-5% remaining error rate

Conclusion (2/3)

Same number of detected genes:

LoRDEC, daccord, PBcR, proovread, (NaS)

Isoform preservation:

LoRDEC, proovread *(tricky to decide; based on lost transcripts, & number of isoforms)*

Conclusion (3/3)

Overall recommendations:

Proovread, PBcR, NaS

If you have to choose a non-hybrid:

daccord/pbdagcon, because they do not lose coverage like LoRMA/MECAT

Conclusion (4/3)

Potential pitfalls:

- Single data type (1D)
- potential aligner bias
- did not track isoforms before/after correction
- couldn’t run Canu (disk hungry)