

An analysis of splicing variation across SRA with Rail-RNA

@AbhiNellore

Johns Hopkins University
Genome Informatics 2015

in genomics

use prior
knowledge



study data
ab initio

in RNA-seq analysis

**use gene
annotation:**

quantify with/
align to known
transcripts



**avoid gene
annotation:**

observe
alternative
splicings

in RNA-seq analysis

**use gene
annotation:**
quantify with/
align to known
transcripts



**avoid gene
annotation:**
alternative
splicings?

study merits/drawbacks
with many RNA-seq samples

Study many RNA-seq samples

SRA (> 4 PB): short reads hard to assemble; missing exons in **60%** of transcripts

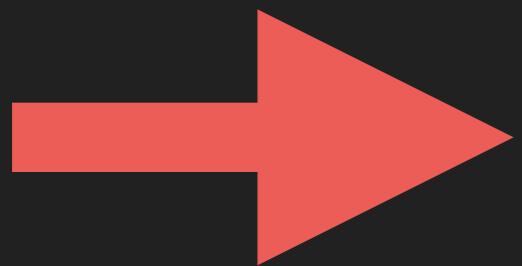
(RGASP 2013 doi:10.1038/nmeth.2714)

=> Compare exon-exon junctions found across SRA RNA-seq with annotated junctions

Filtering SRA

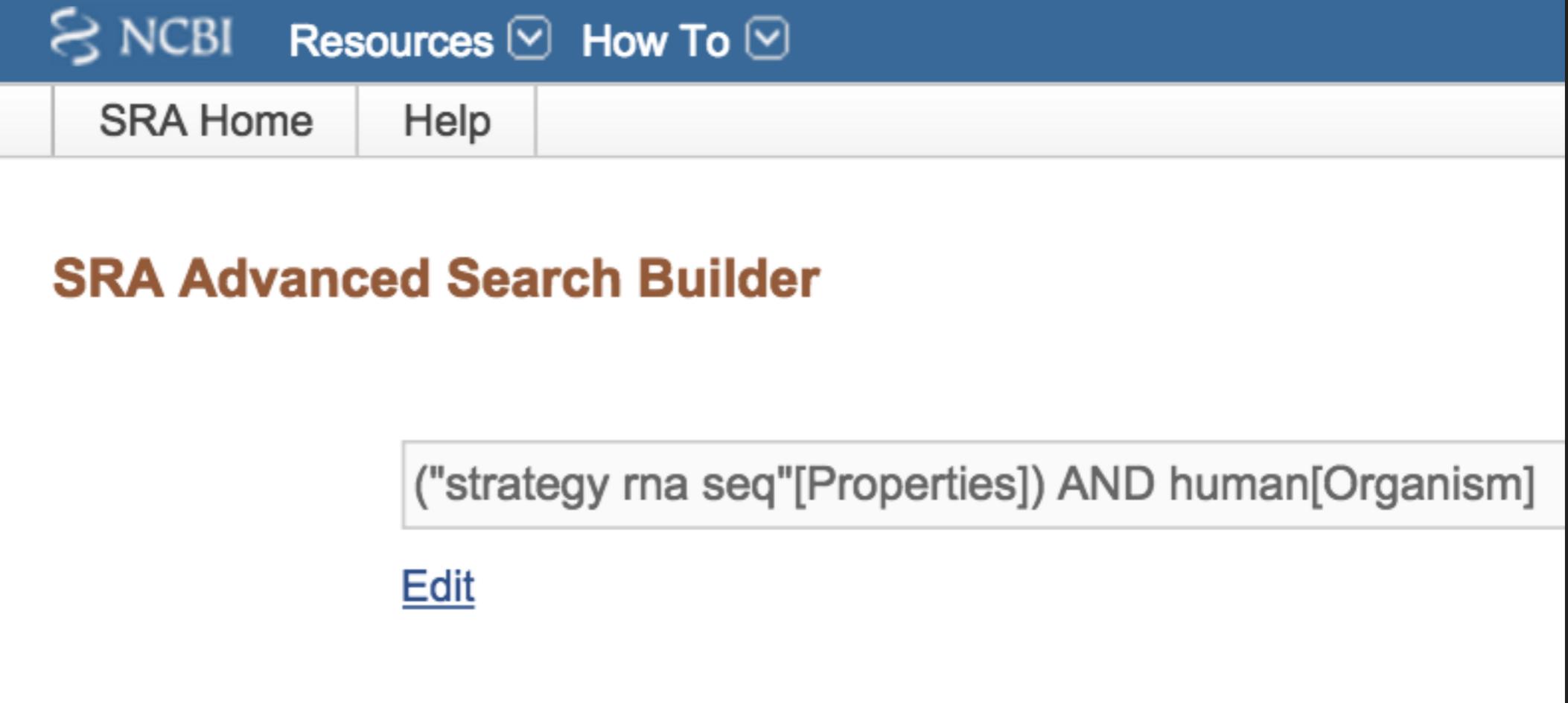
The screenshot shows the NCBI SRA Advanced Search Builder. At the top, there is a blue header bar with the NCBI logo, the word "Resources" with a dropdown arrow, and "How To" with a dropdown arrow. Below the header is a navigation bar with "SRA Home" and "Help". The main content area has a brown header "SRA Advanced Search Builder". Below it is a search input field containing the query "**"strategy rna seq"**[Properties]". Underneath the input field is a blue "Edit" link. The entire interface is set against a white background.

(from <http://www.ncbi.nlm.nih.gov/sra/advanced>)



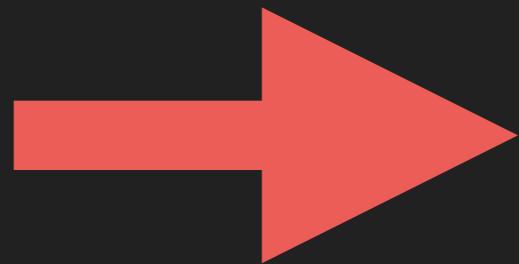
≈ 180k publicly available runs

Filtering SRA



The screenshot shows the NCBI SRA Advanced Search Builder. At the top, there is a blue header bar with the NCBI logo, a "Resources" dropdown, and a "How To" dropdown. Below the header is a navigation bar with "SRA Home" and "Help" buttons. The main content area has a title "SRA Advanced Search Builder" in brown text. Below the title is a search bar containing the query: ("strategy rna seq"[Properties]) AND human[Organism]. Underneath the search bar is an "Edit" link. The entire interface is set against a white background.

(from <http://www.ncbi.nlm.nih.gov/sra/advanced>)



≈36k publicly available runs

Filtering SRA

The screenshot shows the NCBI SRA Advanced Search Builder. At the top, there is a blue header bar with the NCBI logo, a "Resources" dropdown, and a "How To" dropdown. Below the header is a navigation bar with "SRA Home" and "Help" buttons. The main content area has a title "SRA Advanced Search Builder" and a search bar containing the query: ("strategy rna seq"[Properties]) AND human[Organism]. Below the search bar is an "Edit" link.

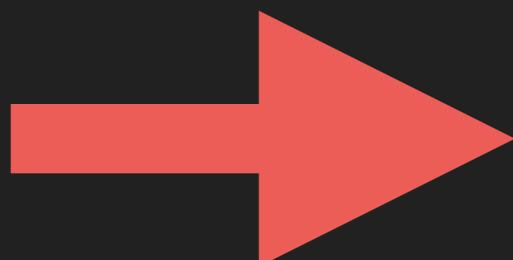
SRA Advanced Search Builder

("strategy rna seq"[Properties]) AND human[Organism]

[Edit](#)

(from <http://www.ncbi.nlm.nih.gov/sra/advanced>)

+ Illumina instruments[Properties]



≈22k runs as of late May '15

How to find junctions across
21,504 RNA-seq runs?

(62 terabases of reads)

annotation-agnostic pipeline

 Rail-RNA

<http://rail.bio>



derfinder

`biocLite("derfinder")`



Leo
Collado-Torres



Alyssa
Frazee

**sidesteps
assembly &
annotation limitations
resolves
isoform-level
features**

derfinder finds unannotated (D)ERs

8.3% of age-associated DERs
outside annotated genes across
72 prefrontal cortex samples:

Jaffe et al. (Nat Neuro, doi:10.1038/nn.3898)

6.9% of ERs outside annotated
genes across 465 GEUVADIS LCLs:
Nellore et al. (j.mp/rail-pre)



Rail-RNA



- No competition for compute
- Rapid: 8 days to results
- Repeatable:

<http://github.com/nellore/gi2015>

for commands and data

- Inexpensive: ~\$0.70/sample

2 commands X 43 batches
gave, across 21,504 samples

One 7-GB tsv.gz

42,882,032

junctions

What gene annotation says

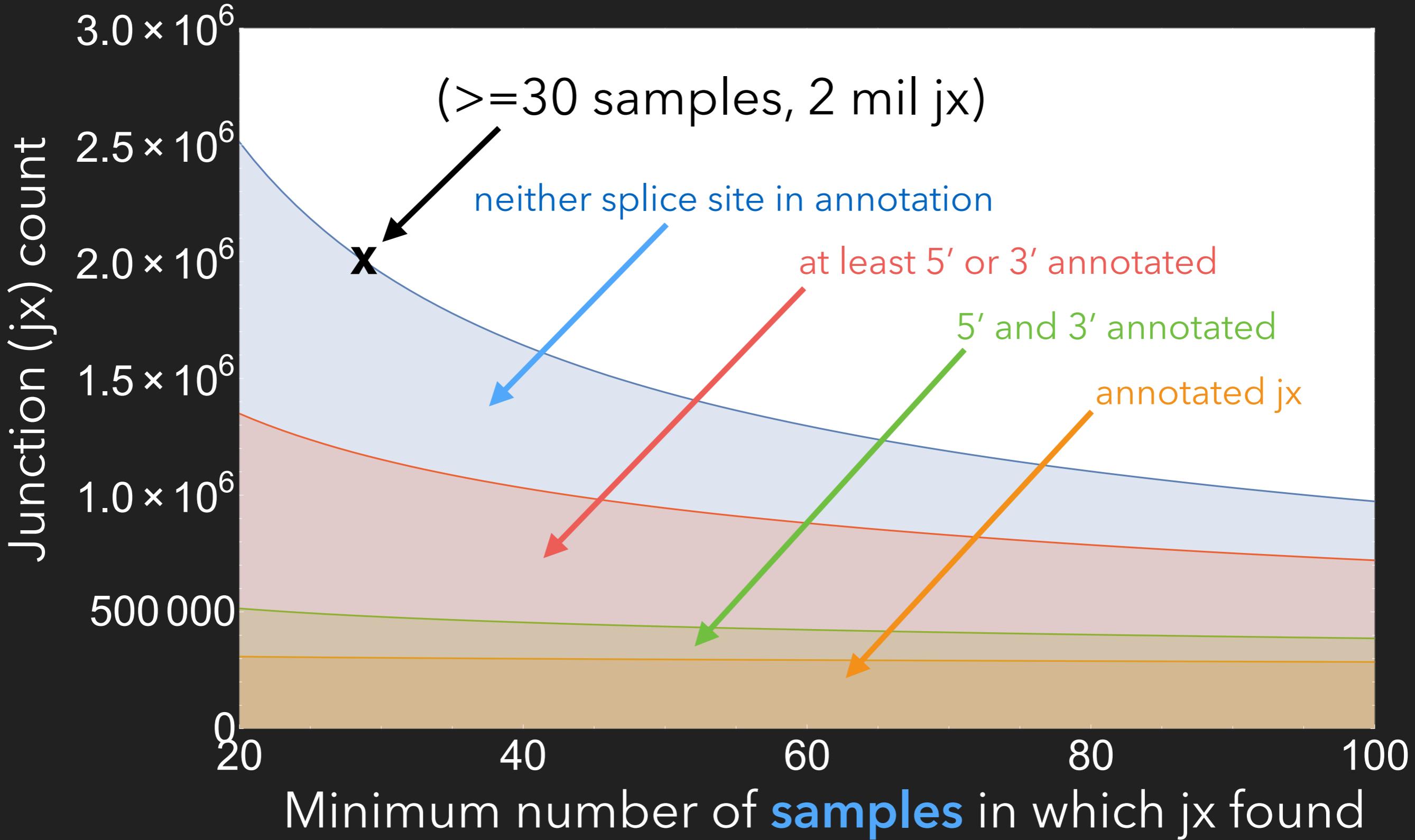
For *hg19*,

Ensembl v75 \cup GENCODE v19 \cup RefSeq
(almost subsumed
by Ensembl v75)

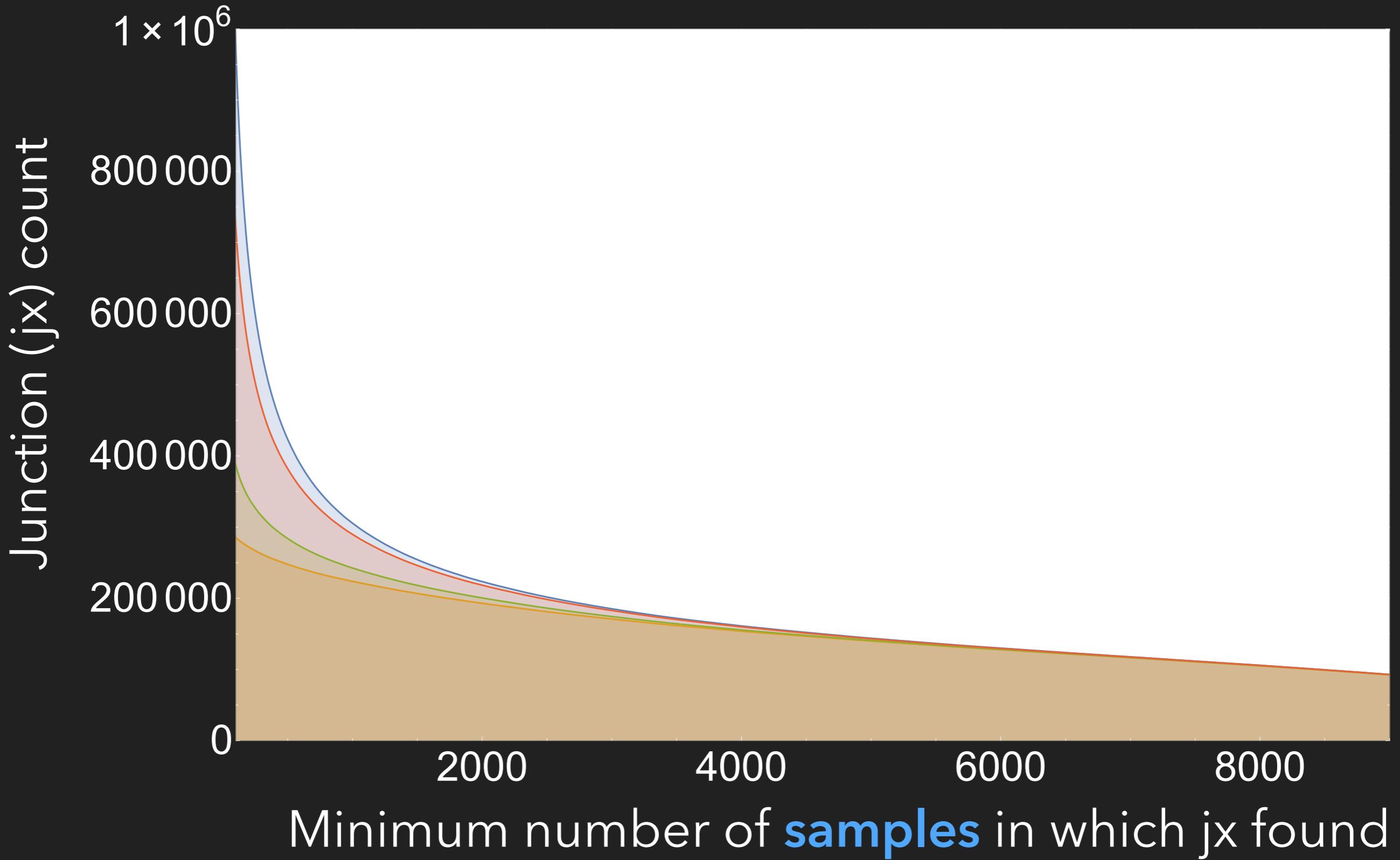
$\approx 350,000$

junctions

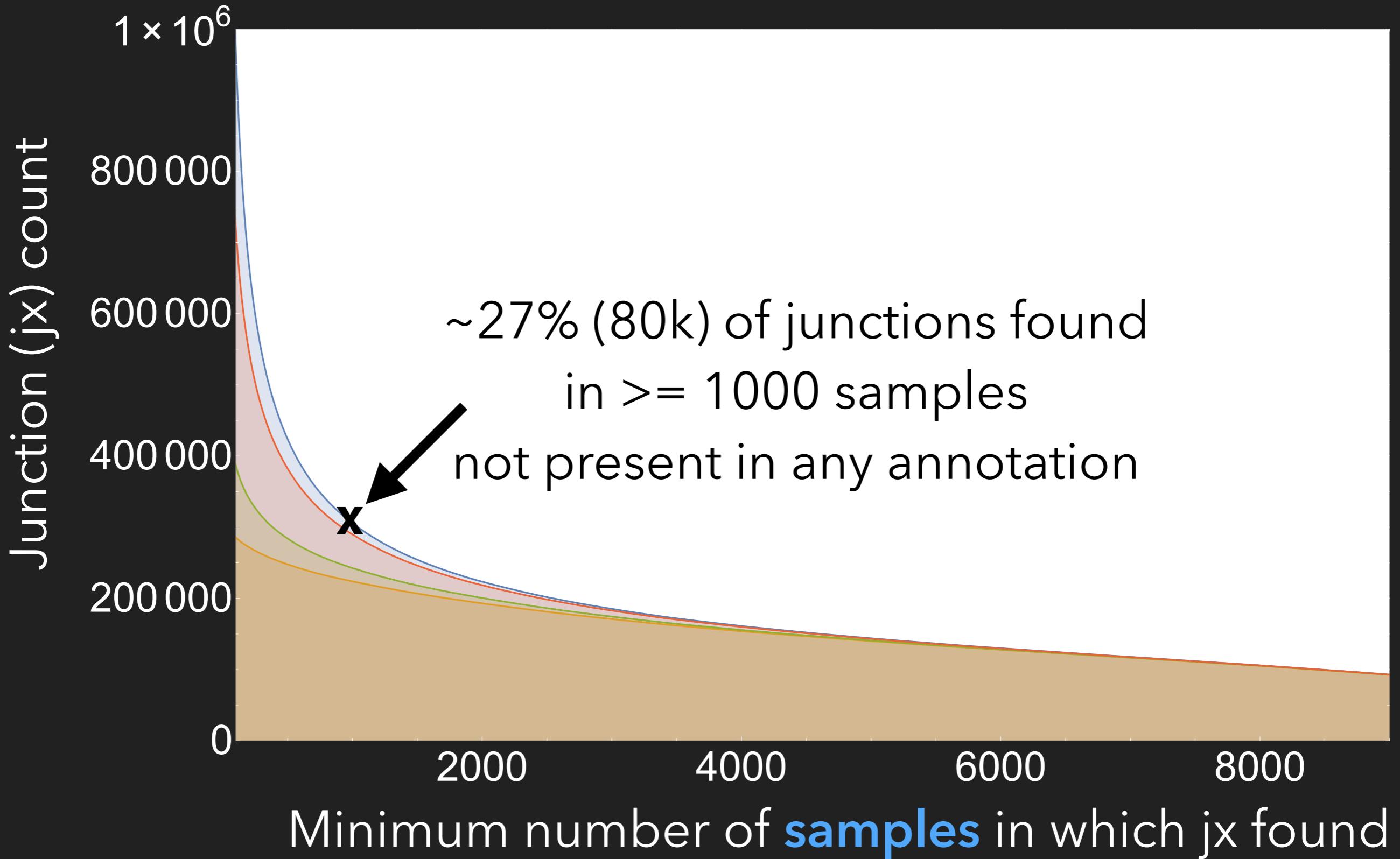
Increasing evidence in annotation



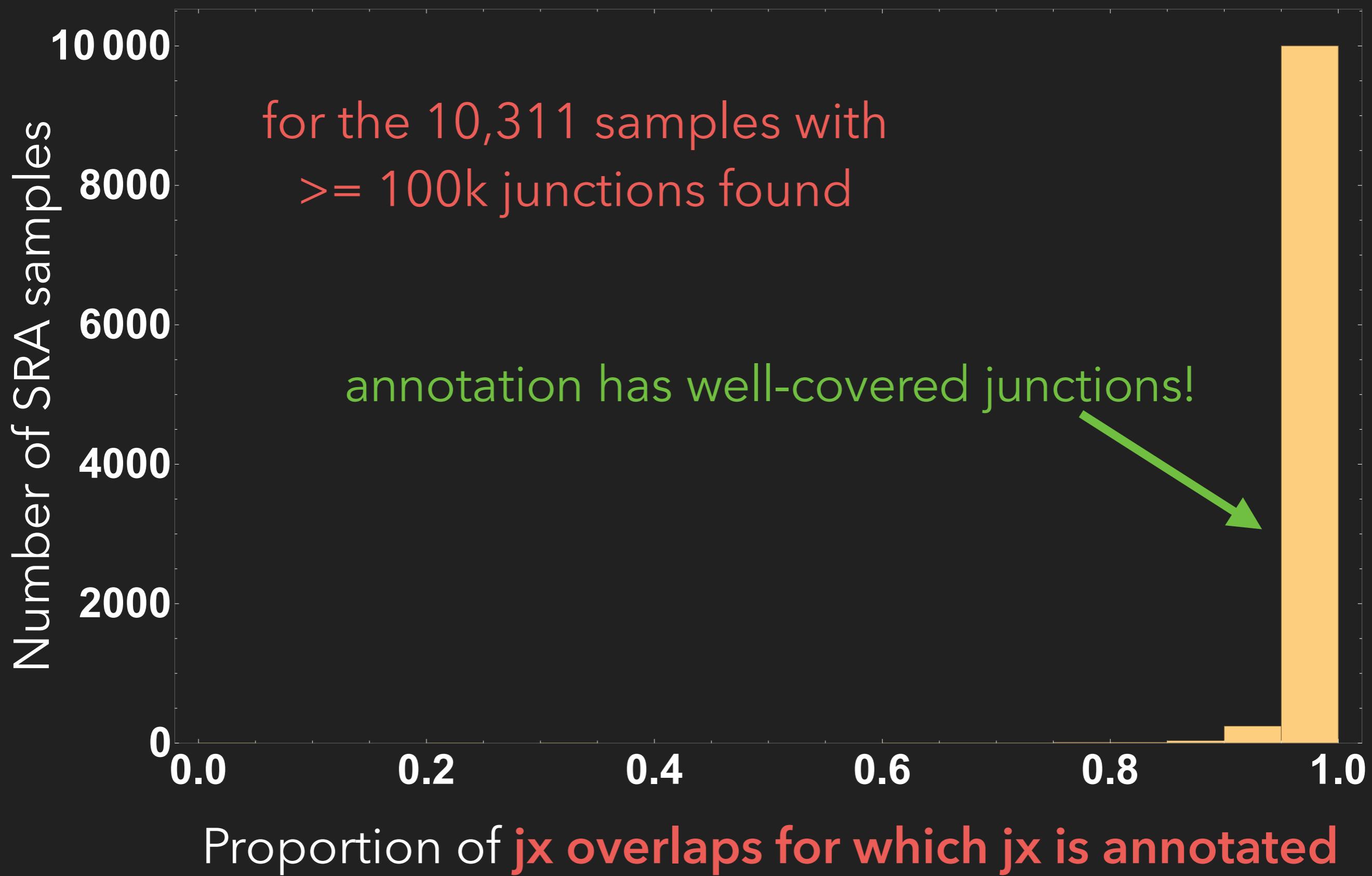
Asymptote to annotation



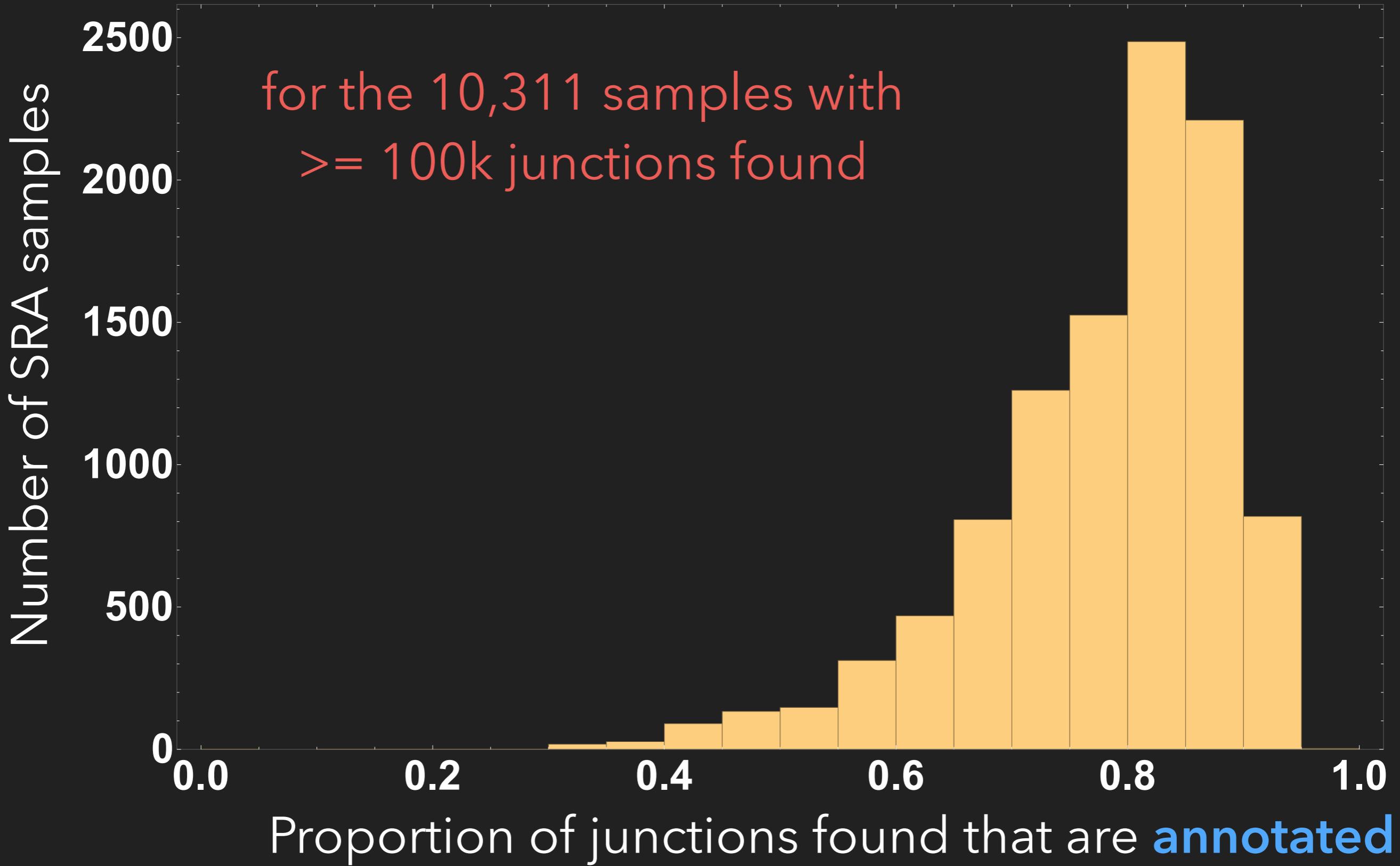
Asymptote to annotation



Junction (jx) overlaps by sample



Annotated junctions by sample

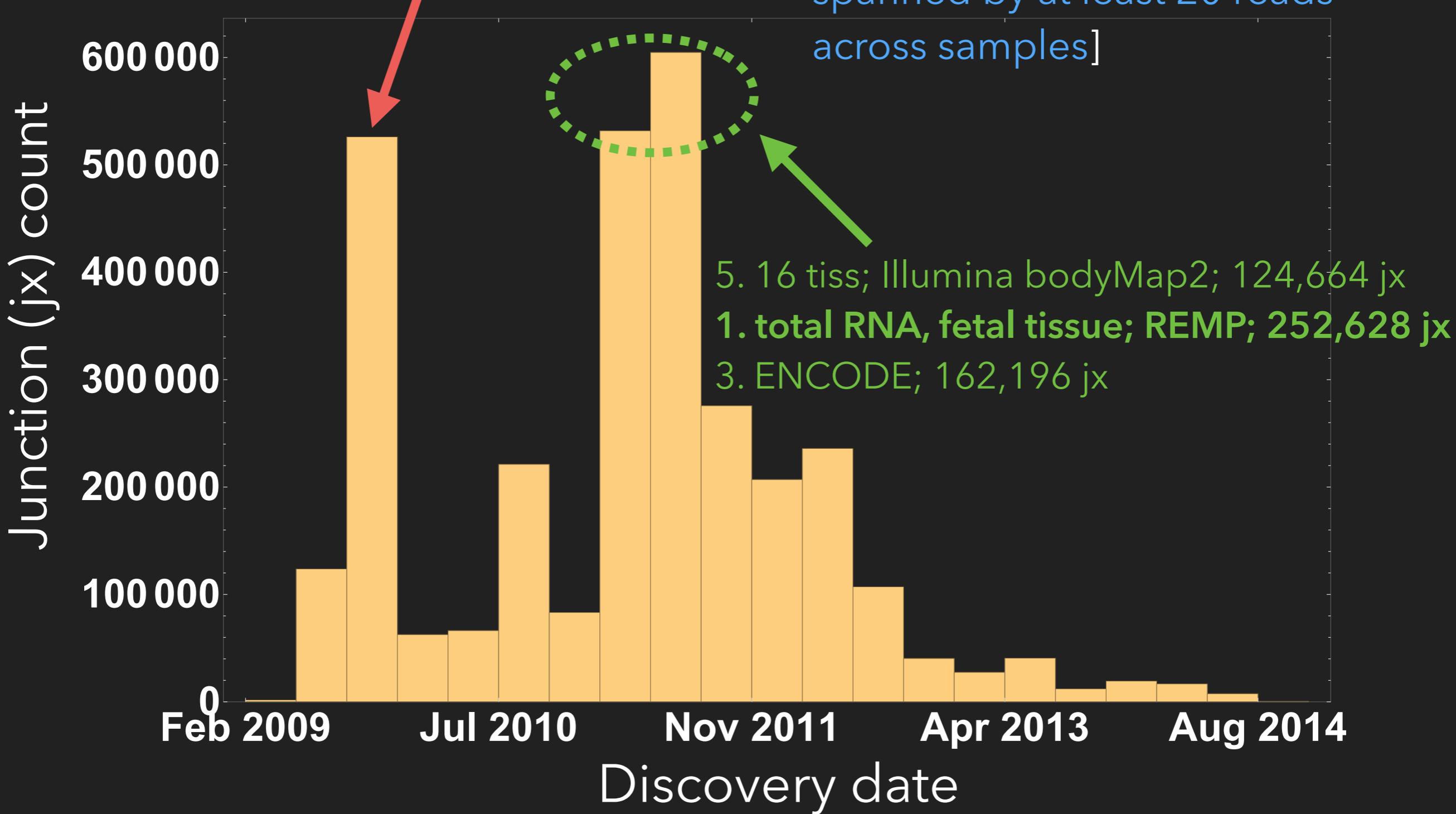


Are we still finding new human jx?

4. 69 LCLs; Pickrell et al.; 155,069 jx

2. 41 Coriell CLs; Cheung et al.; 163,007 jx

[Considers only the 3,211,228 jxns spanned by at least 20 reads across samples]



Mixed news

Junctions in annotation are widespread and well-covered

Junctions in annotation don't capture splicing diversity for many SRA samples

annotation-agnostic pipeline



<http://rail.bio>



bigWigs: order of magnitude
smaller than BAMs

derfinder

`biocLite("derfinder")`

scripts for recovering junctions

&

processed data

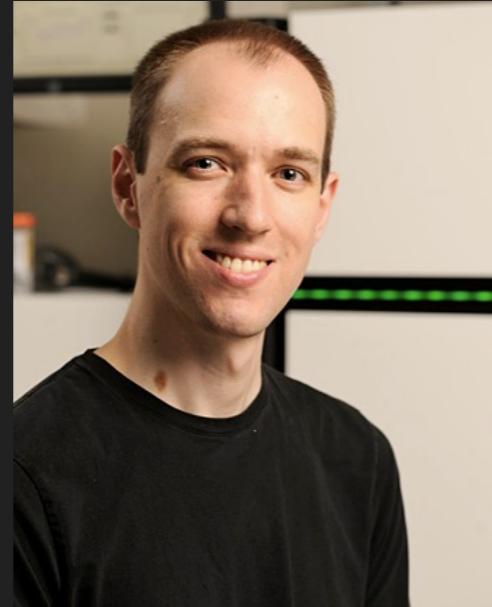
@

<http://github.com/nellore/gi2015>

Collaborators



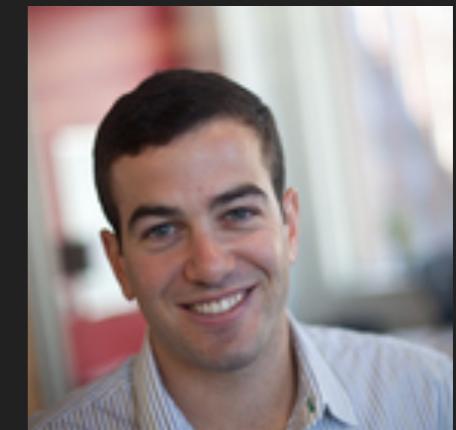
Jeff Leek



Ben Langmead



Leo



Andrew
Jaffe

poster 114; Thu



Jacob Pritt
poster 168; Fri



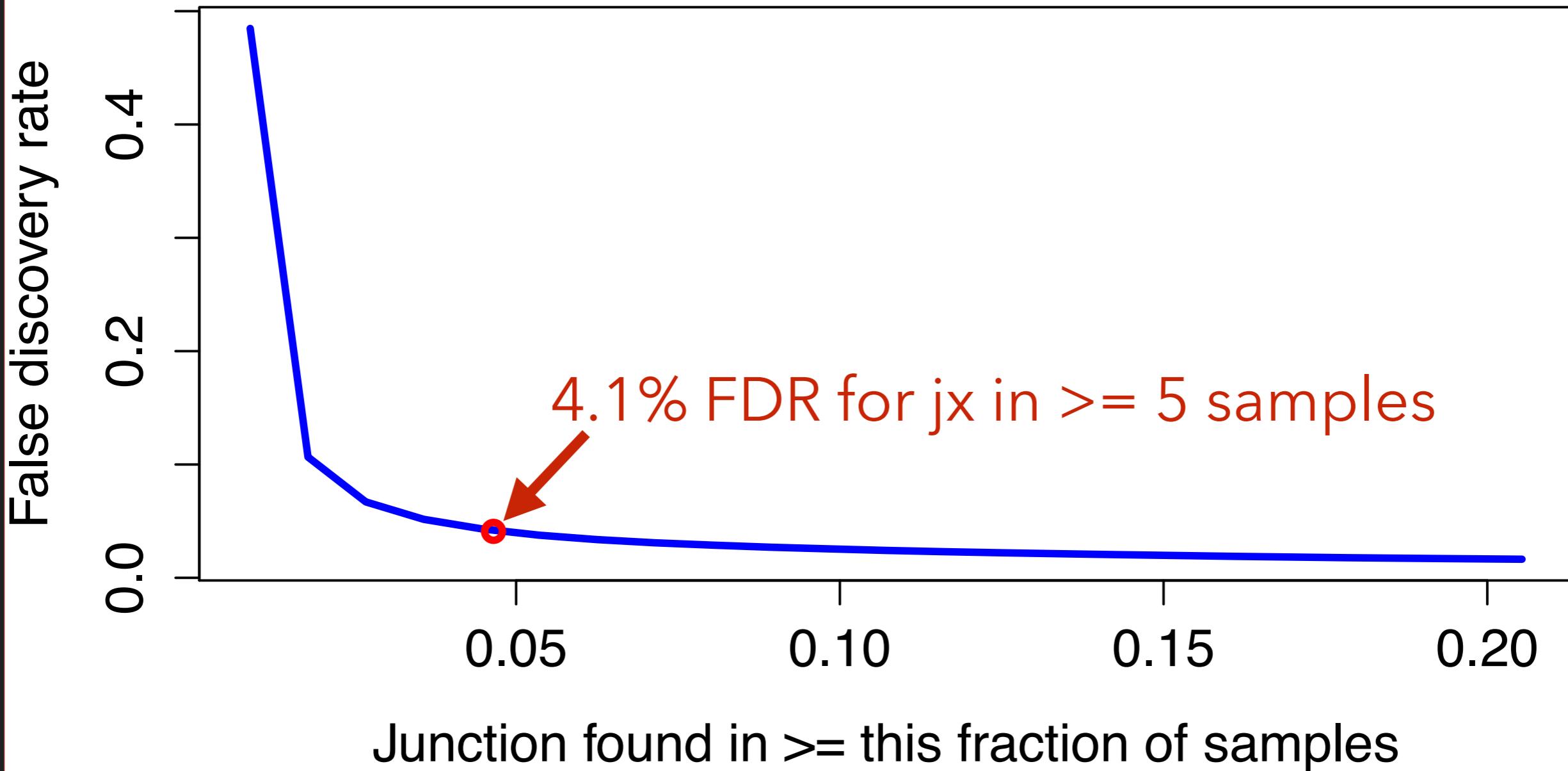
Chris Wilks



José
Alquicira Hernández

Summer interns: Nishika Karbhari, James Morton, Robert Phillips, Sara Wang

First-pass junction–call FDR estimated from 112 GEUVADIS–like simulations



Why so many junctions?

junctions



duds

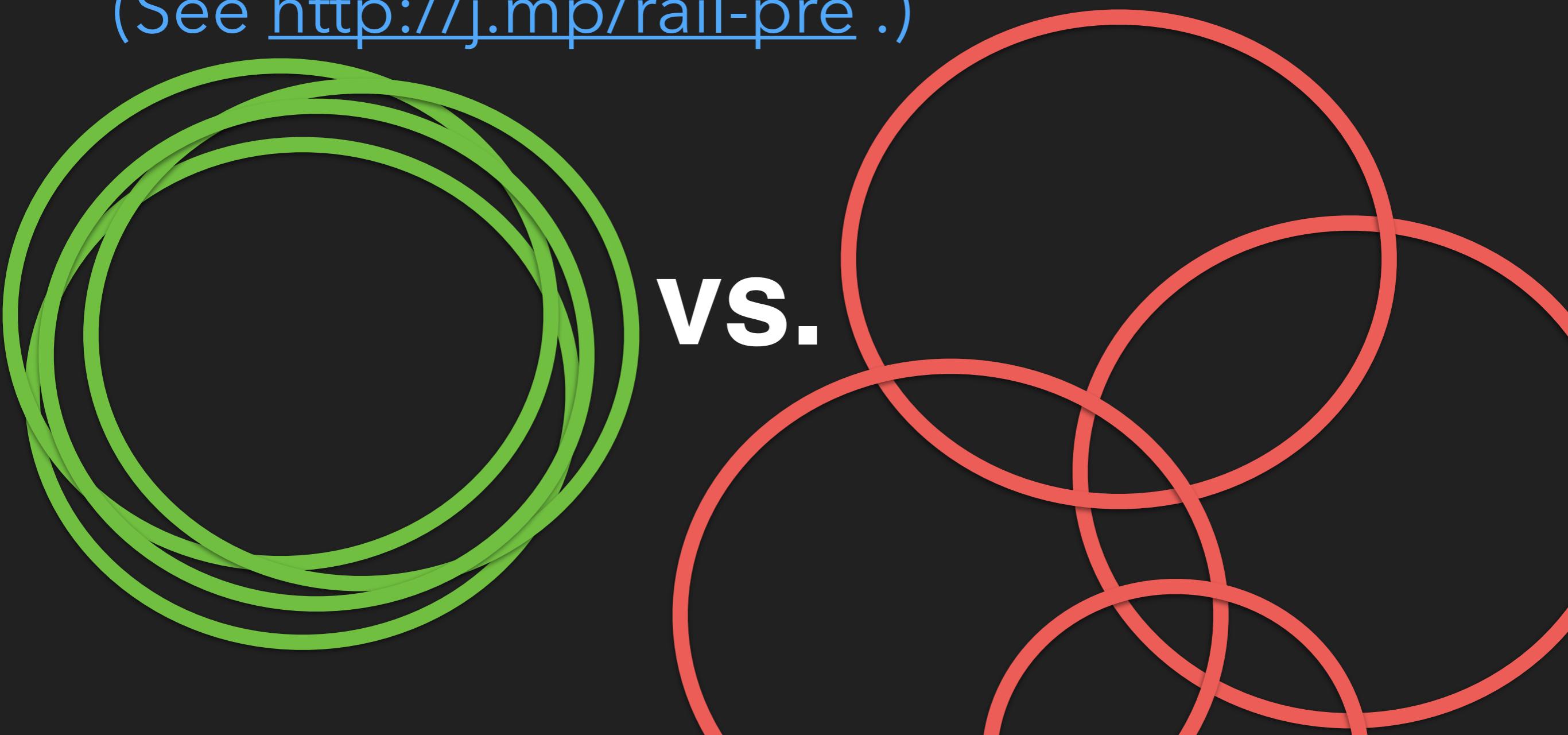
goods

On a single sample,
every aligner will find
some good junctions
and some duds (or
very rare junctions).

Why so many junctions?

Comparing the junctions found in many simulated samples, there is *much more overlap between goods than between duds.*

(See <http://j.mp/rail-pre> .)



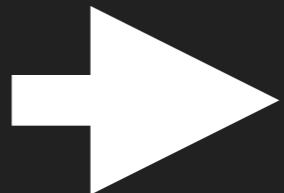
Why so many junctions?

So as you add
samples...

junctions



duds



goods

junctions



duds

goods

We ran

```
rail-rna prep elastic (~500 runs)
--manifest batch_X.tsv
--core-instance-count 20
--output s3://bucket/batch_X_prepended
--core-instance-bid-price 0.13
--master-instance-bid-price 0.13
--core-instance-type c3.2xlarge
--master-instance-type c3.2xlarge
```

for $X \in \{0, \dots, 42\}$

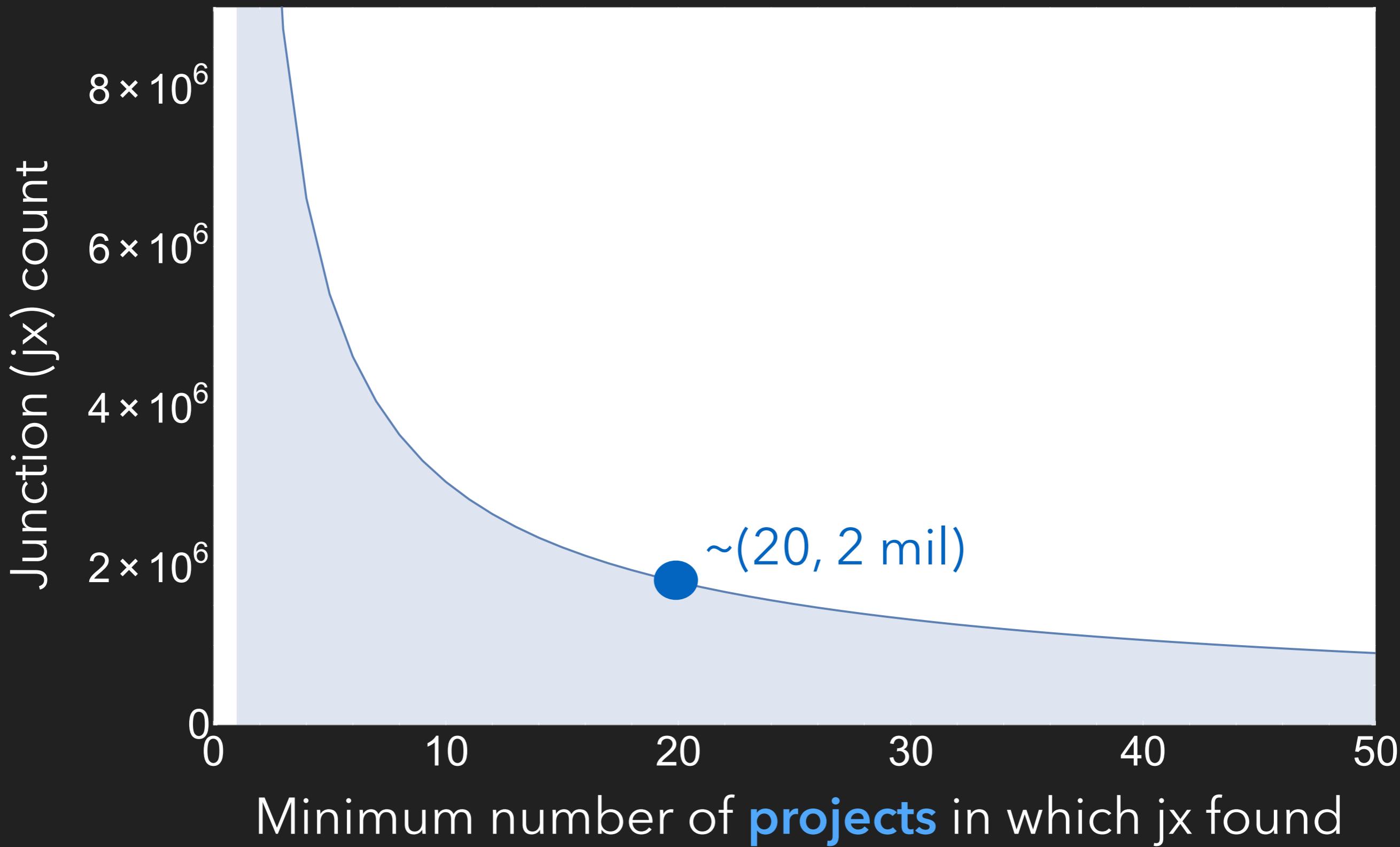
to download/preprocess data, copy to S3

We ran

```
rail-rna align elastic  
--manifest batch_X.tsv  
--input s3://bucket/batch_X_prepended  
--output s3://bucket/batch_X_itn  
--core-instance-bid-price 0.60  
--master-instance-bid-price 0.60  
--core-instance-count 60  
--core-instance-type c3.8xlarge  
--master-instance-type c3.8xlarge  
--deliverables itn
```

to detect junctions from one pass of
alignment

A steep dropoff: project-level



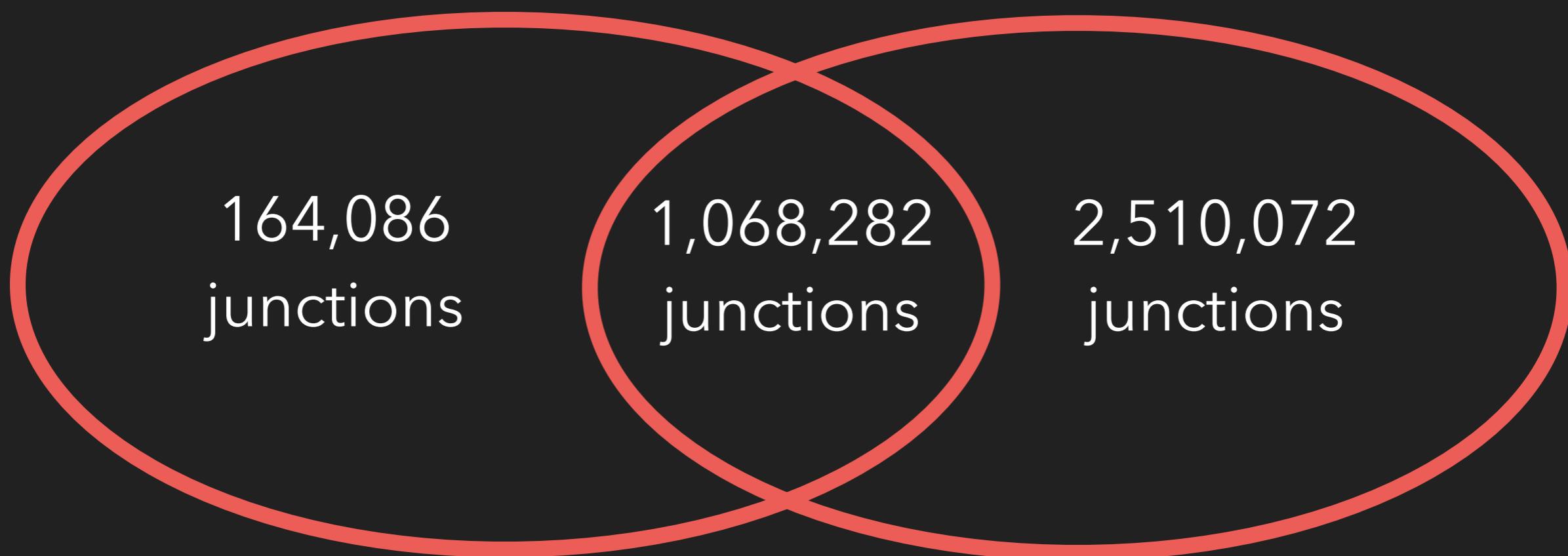
Comparison with SEQC

SEQC/MAQC-III (Nat Biotech, doi:10.1038/nbt.2957)

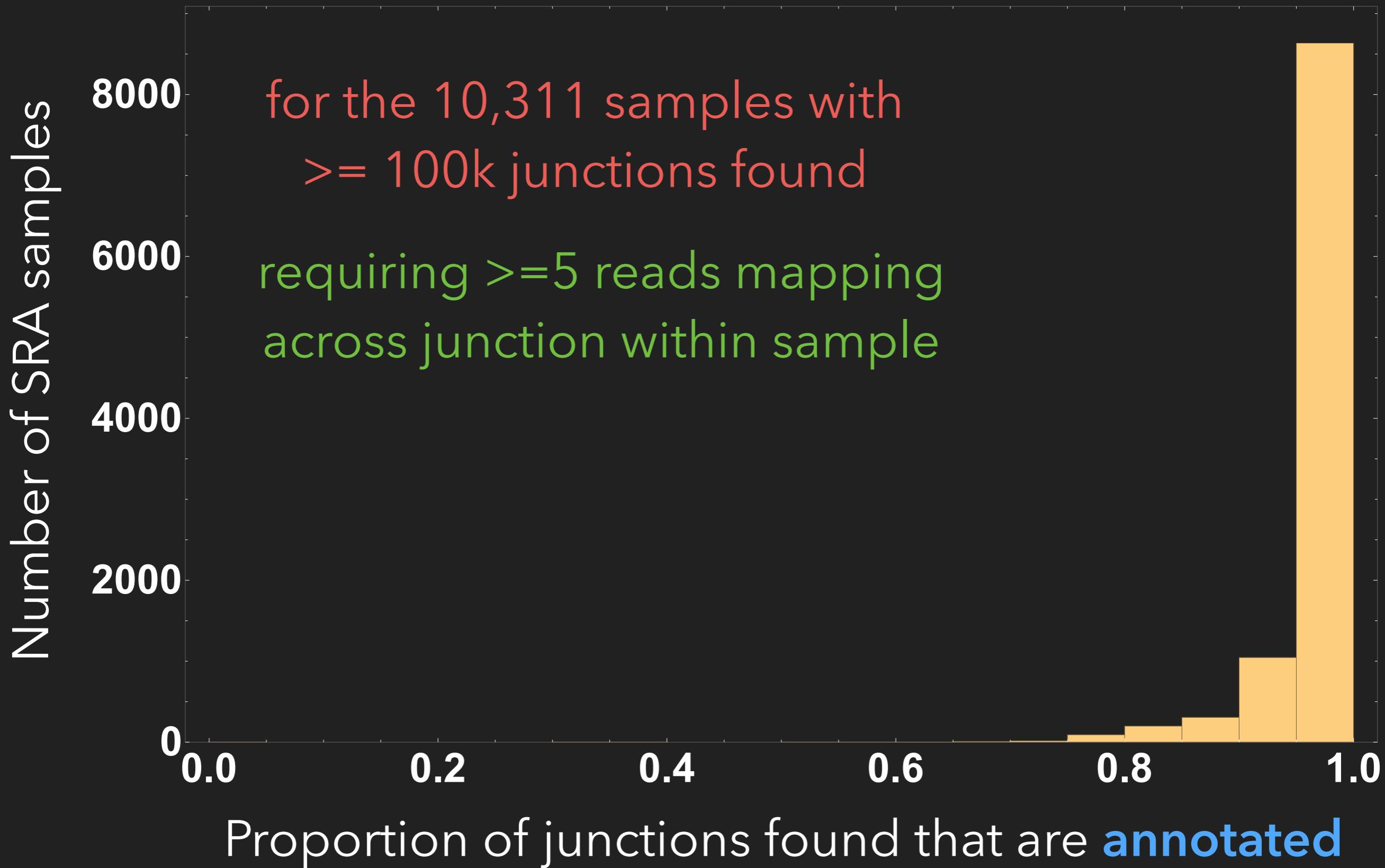
1720 samples in common with Rail;
universal human & brain reference samples

Rail-RNA
in at least
5 SEQC samples

One of rmake,
magic,
and subread



Annotated junctions by sample



Gedankenexperiments

More complete annotation = **better!**

Increased sensitivity

(Can detect isoform 2 now!)

isoform 1



(new) isoform 2



More complete annotation = **worse!**

Decreased specificity

(What if isoform 2 is really rare?)

Actual experiment

RGASP simulated sample 1
(40 mil read pairs)

HISAT2 2.0.0-beta

		junctions	junction overlaps
fed true jx	prec:	0.94	0.98
	rec:	0.99	0.93
fed union of annotated jx	prec:	0.80	0.97
	rec:	0.95	0.92

Actual experiment

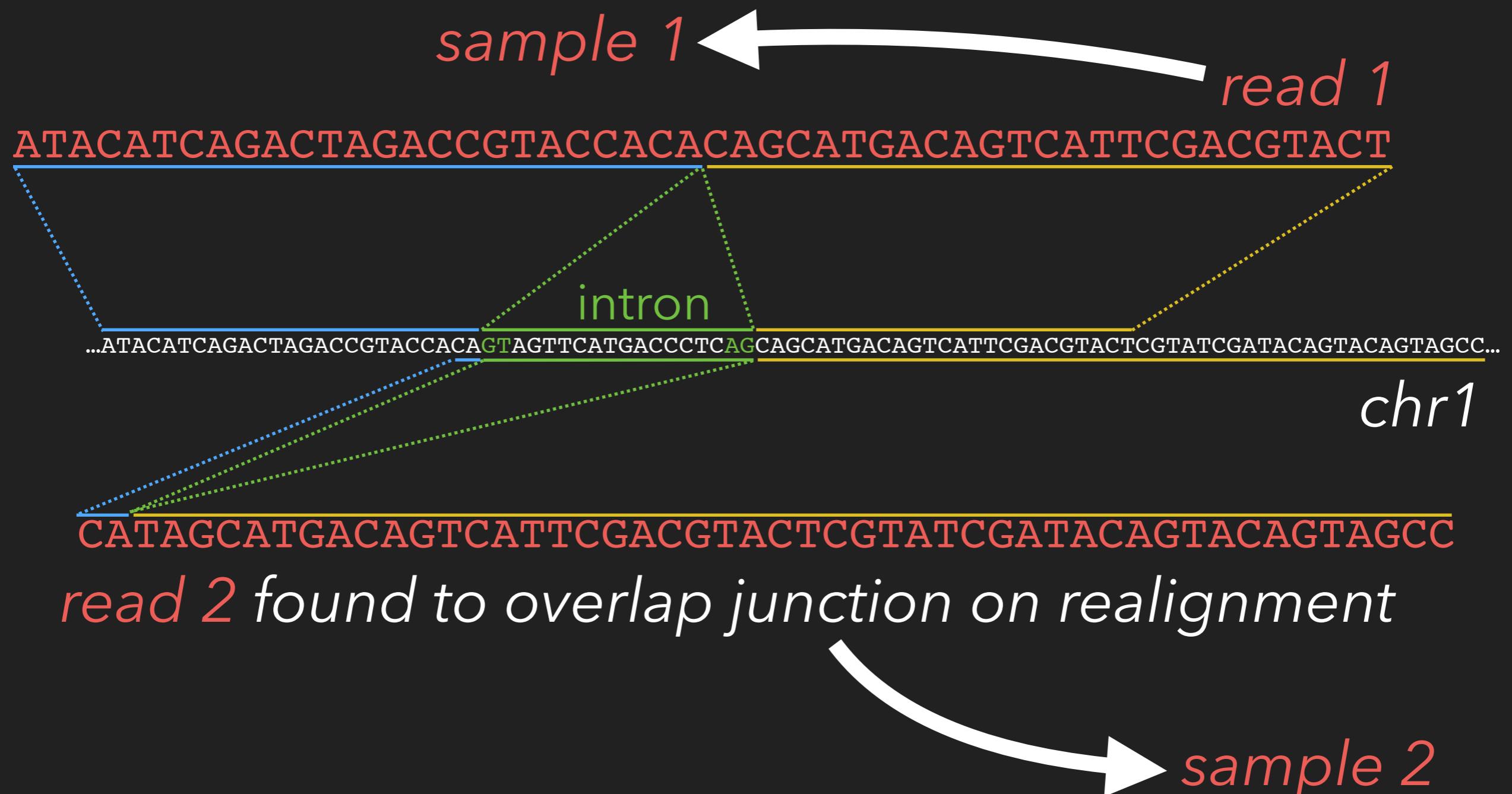
RGASP simulated sample 1
(40 mil read pairs)

STAR 2.4.2a

		junctions	junction overlaps
fed true jx	prec:	0.98	0.995
	rec:	0.99	0.91
fed union of annotated jx	prec:	0.90	0.98
	rec:	0.97	0.87

Rail-RNA's approach

Realign after collecting and filtering a list of junctions across **SIMILAR** samples.



similar: same feature to which you
want to be sensitive

cell line, tissue type, population, experimental condition...

Junction (jx) filter

Keep a junction if and only if
it's initially detected in:

(1) 5% of samples

← grabs common jx

OR

(2) at least 5 reads in any
one sample

← so we don't miss jx
that are probably
there but unique to
a sample

Comparison

Simulate from annotation,
then give competitors annotation

112 simulated LCLs (based on GEUVADIS)

mean overlap accuracy value | mean junction accuracy value

	Precisions	Recalls	F-scores
TopHat 2 ann	.815 .947	.839 .982	.826 .964
STAR ann	.882 .977	.874 .980	.878 .979
HISAT ann	.895 .922	.857 .982	.875 .951
Rail	.969 .976	.858 .939	.910 .957

(<http://j.mp/rail-pre>)

A steep dropoff

