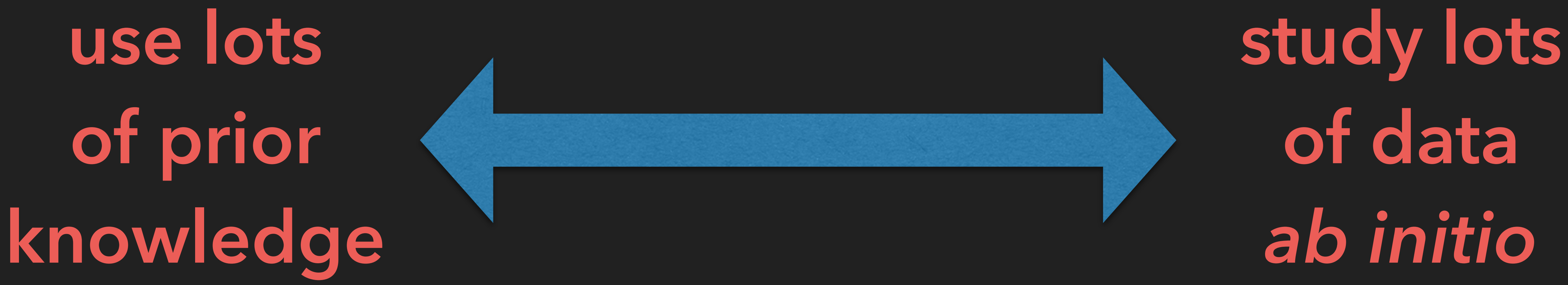
# An analysis of splicing variation across SRA with Rail-RNA

@AbhiNellore

Johns Hopkins University

Genome Informatics 2015

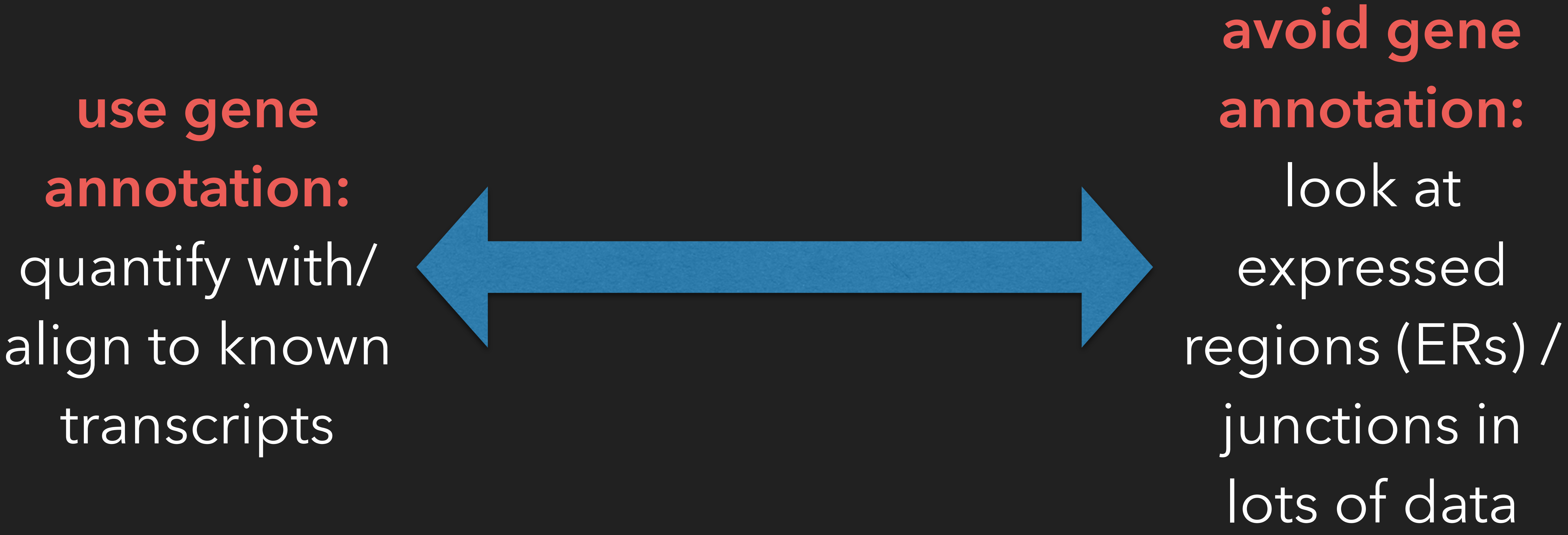
# in genomics



use lots of prior knowledge

study lots of data *ab initio*

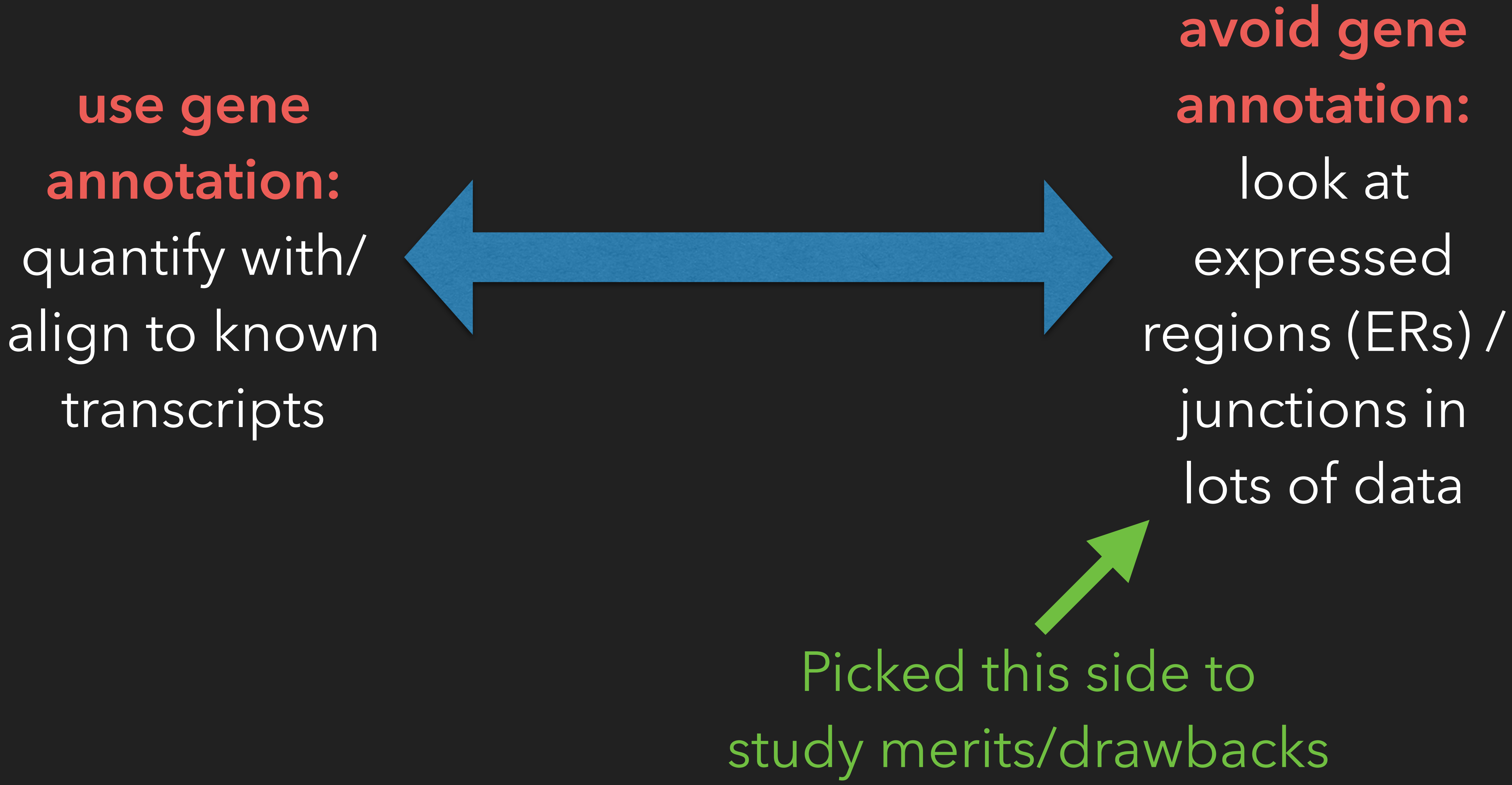
# in RNA-seq analysis



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

**avoid gene annotation:** look at expressed regions (ERs) / junctions in lots of data

# in RNA-seq analysis



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

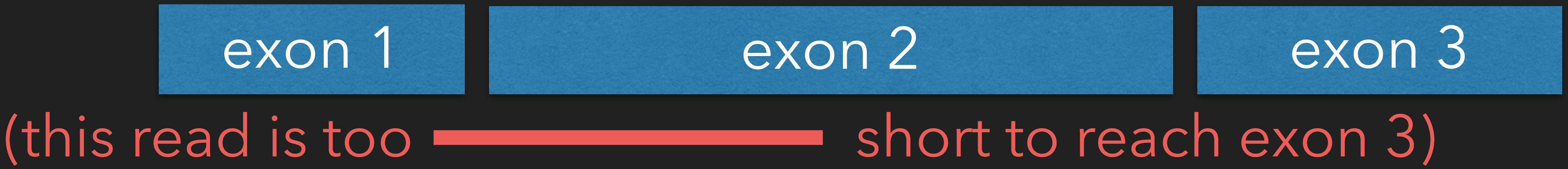
**avoid gene annotation:** look at expressed regions (ERs) / junctions in lots of data

Picked this side to study merits/drawbacks

# Study many RNA-seq samples

SRA: short reads hard to assemble; missing exons in **60%** of transcripts

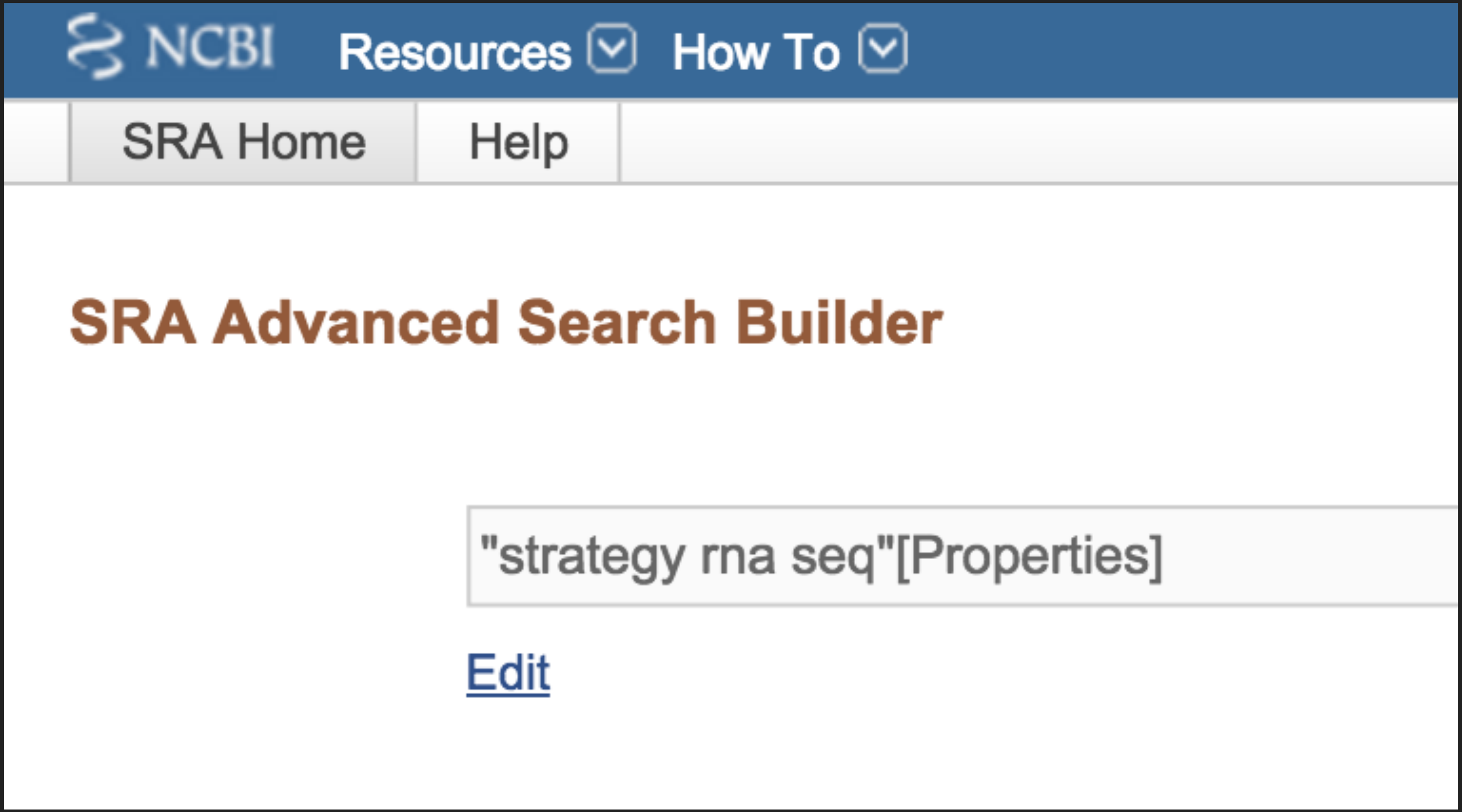
(RGASP 2013 doi:10.1038/nmeth.2714)



(this read is too short to reach exon 3)

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

# Filtering SRA



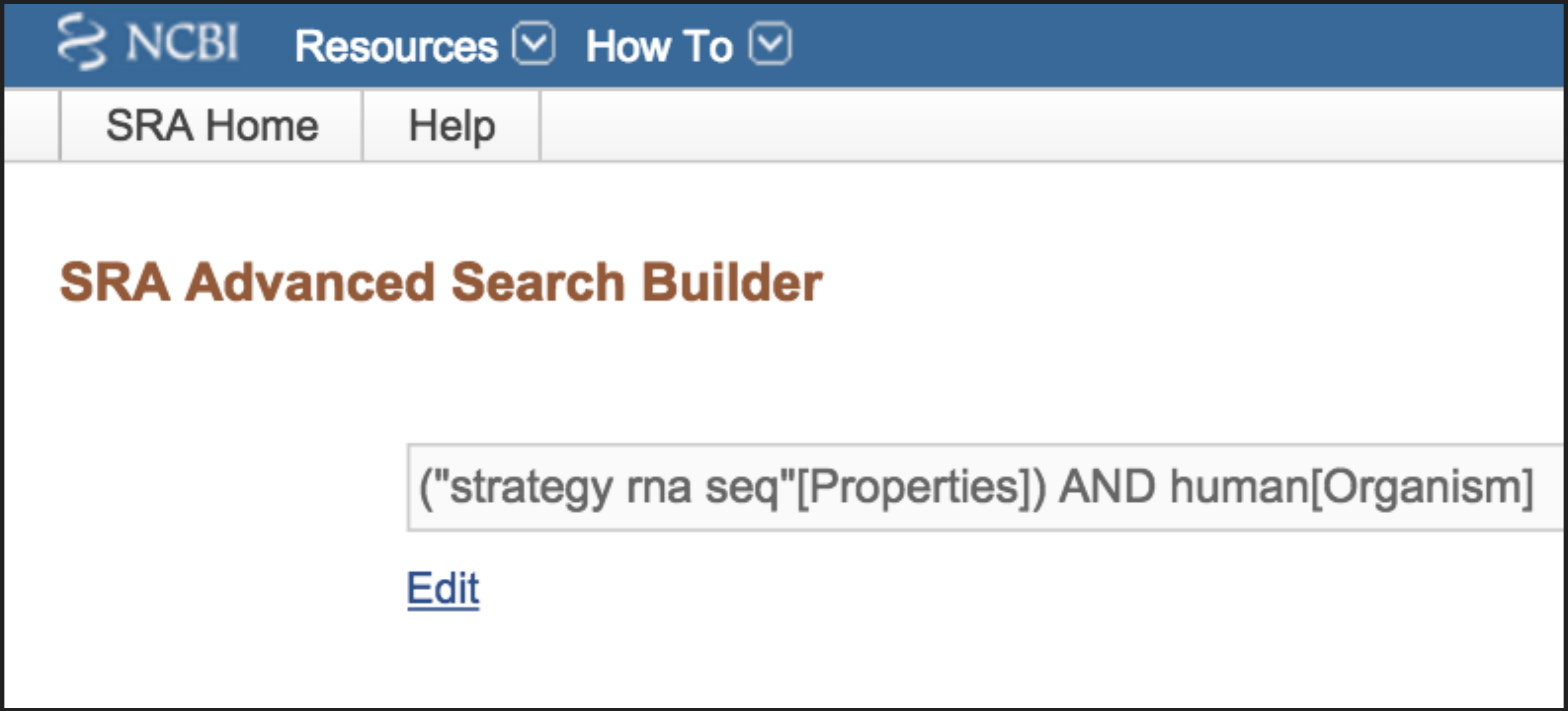
SRA Advanced Search Builder

"strategy rna seq"[Properties]

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

≈180k publicly available runs

# Filtering SRA



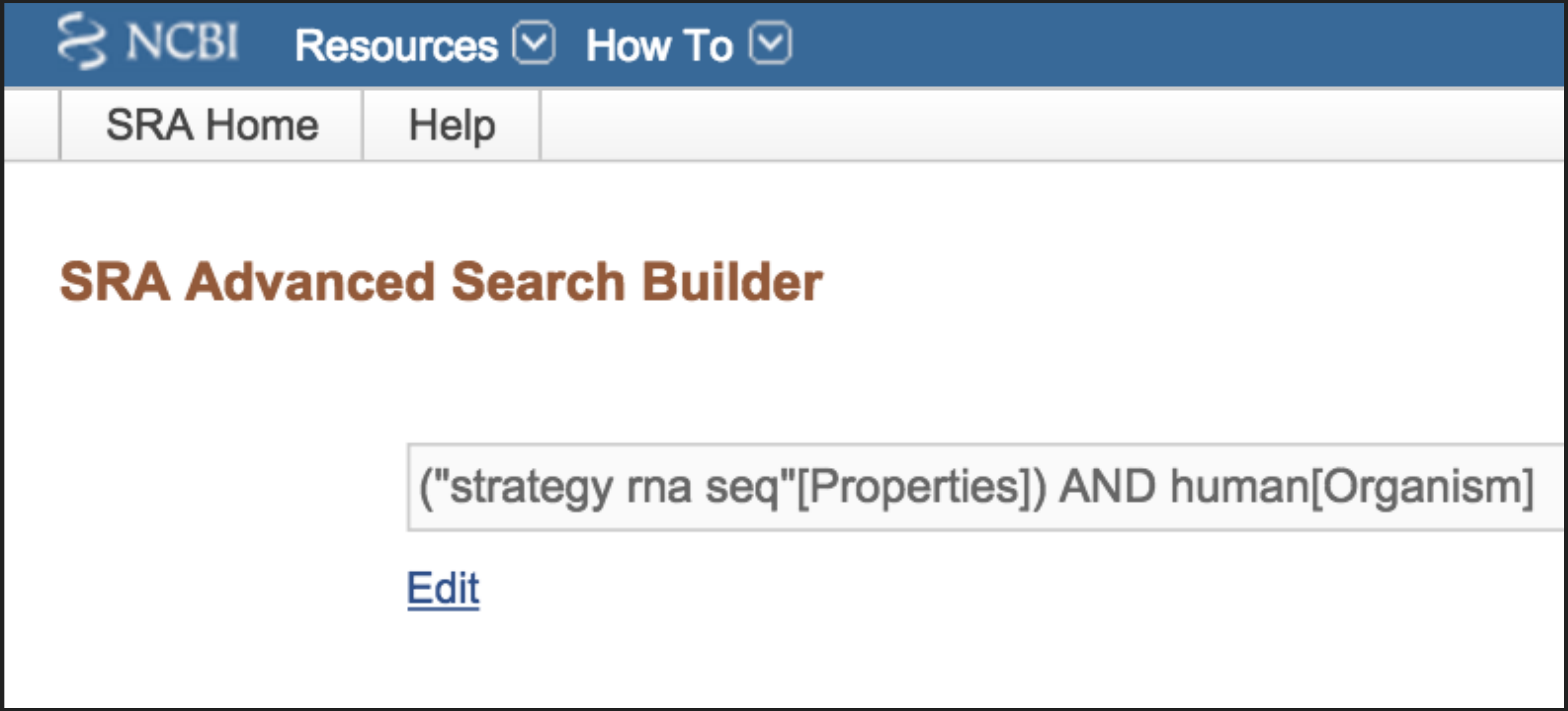
SRA Advanced Search Builder

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

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

≈36k publicly available runs

# Filtering SRA



SRA Advanced Search Builder

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

(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)

Rail-RNA + amazon webservices™

- No competition for compute
- Rapid: 8 days to data
- Reproducible:

  http://github.com/nellore/gi2015

  for commands (& goodies!)
- Cheap: ~$0.70/sample

# What gene annotation says

For *hg19*,

$\text{Ensembl v75} \cup \text{GENCODE v19} \cup \text{RefSeq}$

(almost subsumed
by Ensembl v75)

≈350,000

junctions

2 commands x 43 batches

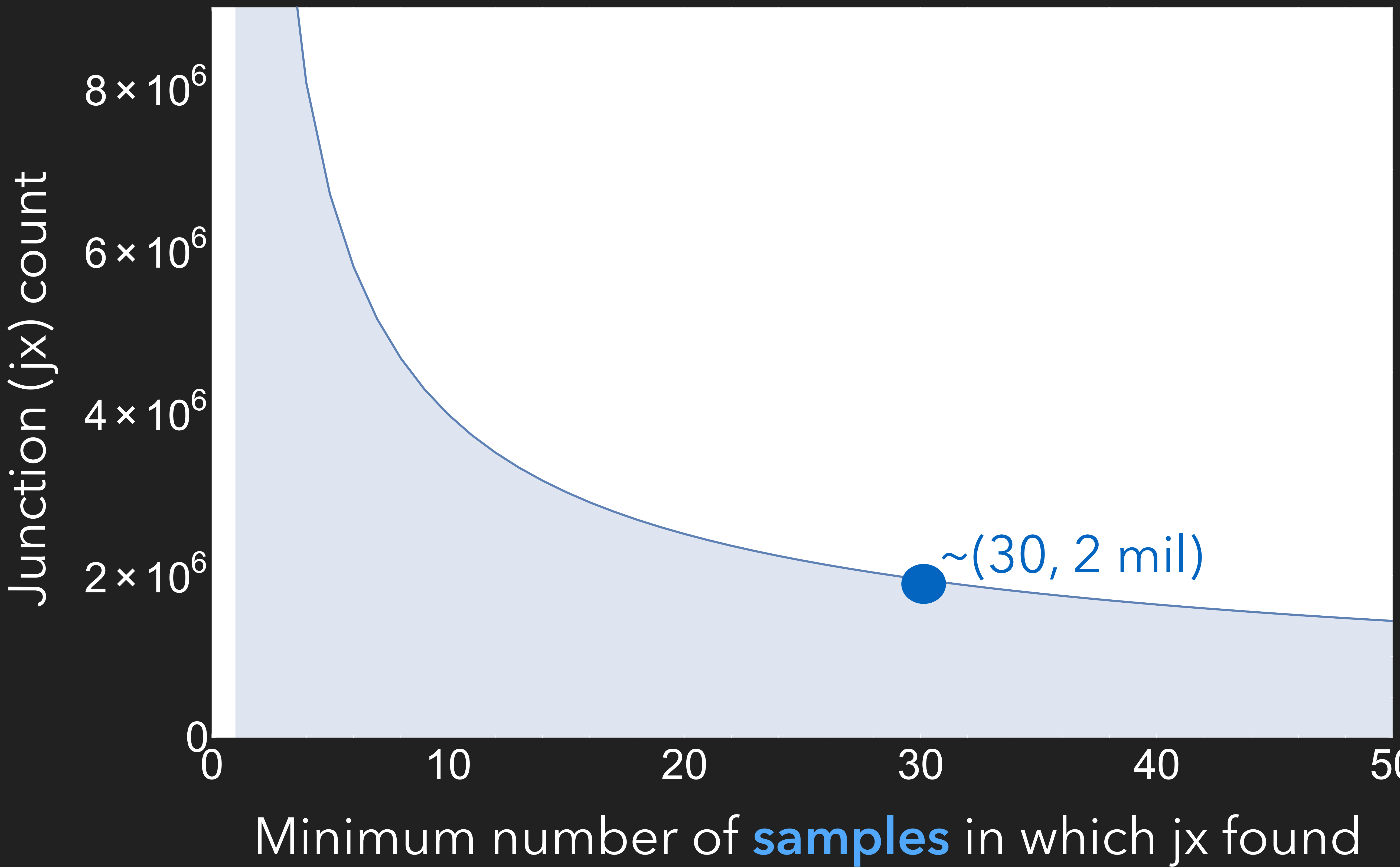
gave, across 21,504 samples

One 7-GB `tsv.gz`

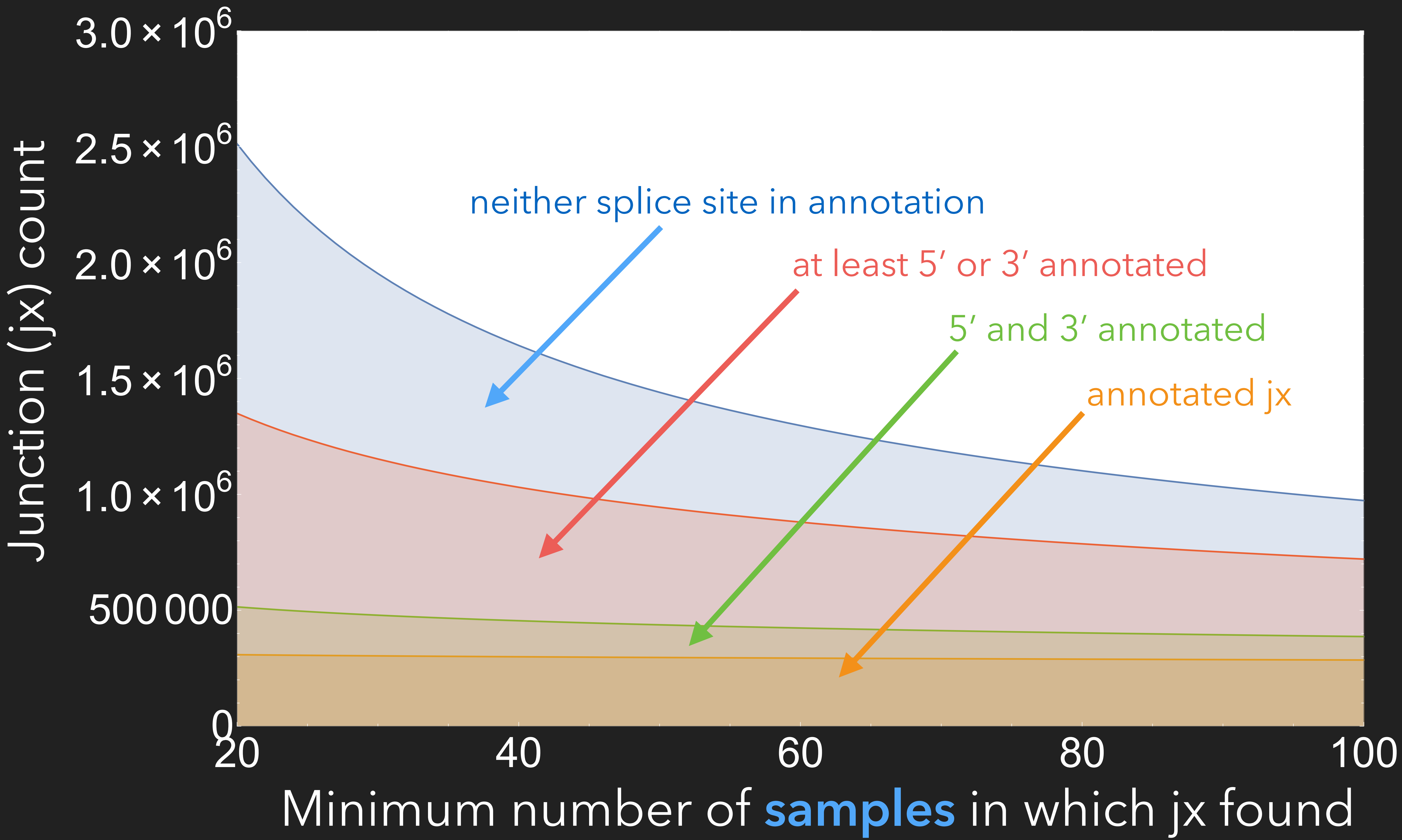
# 42,882,032

junctions

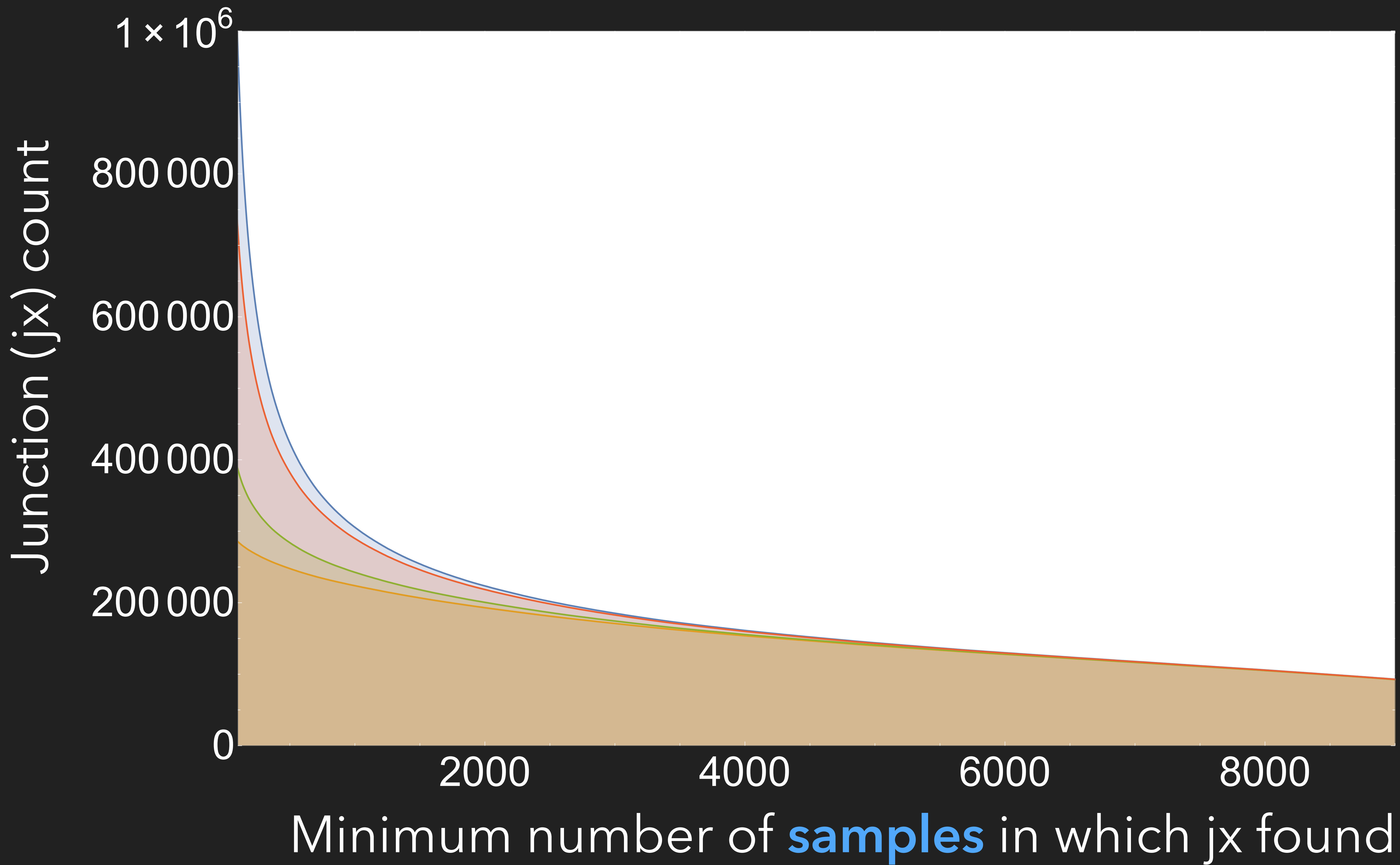
# A steep dropoff



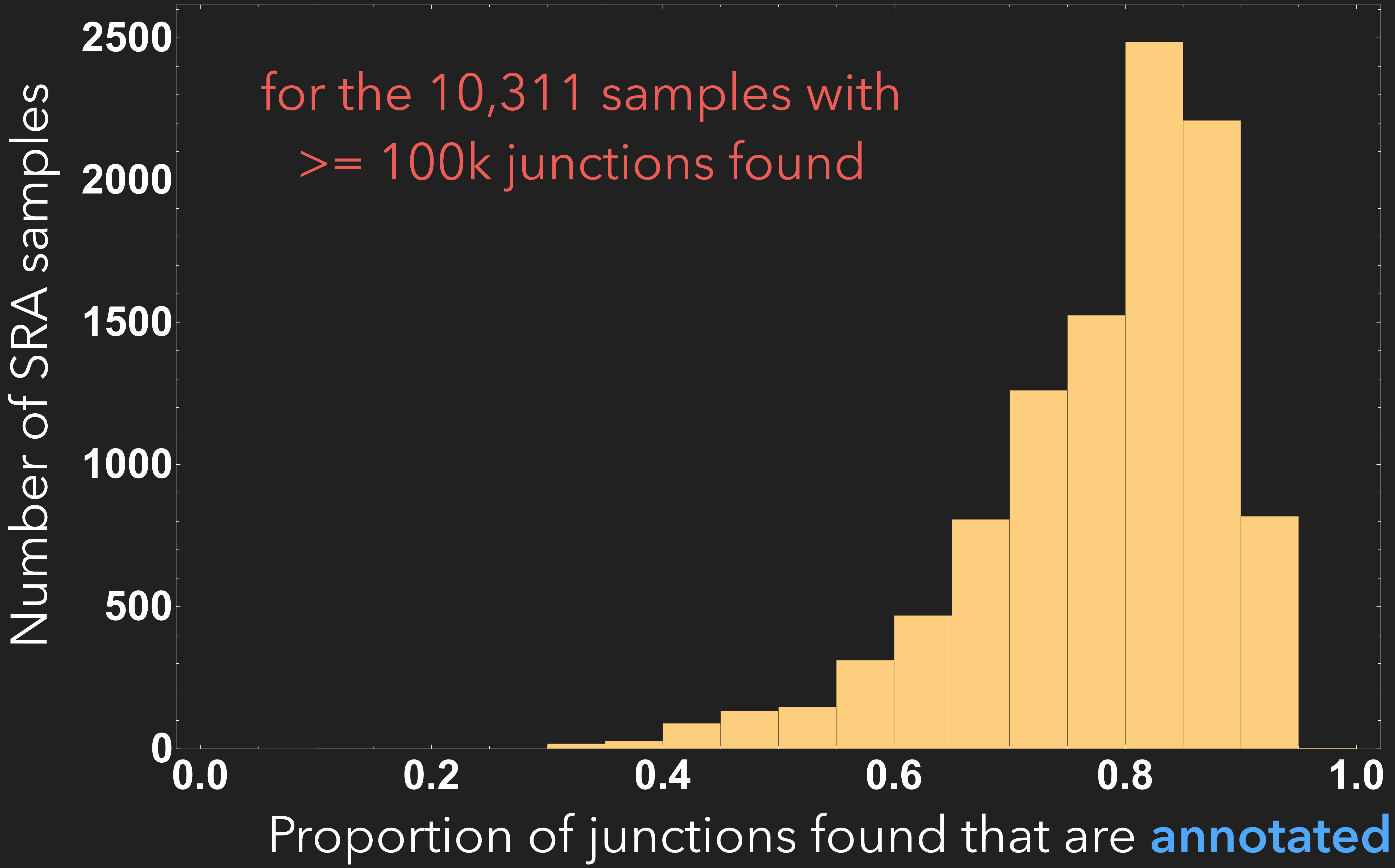
# Increasing evidence in annotation



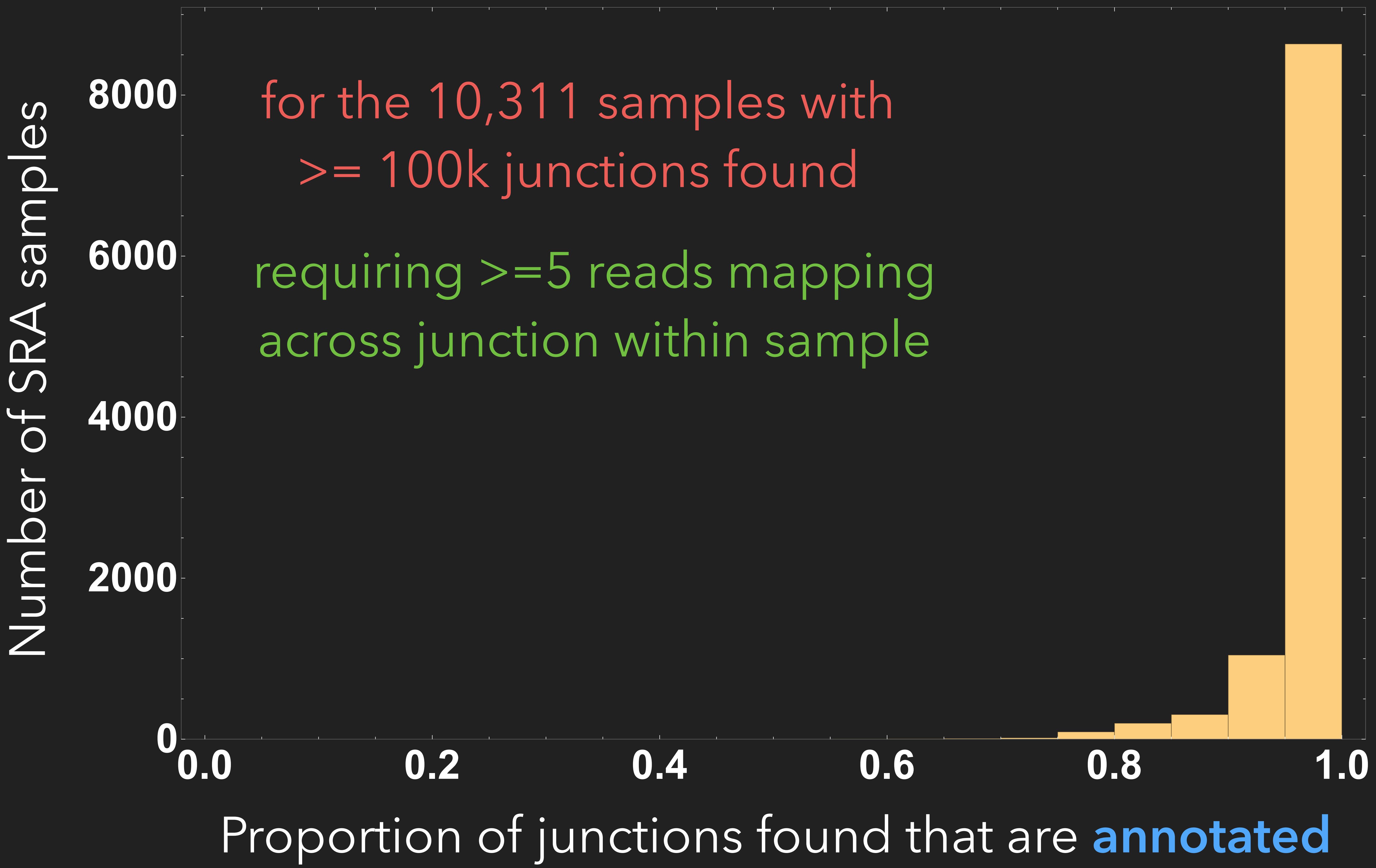
# Asymptote to annotation



# Annotated junctions by sample



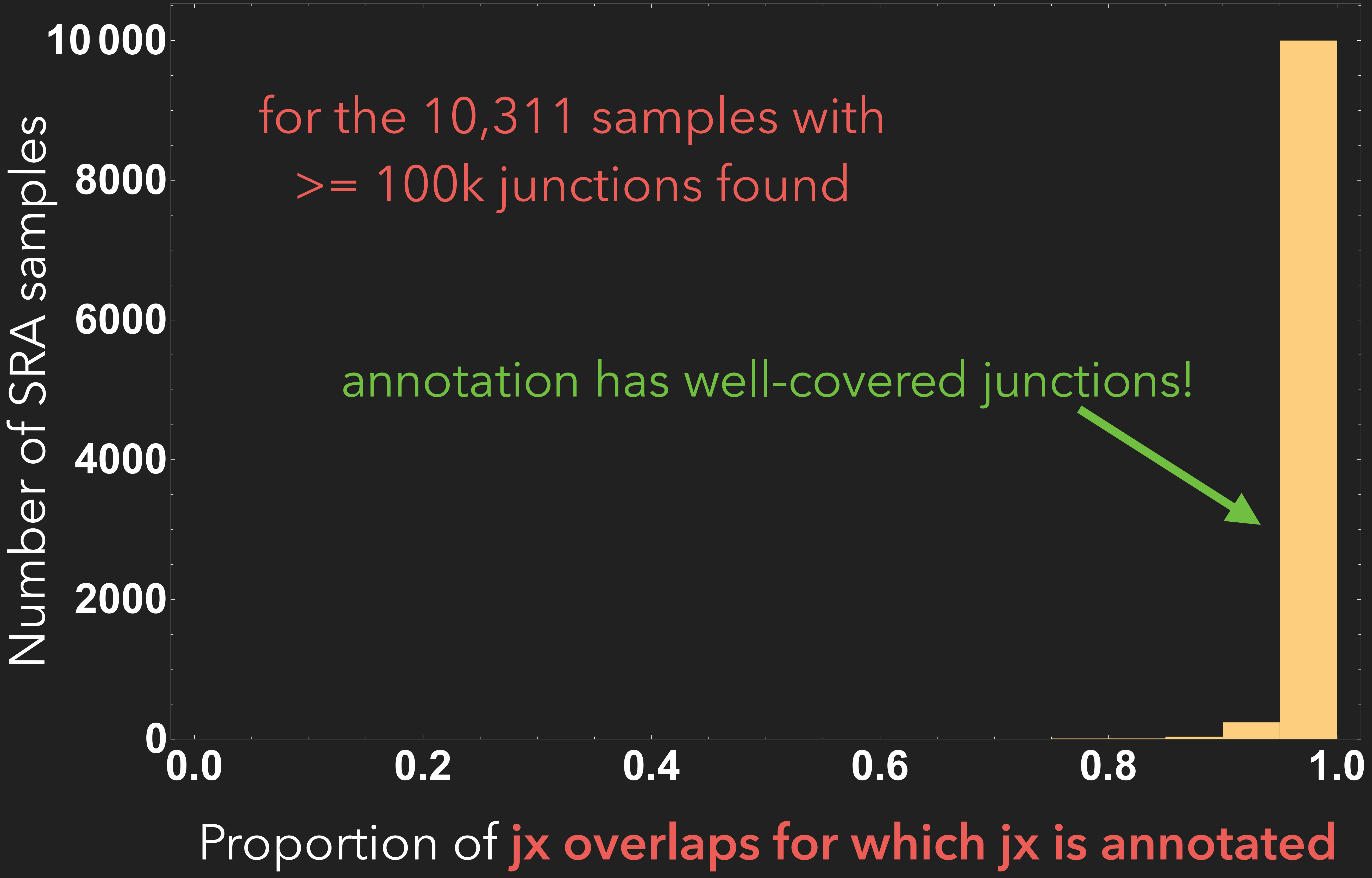
# Annotated junctions by sample



for the 10,311 samples with
>= 100k junctions found

requiring >=5 reads mapping
across junction within sample

# Junction (jx) overlaps by sample



for the 10,311 samples with >= 100k junctions found

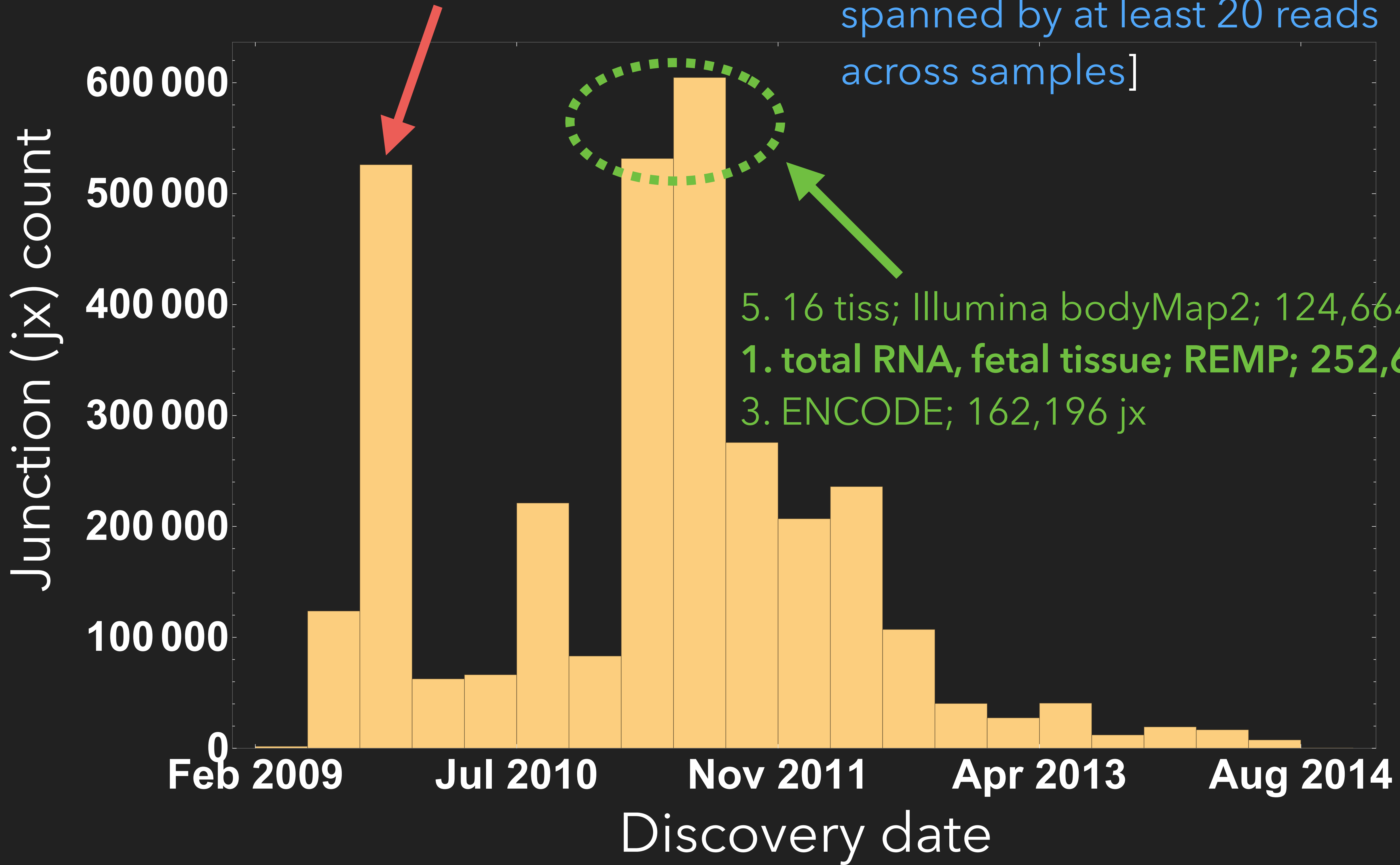
annotation has well-covered junctions!

# Are we still finding new junctions?

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]



5. 16 tiss; Illumina bodyMap2; 124,664 jx

**1. total RNA, fetal tissue; REMP; 252,628 jx**

3. ENCODE; 162,196 jx

So just make annotation better!

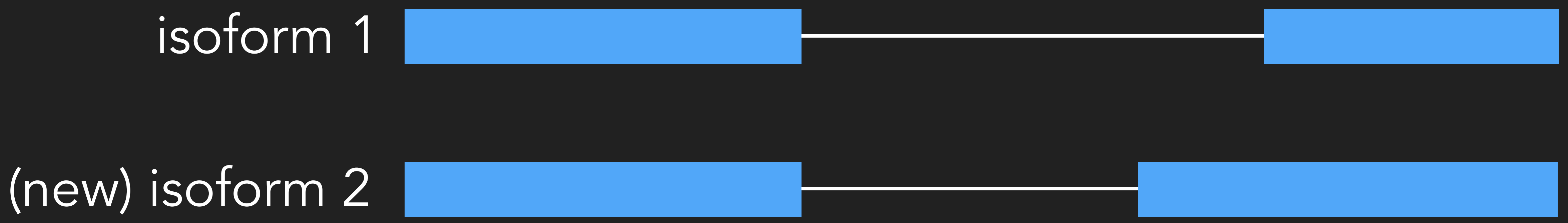
Not so fast.

# Gedankenexperiments

More complete annotation = **better!**

Increased sensitivity

(Can detect isoform 2 now!)



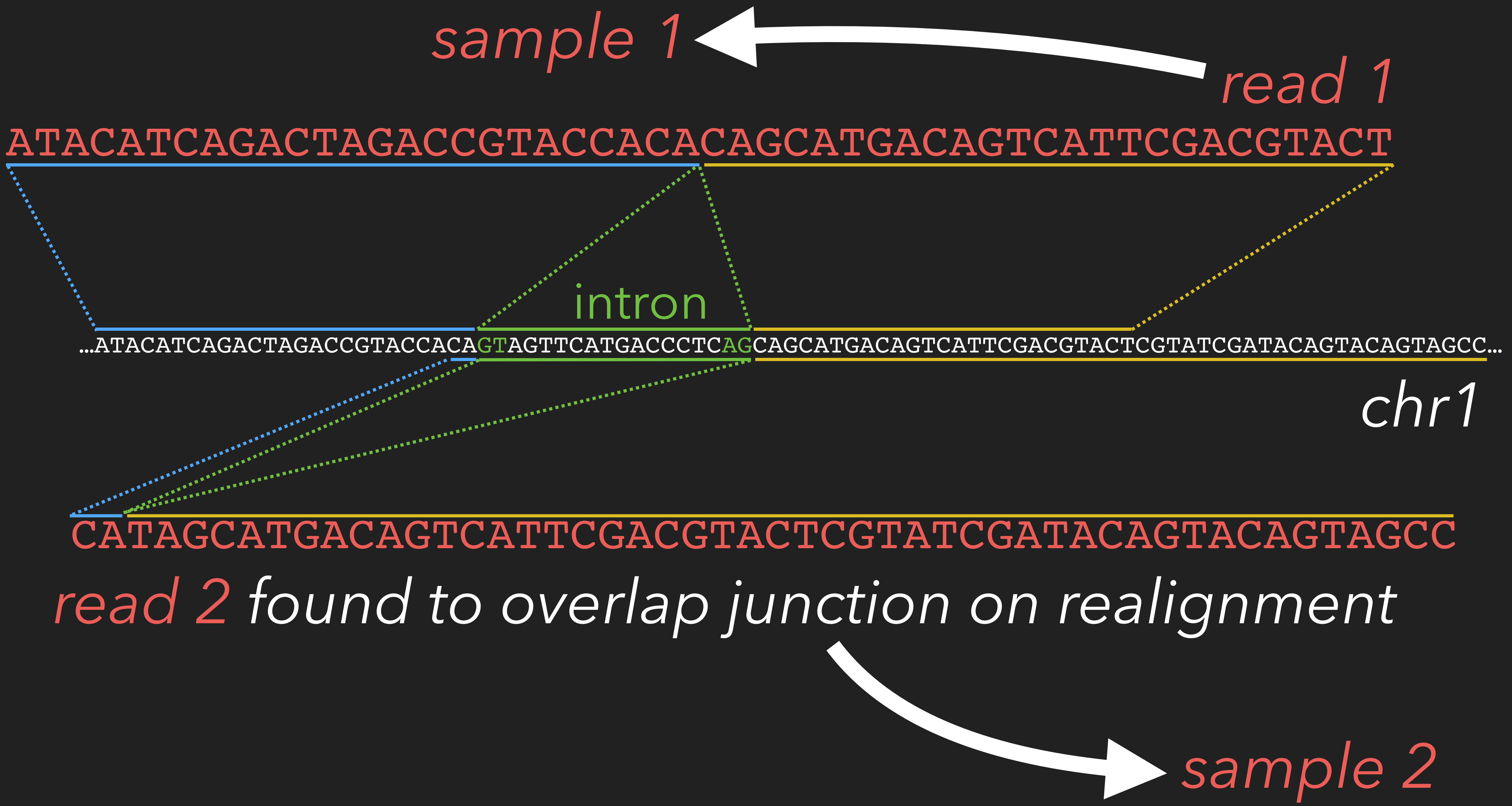
More complete annotation = **worse!**

Decreased specificity

(What if isoform 2 is really rare?)

# Rail-RNA’s approach

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



*sample 1*

*read 1*

ATACATCAGACTAGACCGTACCACACAGCATGACAGTCATTCGACGTACT

intron

…ATACATCAGACTAGACCGTACCACAGTAGTTCATGACCCTCAGCAGCATGACAGTCATTCGACGTACTCGTATCGATACAGTACAGTAGCC…

*chr1*

CATAGCATGACAGTCATTCGACGTACTCGTATCGATACAGTACAGTAGCC

*read 2 found to overlap junction on realignment*

*sample 2*

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)

# annotation-agnostic pipeline

Rail-RNA http://rail.bio

↓

derfinder

```
biocLite("derfinder")
```

Leo
Collado-Torres

Alyssa
Frazee

# `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)

scripts for recovering junctions

&

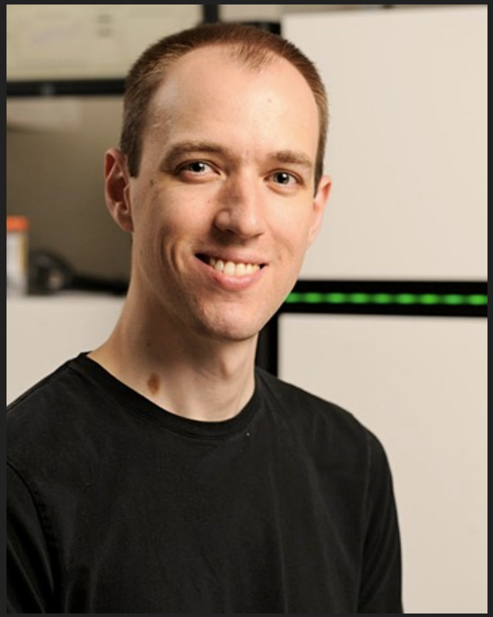
processed data

@

**http://github.com/nellore/gi2015**

# Collaborators

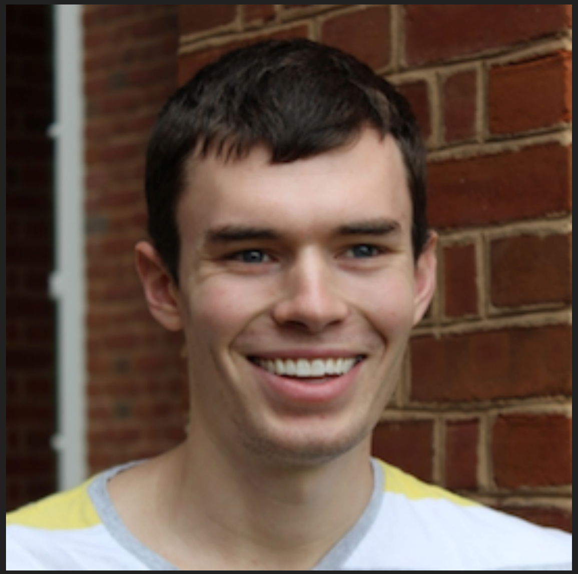
Jeff Leek



Ben Langmead

Leo Collado-Torres

Andrew Jaffe



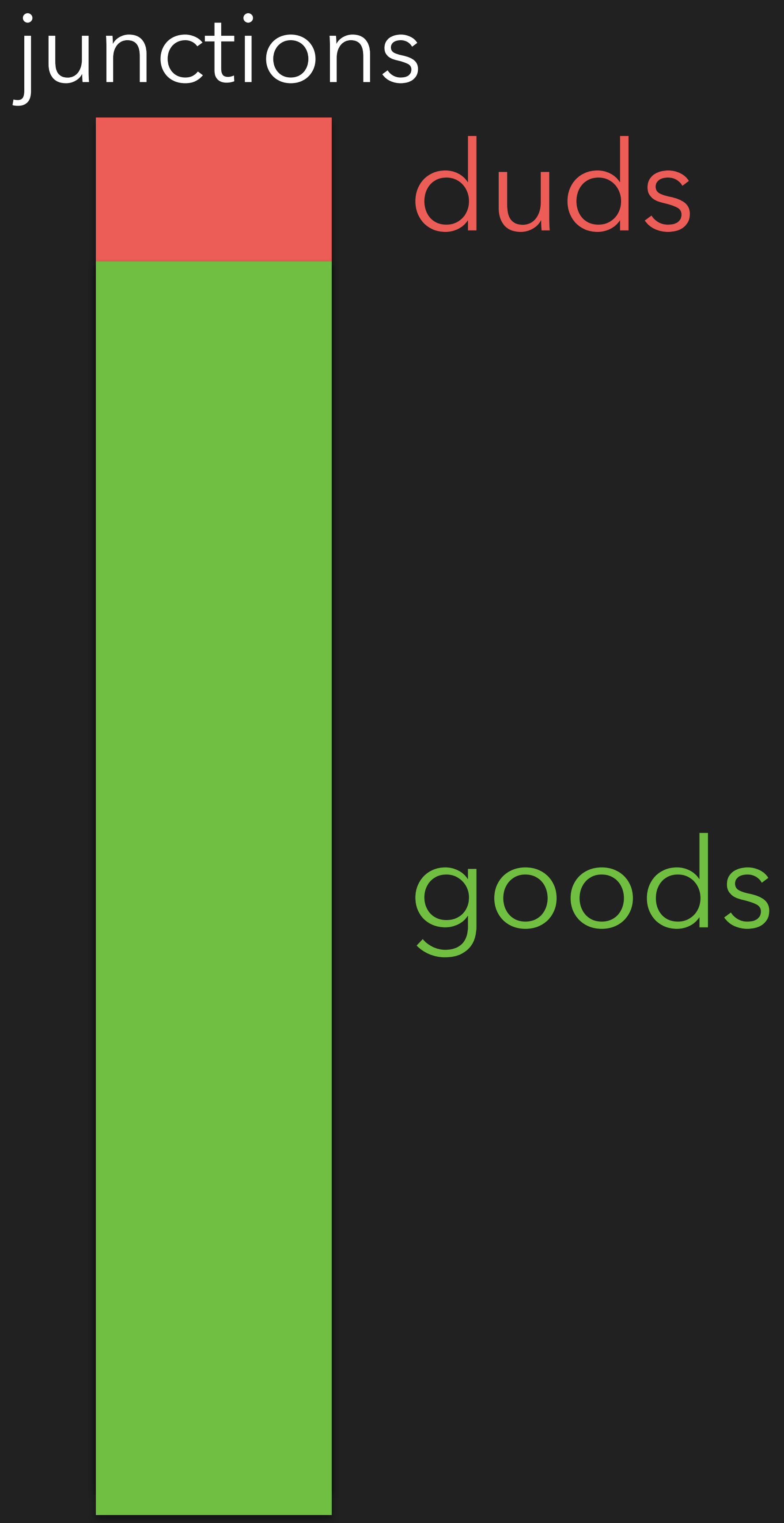
Jacob Pritt

Chris Wilks

José Alquicira Hernández

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

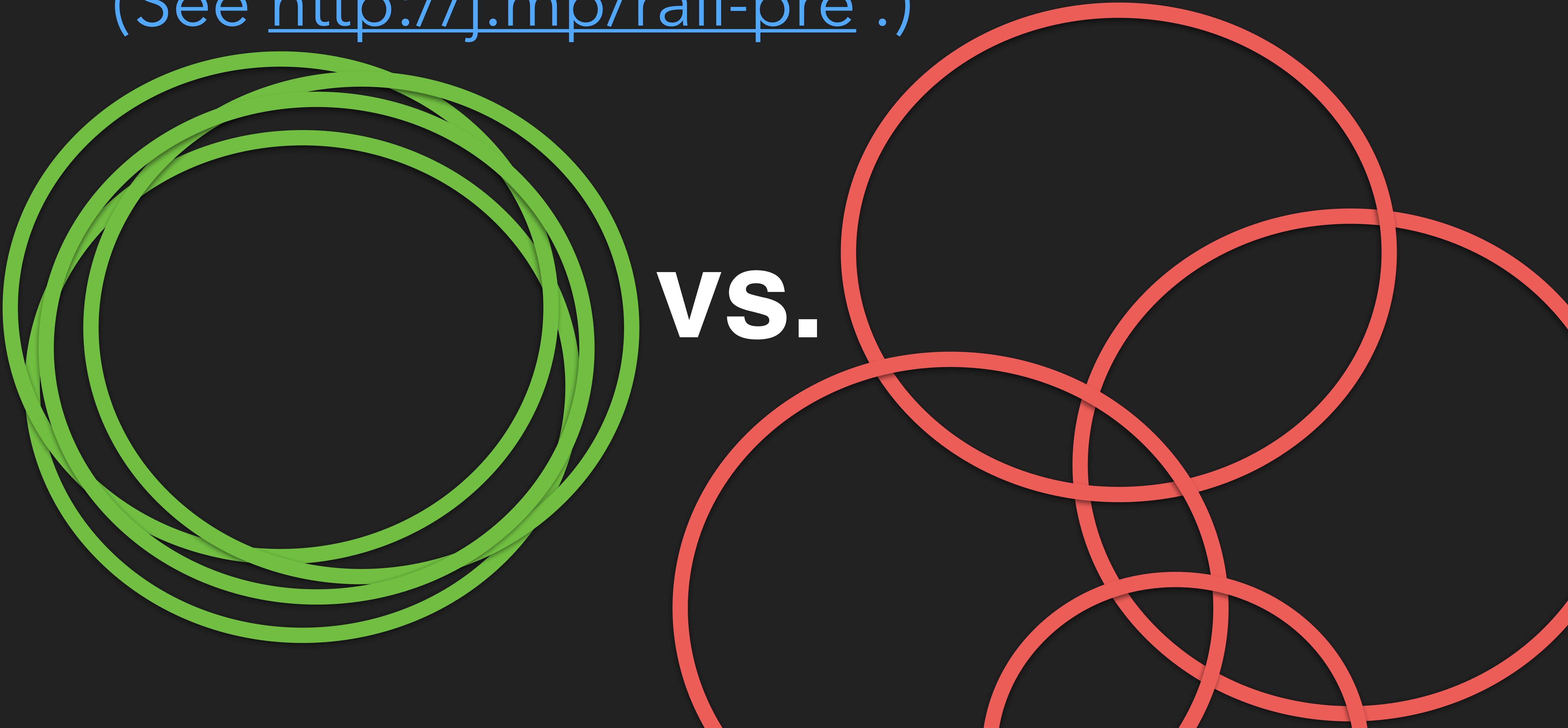
# Why so many junctions?



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

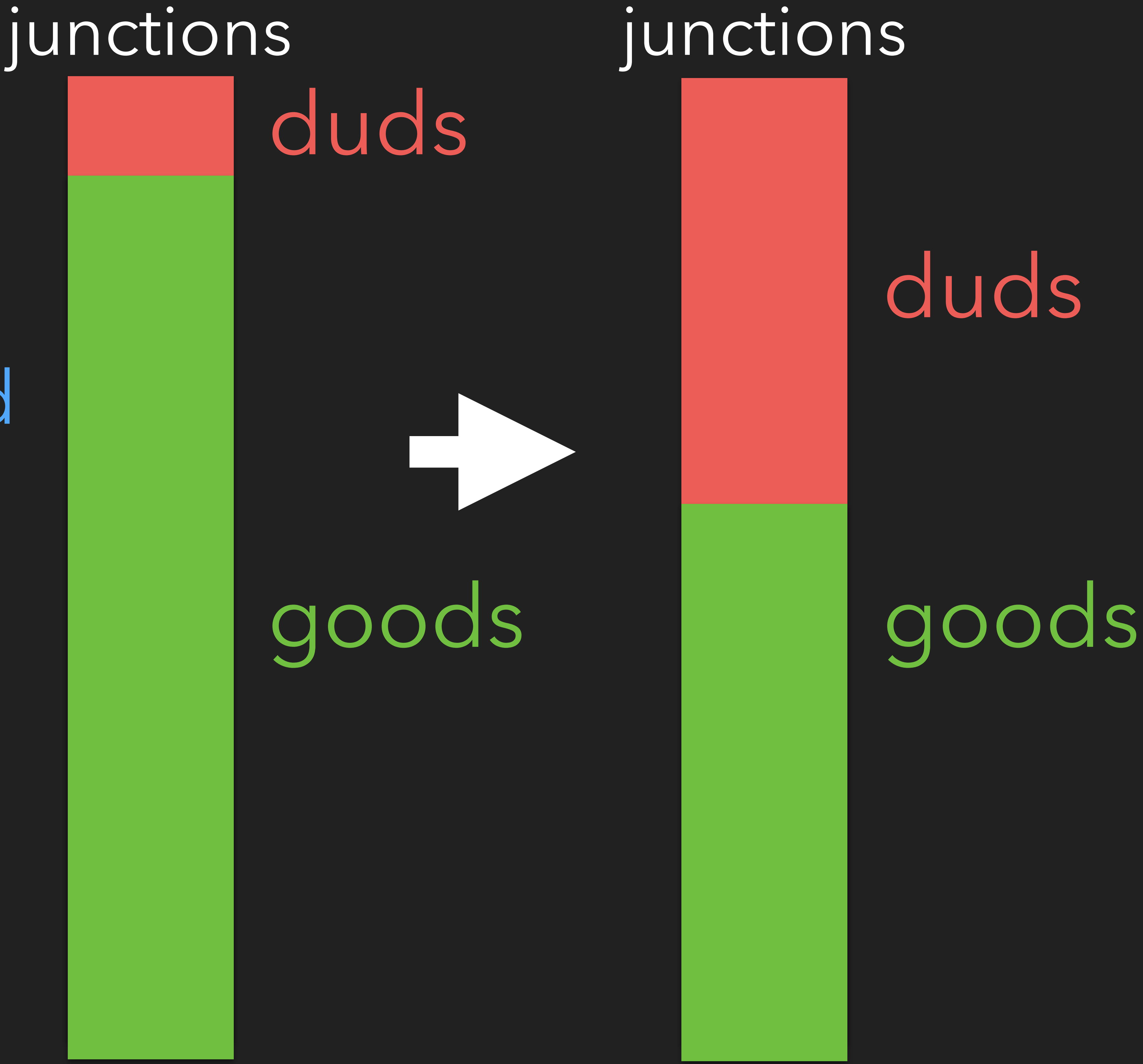
# Why so many junctions?

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…



# We ran

(~500 runs)

```
rail-rna prep elastic
—-manifest batch_x.tsv
—-core-instance-count 20
—-output s3://bucket/batch_x_prepped
--core-instance-bid-price 0.13
--master-instance-bid-price 0.13
—-core-instance-type c3.2xlarge
—-master-instance-type c3.2xlarge
```

for $x \in \{0, \ldots, 42\}$

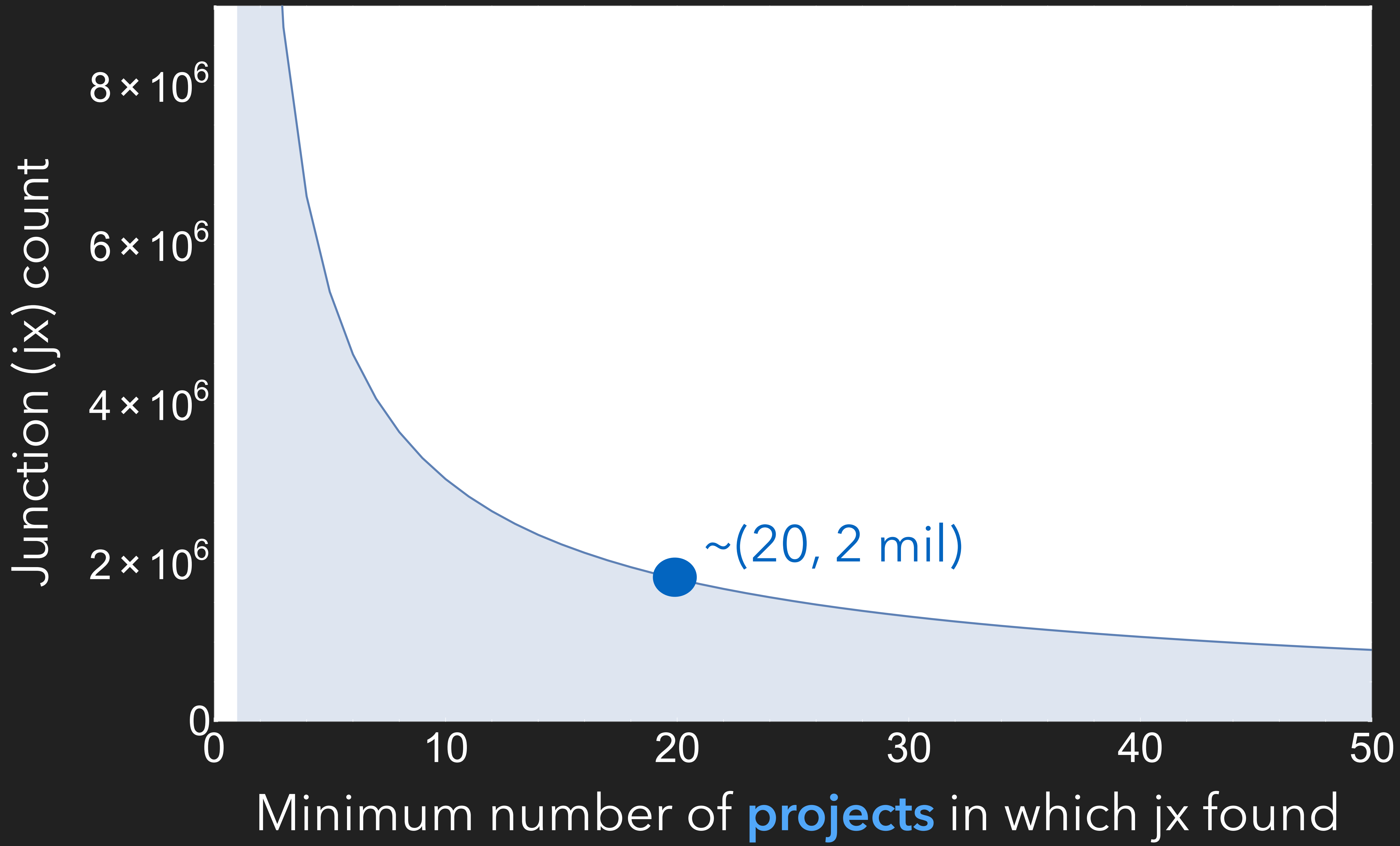
to download/preprocess data, copy to S3

# We ran

```
rail-rna align elastic
—-manifest batch_X.tsv
—-input s3://bucket/batch_X_prepped
—-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



# 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)

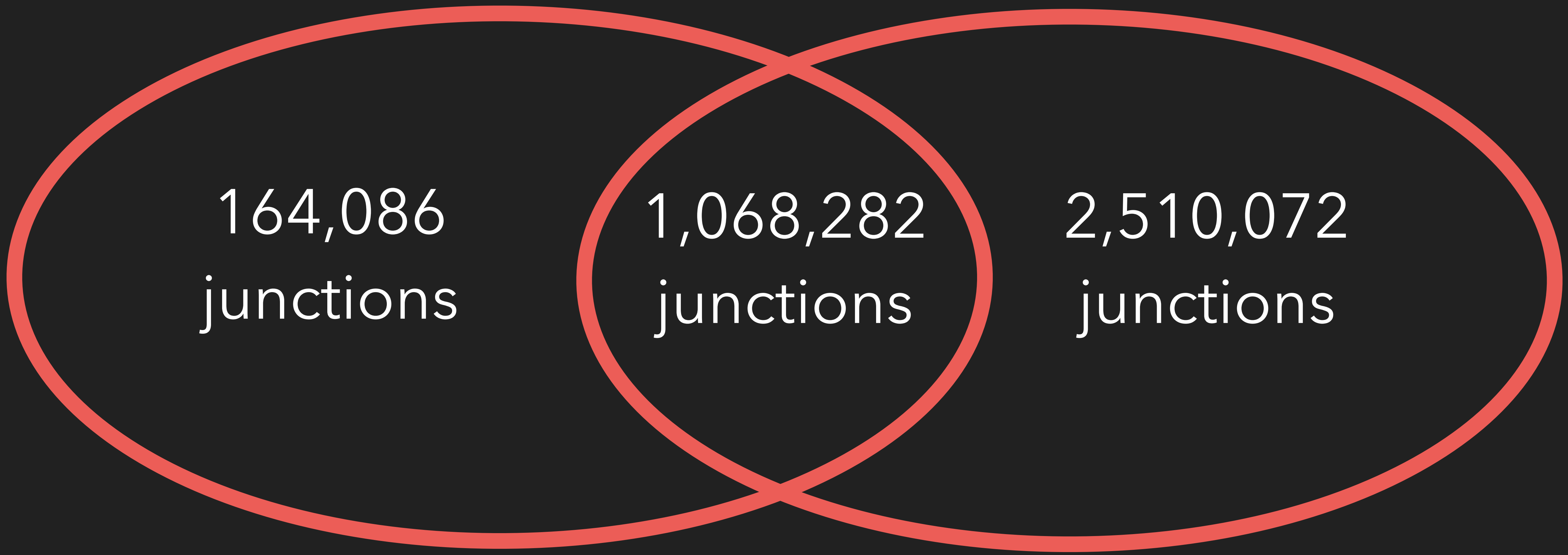
## STAR 2.4.2a

| | | 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 |

# 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

164,086
junctions

1,068,282
junctions

2,510,072
junctions