An analysis of splicing variation across SRA with Rail-RNA

@AbhiNellore
Johns Hopkins University

Genome Informatics 2015

How many introns are there in human?

How many exon-exon junctions are there in human RNA samples we’ve studied?

not-too-rare, naturally occurring

How many ▲ exon-exon junctions are there in human RNA samples we’ve studied?

How many exon-exon junctions are in at least *K* publicly available human RNA-seq samples on SRA?

How many exon-exon junctions are in
at least *K* publicly available
Illumina
human ▲ RNA-seq samples on SRA?

What gene annotation says

For *hg19*,

Ensembl v75 $\cup$ GENCODE v19 $\cup$ RefSeq

(almost subsumed
by Ensembl v75)

≈350,000

junctions

SRA classification basics (raw data)

project ([SED]RP\d+) : study; “unit of research”

experiment ([SED]RX\d+) : library, platform, processing parameters

sample ([SED]RS\d+) : physical sample

run ([SED]RR\d+) : one or pair of FASTQs

SRA classification basics (raw data)

We refer to this.

project ([SED]RP\d+) : study; “unit of research”

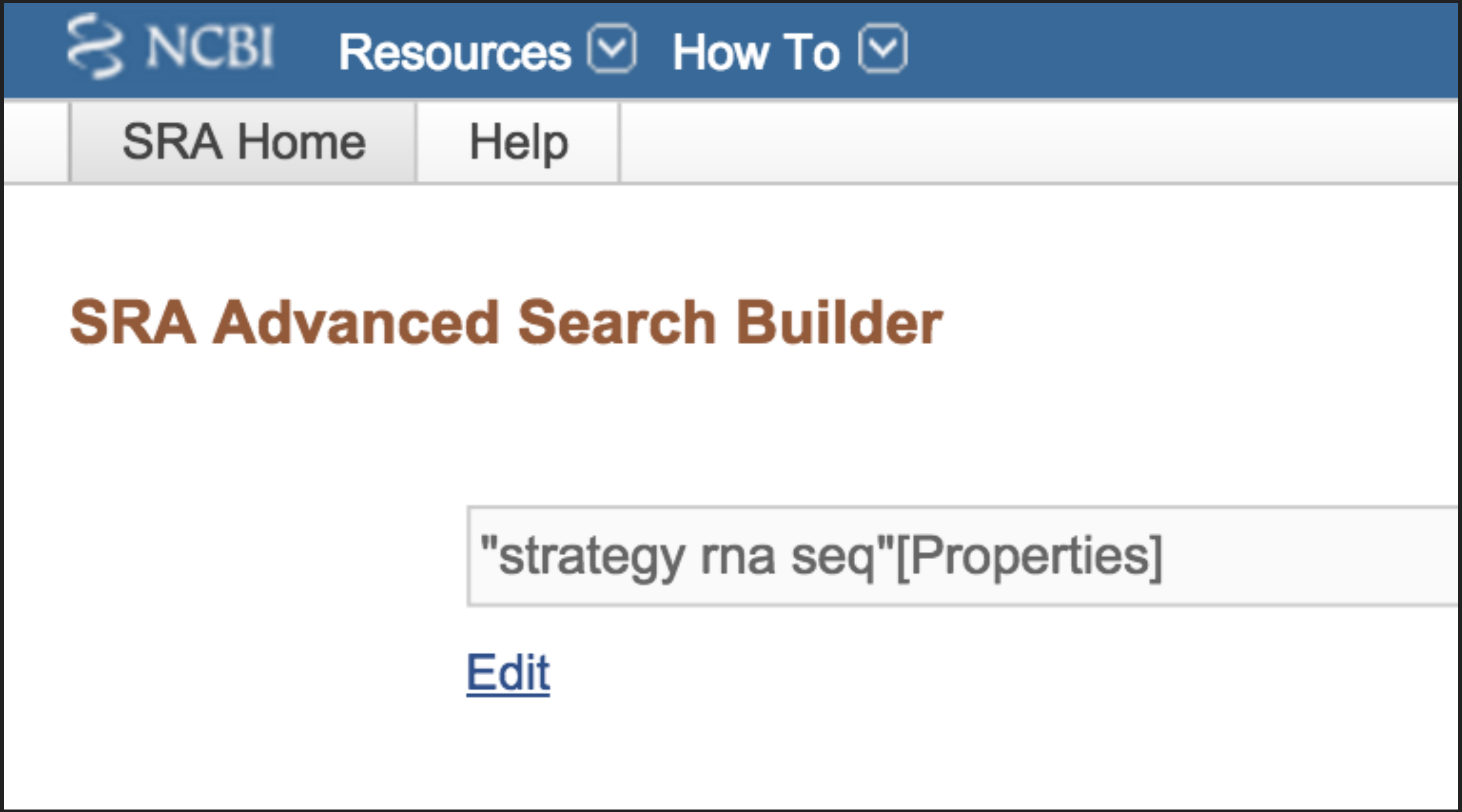
experiment ([SED]RX\d+) : library, platform,
processing
parameters

sample ([SED]RS\d+) : physical sample

run ([SED]RR\d+) : one or pair of FASTQs

We refer to this.

Filtering SRA



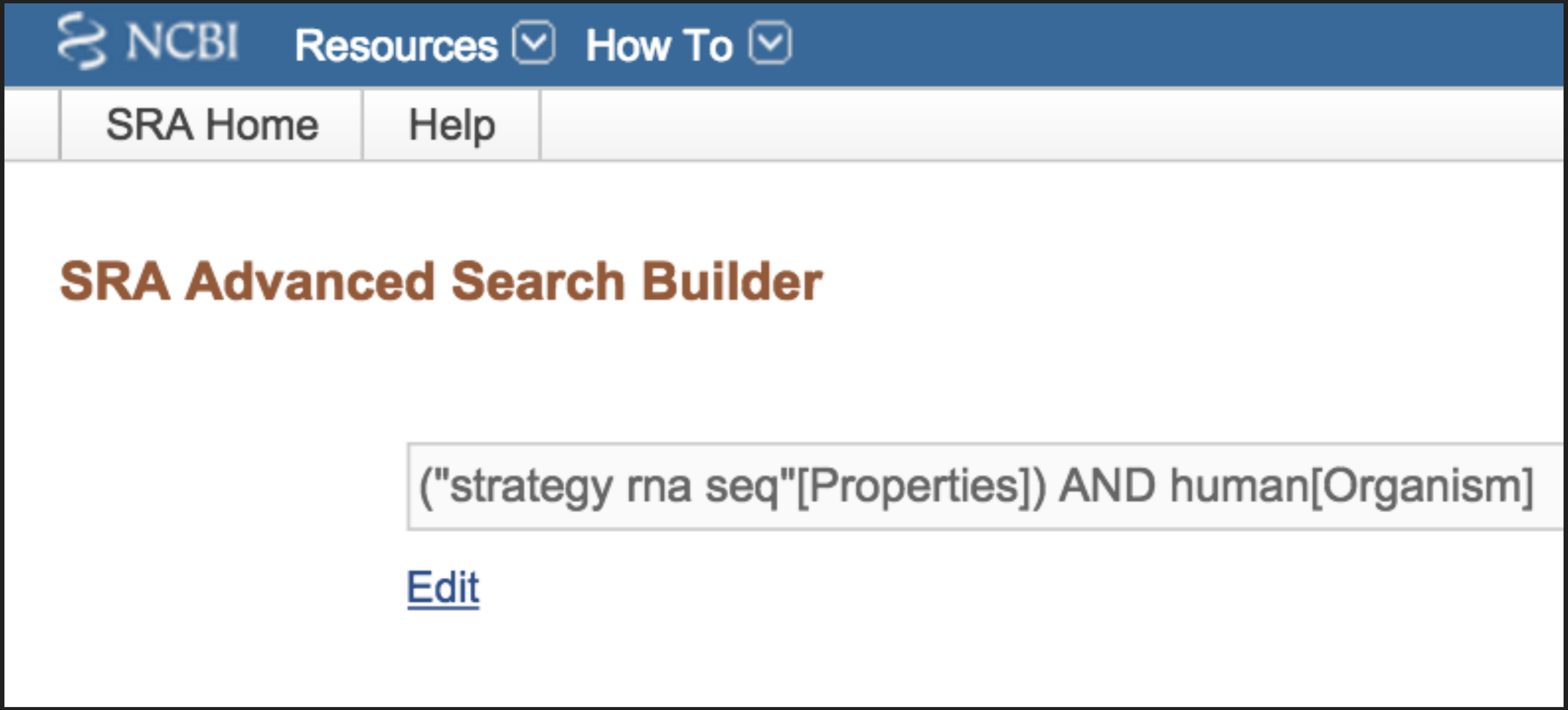
SRA Advanced Search Builder

"strategy rna seq"[Properties]

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

≈180k publicly available runs

Filtering SRA



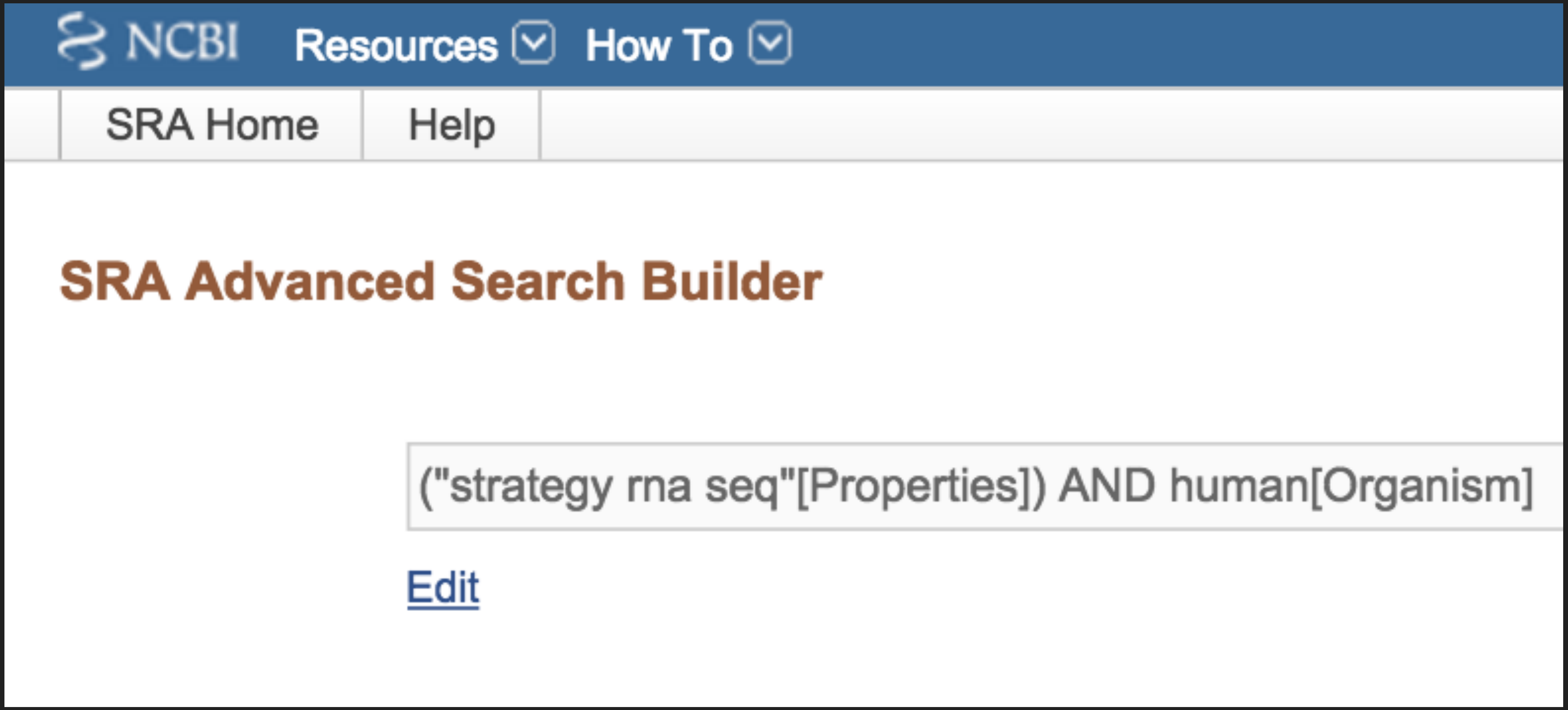
SRA Advanced Search Builder

("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 web services™

= freedom

from

downloading hassles

cluster administrators

competition for compute

irrecoverable results

bioinformaticians

Download Rail-RNA at

http://rail.bio.

#FreedomThroughTheCloud #gi2015

Read the preprint at http://j.mp/rail-pre .

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

to download/preprocess data, copy to S3

We ran

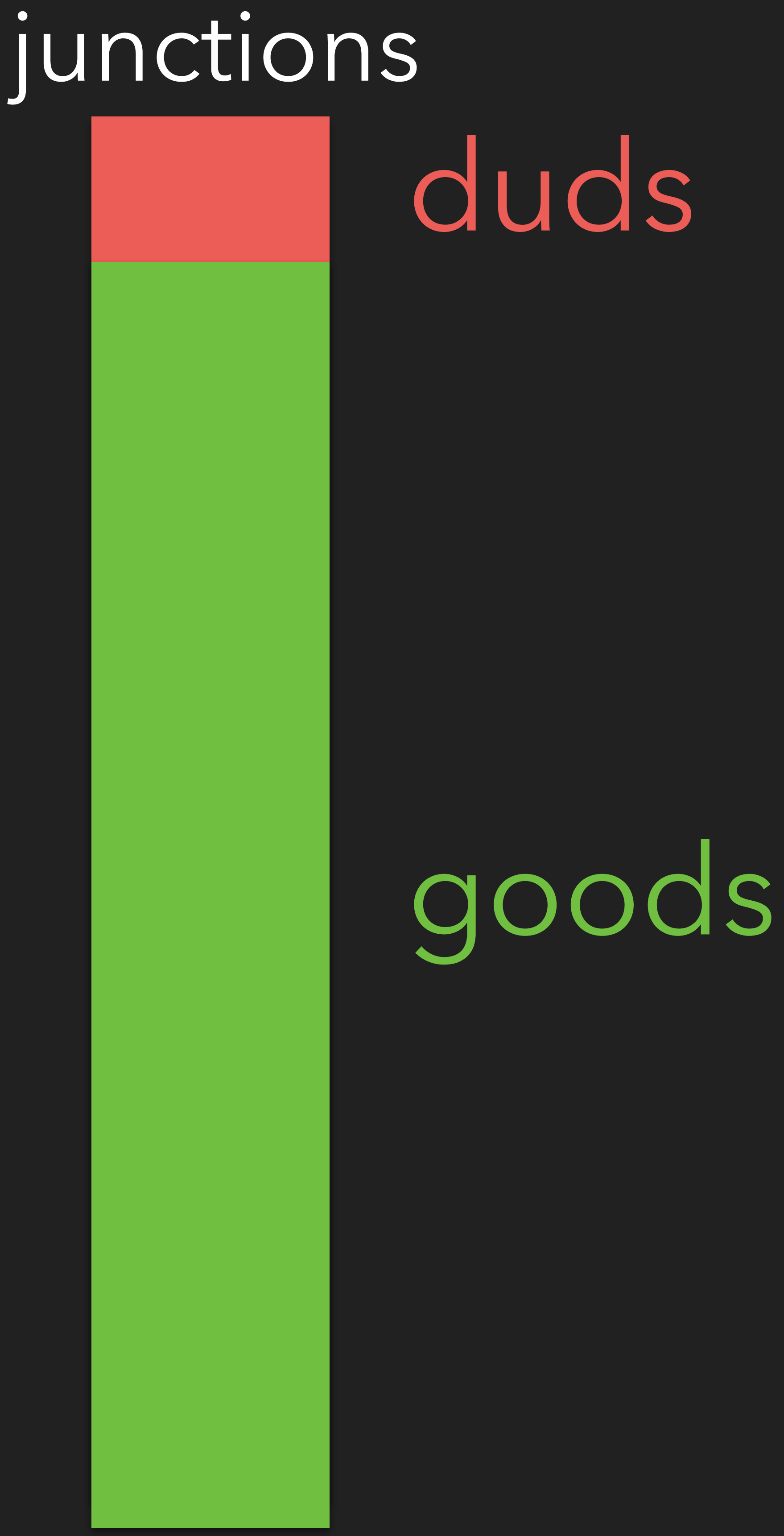
```
rail-rna align elastic
—-manifest batch_X.tsv
—-input s3://bucket/batch_X_prepped
—-output s3://bucket/batch_X_itn
--core-instance-bid-price 0.60
--master-instance-bid-price 0.60
—-core-instance-count 60
—-core-instance-type c3.8xlarge
—-master-instance-type c3.8xlarge
—-deliverables itn
```

to detect junctions from one pass of alignment

2 commands x 43 batches gave, after merging

- One 7-GB `tsv.gz`
- 42,882,032 junctions
- number of reads in which each junction was detected after 1-pass alignment in each sample

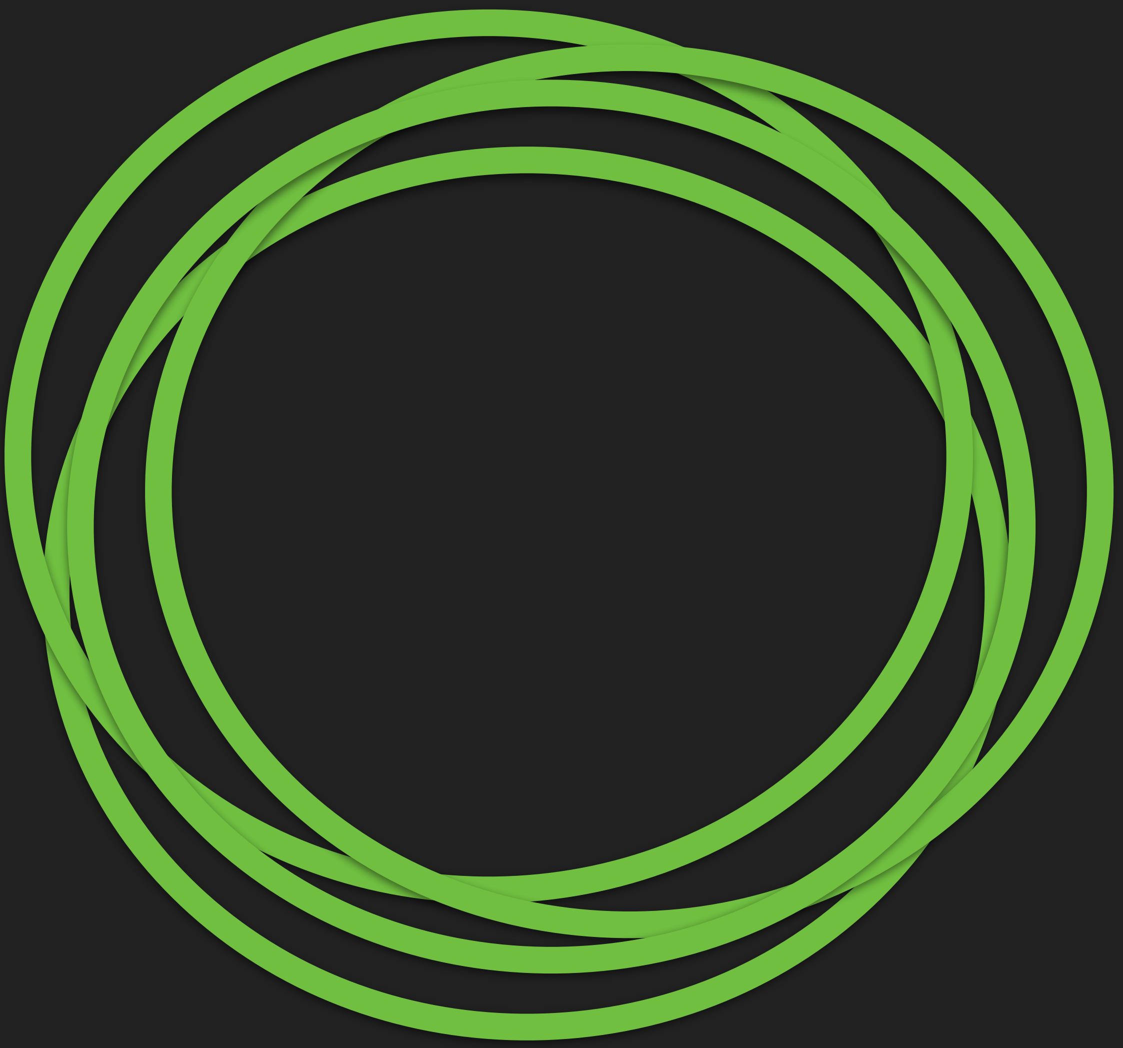
Why so many junctions?



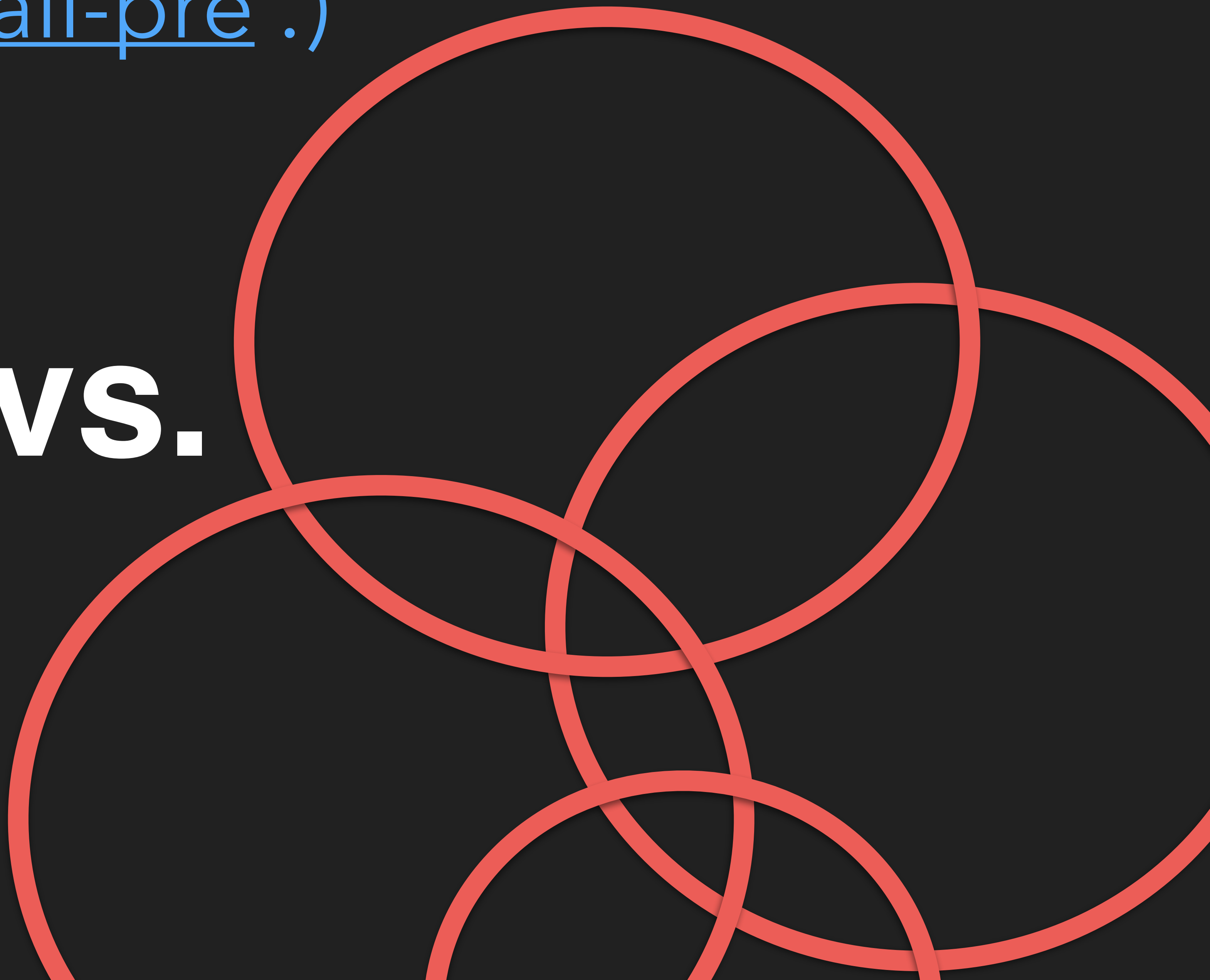
On a single sample, *every* aligner will find some good junctions and some duds (or very rare junctions).

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 .)

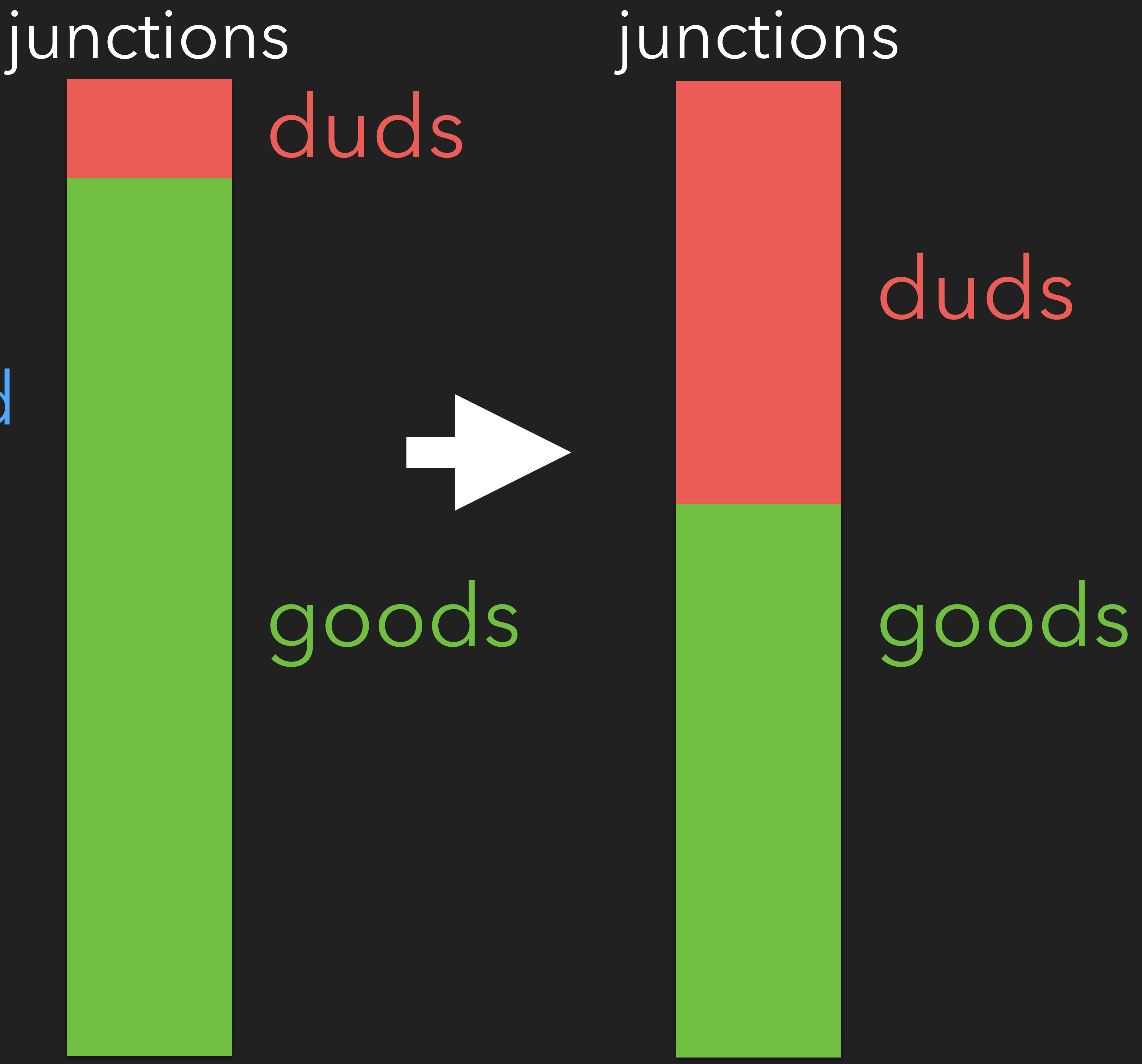


vs.

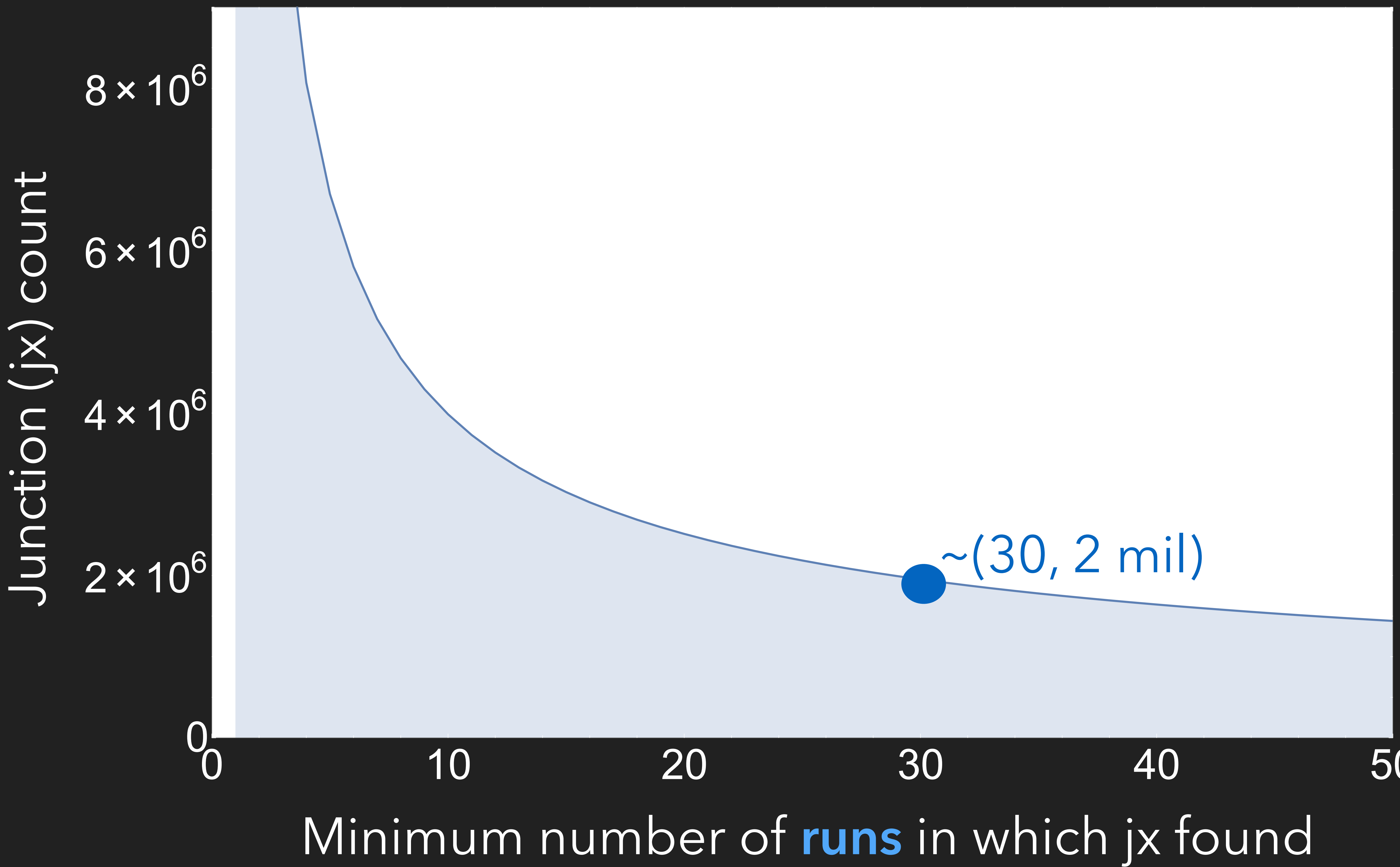


Why so many junctions?

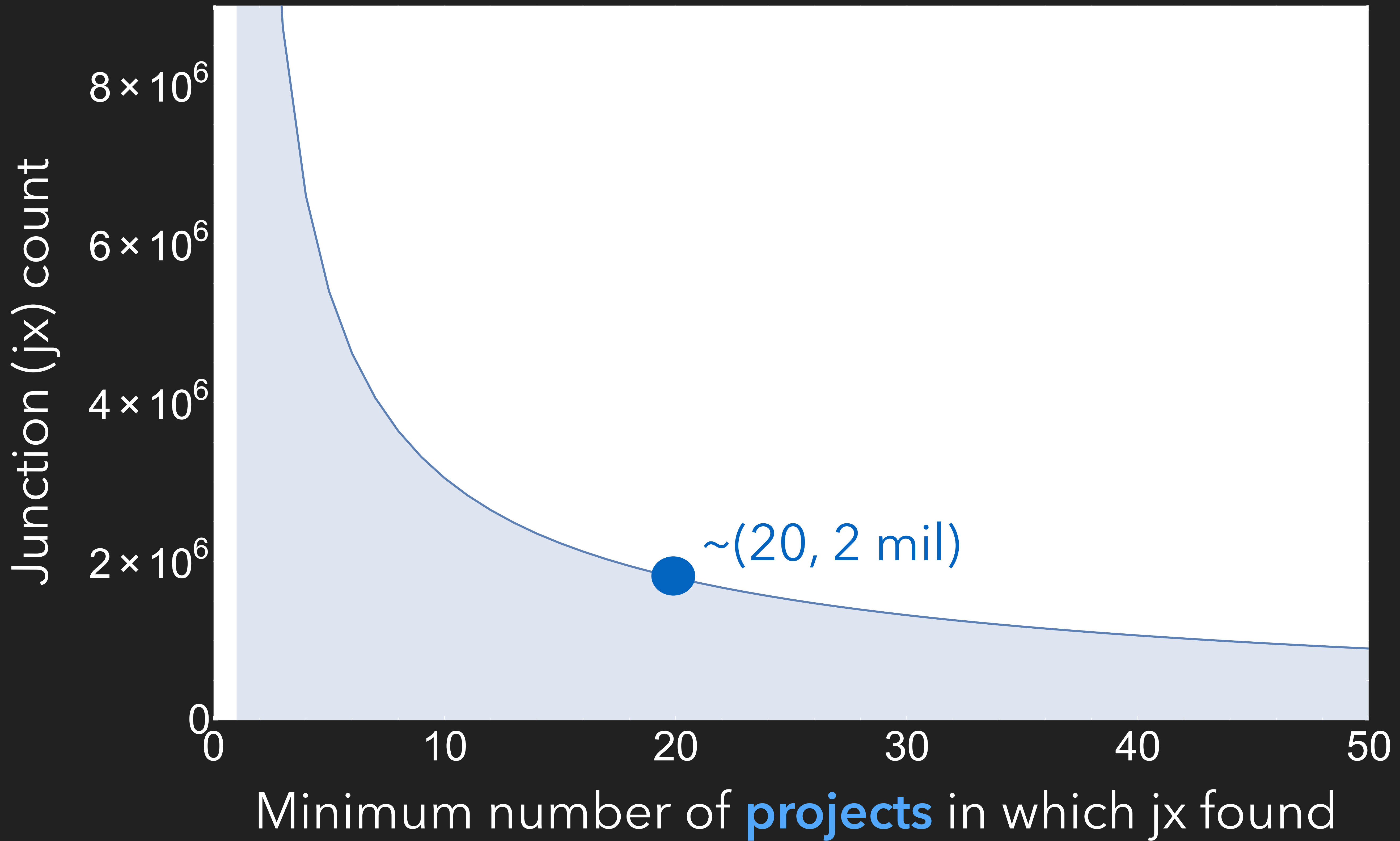
So as you add samples...



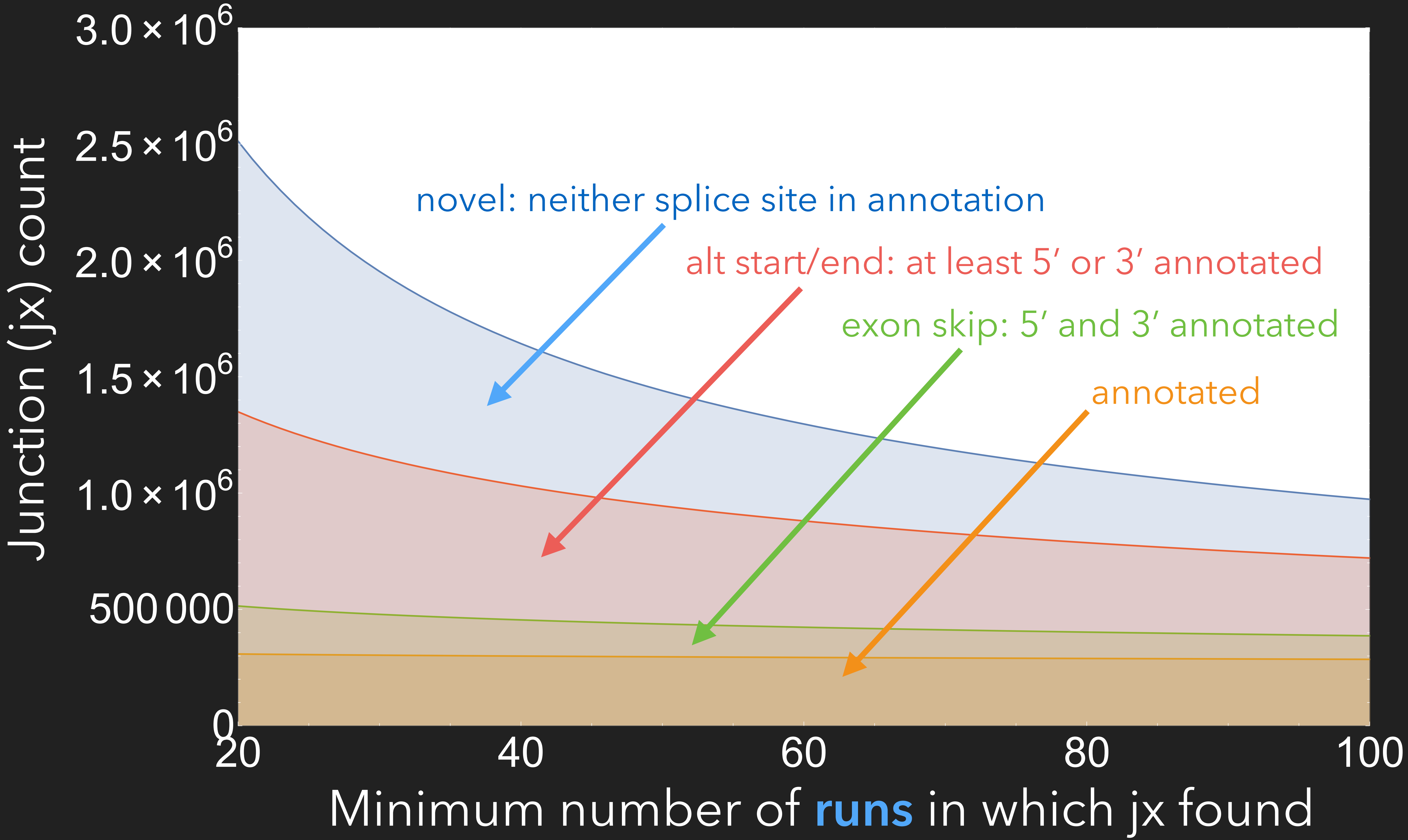
A steep dropoff: run-level



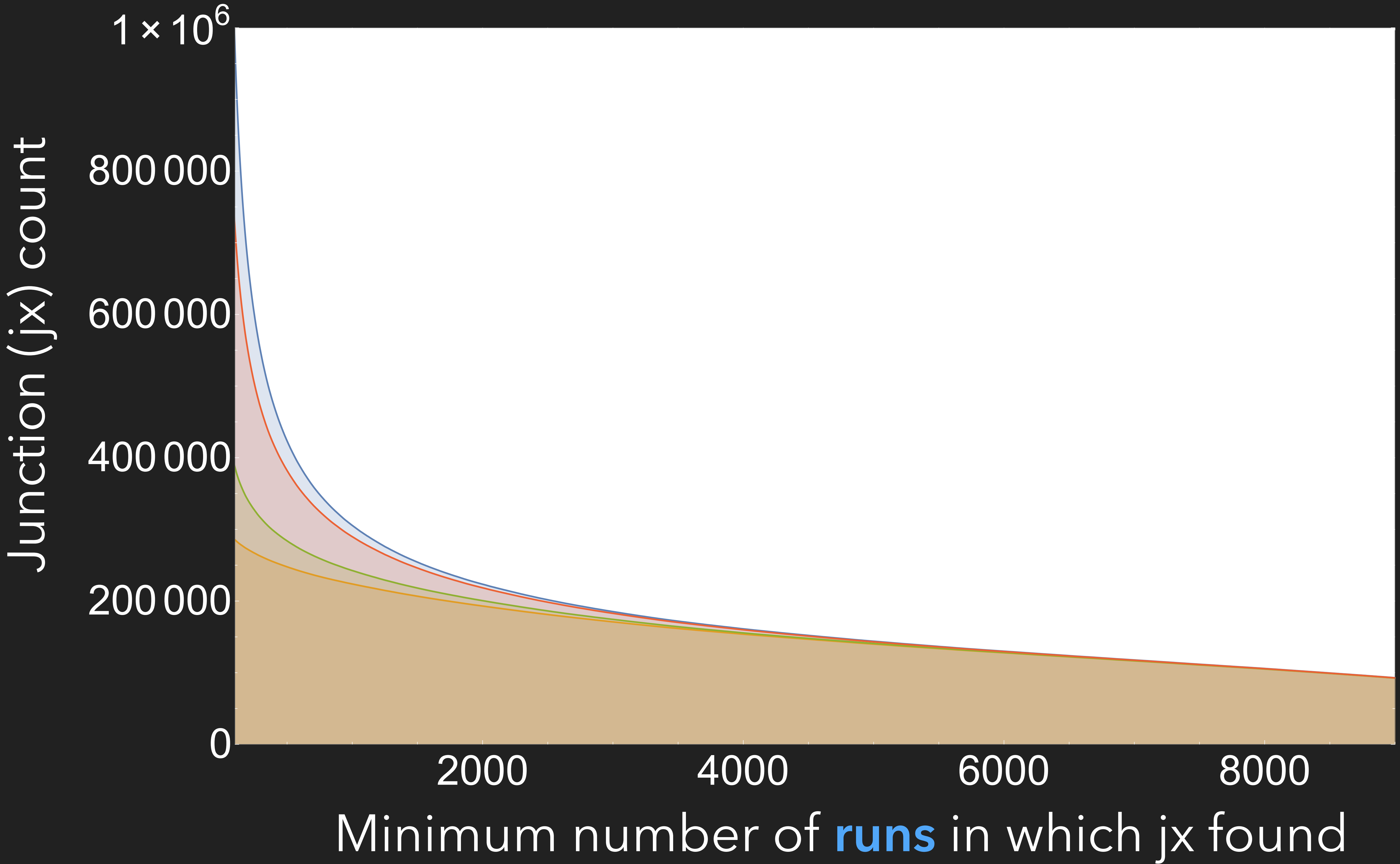
A steep dropoff: project-level



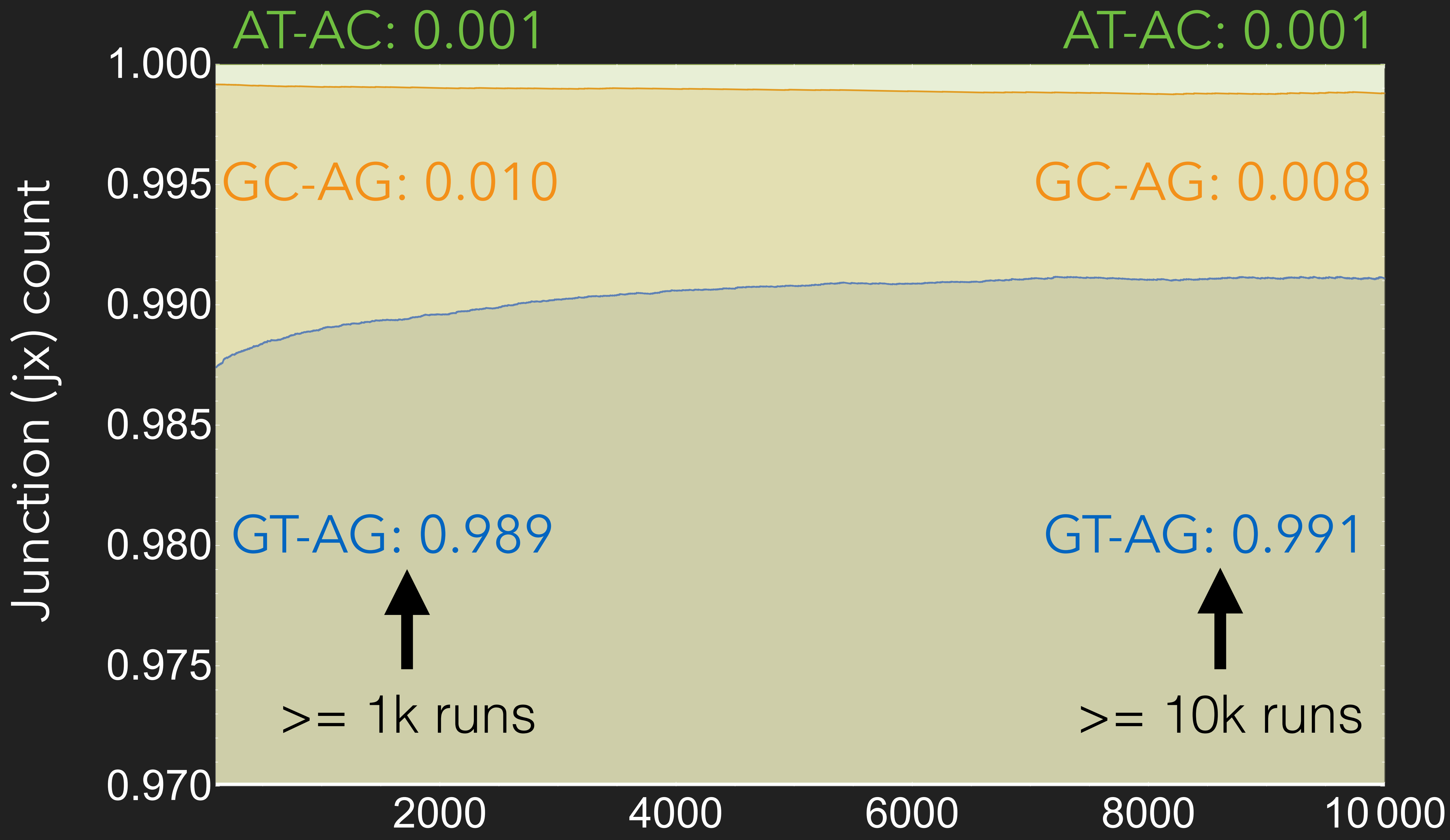
Levels of evidence in annotation



Asymptote to annotation

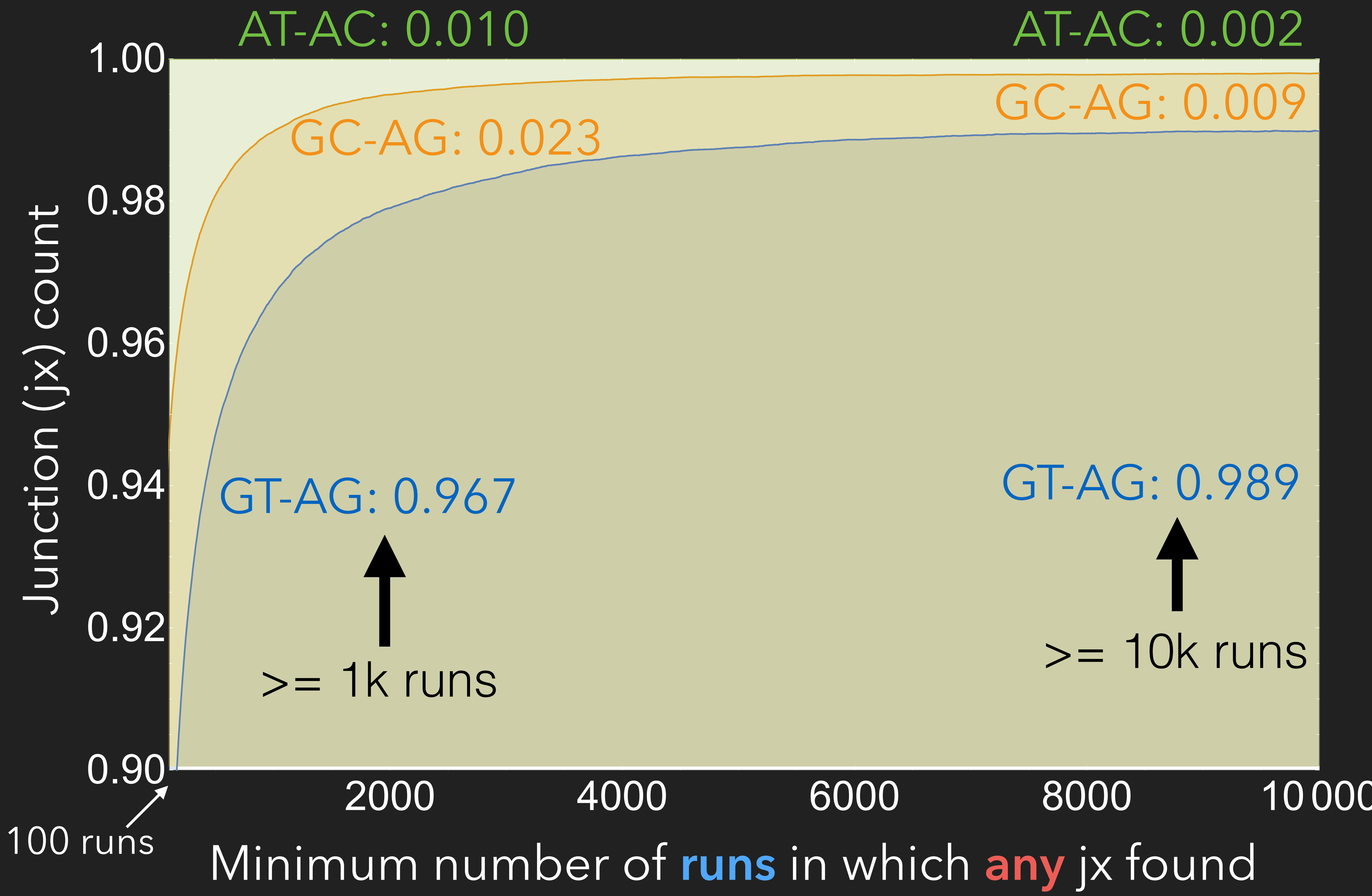


Motif ratios for annotated junctions

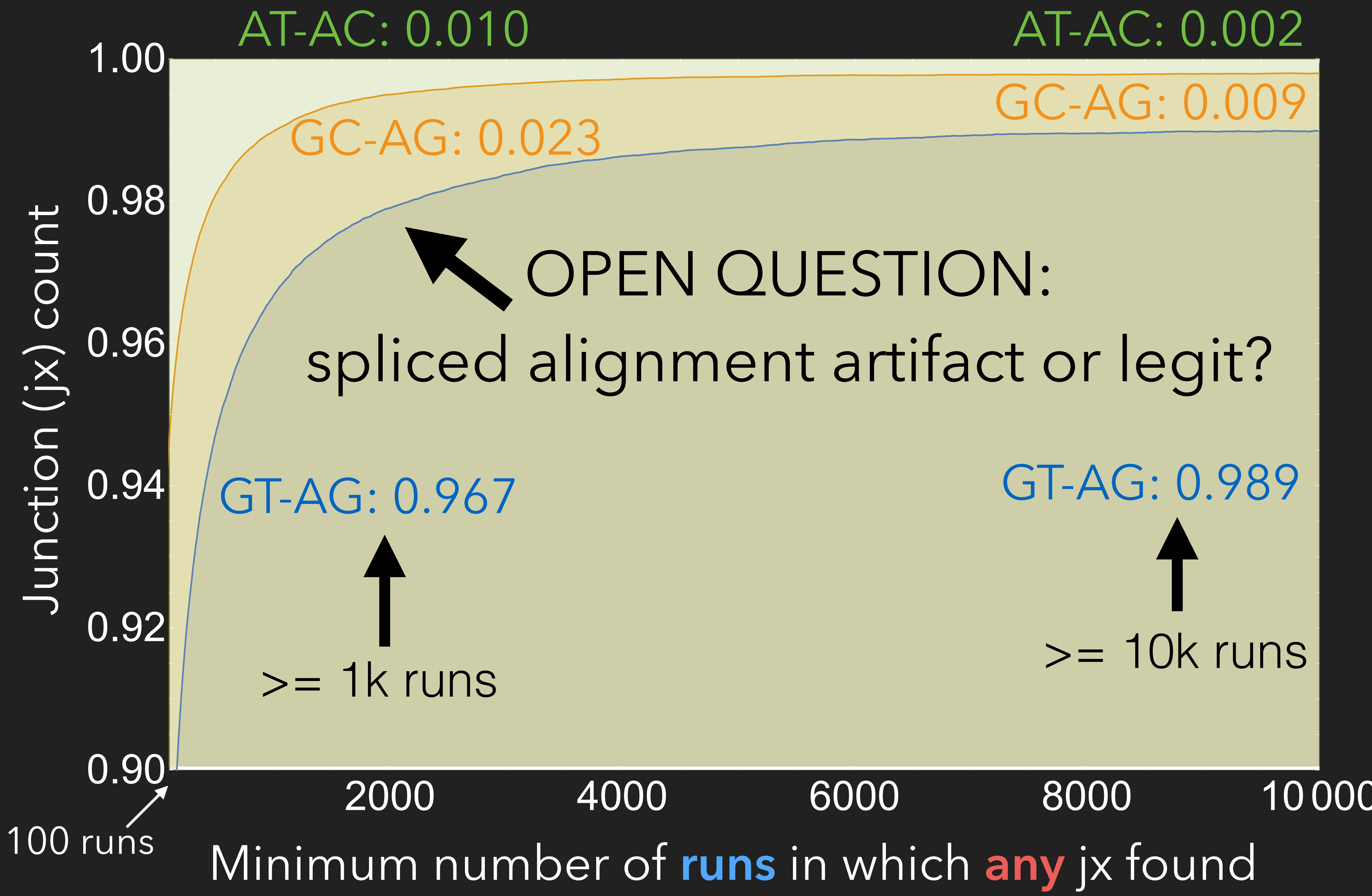


Minimum number of **runs** in which **annotated** jx found

Motif ratios for all junctions



Motif ratios for all junctions



AT-AC: 0.010

AT-AC: 0.002

GC-AG: 0.023

GC-AG: 0.009

GT-AG: 0.967

GT-AG: 0.989

OPEN QUESTION:

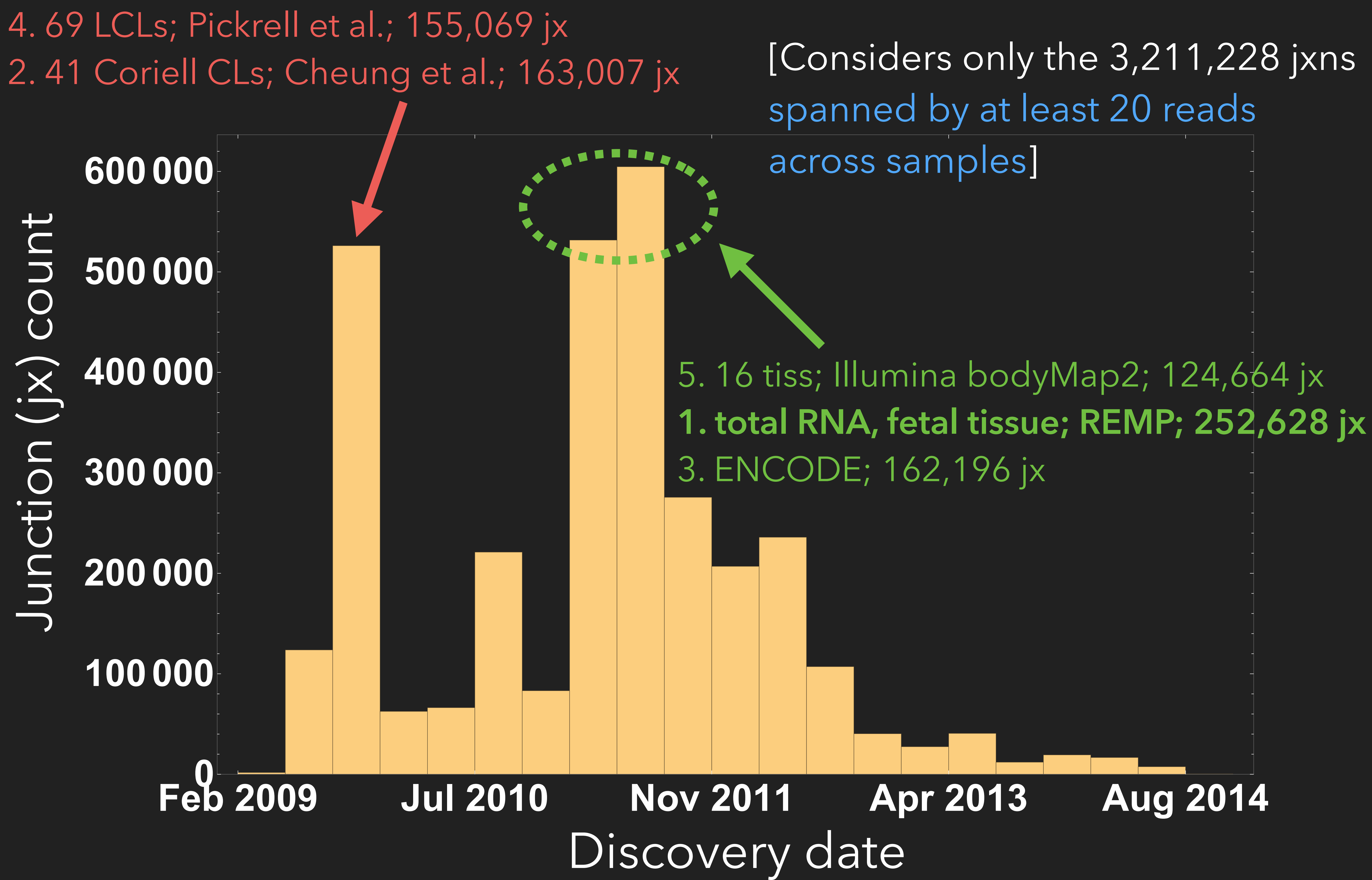
spliced alignment artifact or legit?

>= 1k runs

>= 10k runs

100 runs

Are we still finding new junctions?



How many exon-exon junctions are there in human RNA that we'll care about?

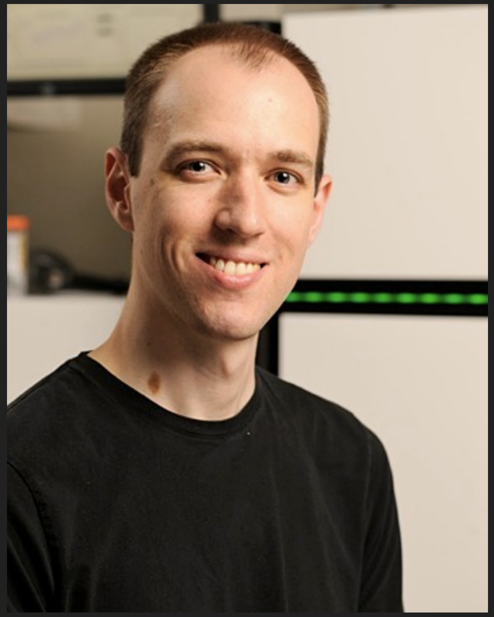
Between 1 and 5 million. Probably.

Junction list unreleased, but processed data for generating these results and more available at

http://github.com/nellore/gi2015

Collaborators

Jeff Leek



Ben Langmead

Summer interns

Nishika Karbhari
Robert Phillips
Sara Wang

Leo
Collado-Torres

Andrew
Jaffe

Chris Wilks

José
Alquicira Hernández