OPEN ACCESS Freely available online PLoS GENETICS

Noisy Splicing Drives mRNA Isoform Diversity in Human Cells

Joseph K. Pickrell[1]*, Athma A. Pai[1]*, Yoav Gilad[1]*, Jonathan K. Pritchard[1,2]*

1 Department of Human Genetics, The University of Chicago, Chicago, Illinois, United States of America, 2 Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois, United States of America

Journal Club
February 16, 2011
Gerstein Lab

Lukas Habegger

Introduction

- Majority of multi-exonic human genes show evidence of alternative splicing
- Unclear what fraction of splice forms are functionally relevant
- Two alternatives:
  - Most genes have functionally relevant isoforms
  - Many transcripts are non-functional noise

Evidence of noisy splicing

- Large fraction of exon-skipping events in human genes is not observed in mice
  - Not conserved
- Number of transcript isoforms correlates with the number of exons
- Short introns in humans have evolved to preferentially trigger degradation via nonsense-mediated decay (NMD)
  - Evidence that errors are common enough to exert detectable selective pressure

Evidence of noisy splicing

- Splicing factors play an important role in exon recognition
- Binding sites for such splicing factors comprise a large mutational target
  - Large size of introns compared to exons provides ample opportunity for mutations
- Although mutations that create or disrupt binding sites are slightly deleterious, the large number of possible mutations makes it inevitable that some will be fixed in the population
- Hypothesis: the human genome carries a substantial load of suboptimal sequences giving rise to aberrant transcript isoforms

Data set

- 75 lymphoblastoid cell lines derived from Nigerian individuals as part of the International HapMap Project (Pickrell et al., *Nature*, 2010)
- 1.4 billion sequencing reads (35 or 46nt)
- 1.2 billion reads mapped to the human reference genome sequence

Splice junction reads

- The remaining reads were used for splice junction identification
  - *de-novo* approach
  - Split sequence read into two segments and map each end separately

- 48 million reads mapped to 392,612 putative splice junctions

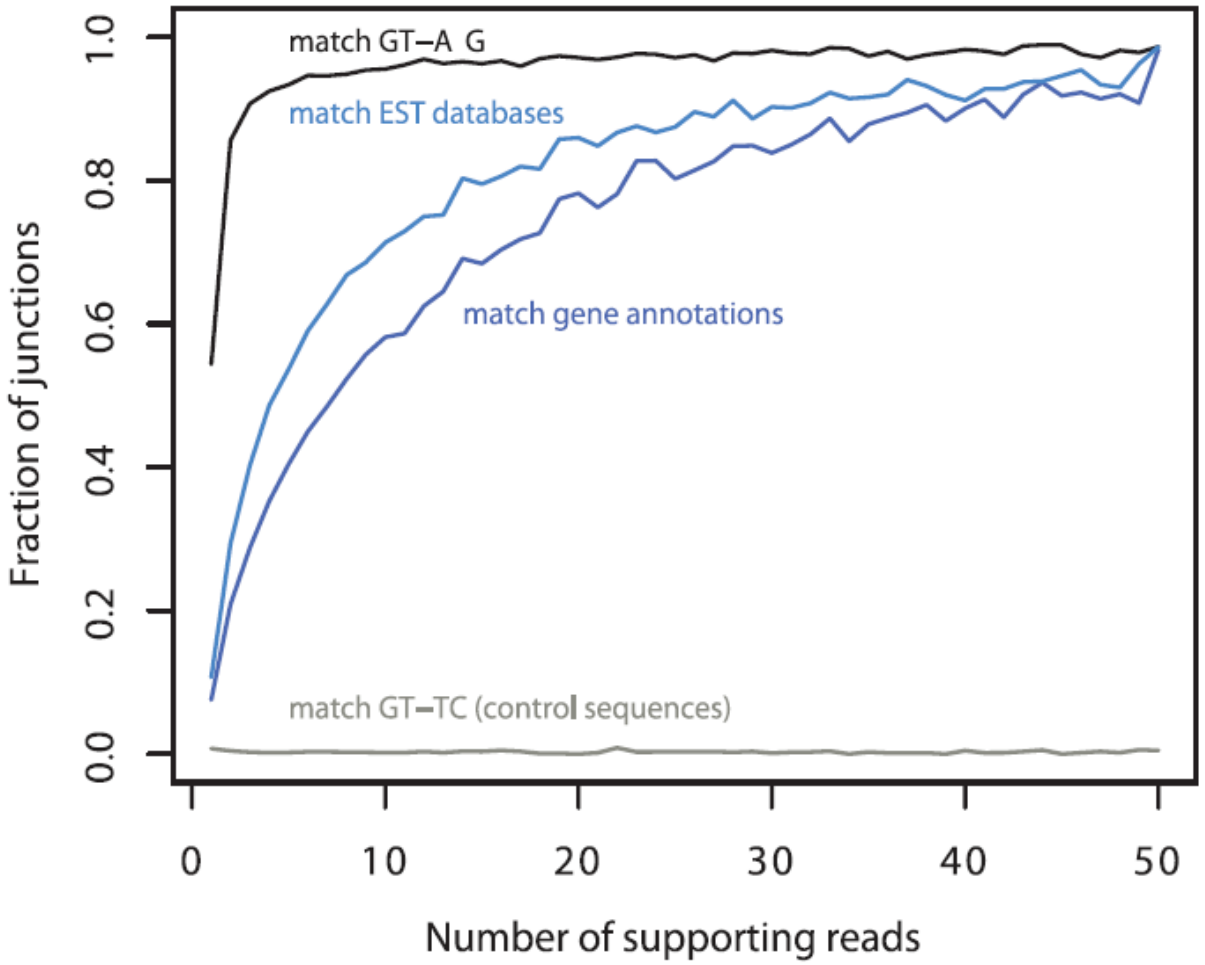
Assessment of splice junctions

- Splice junction identification method does not rely on a gene annotation set
  - Concern: identification of spurious junctions due to mapping artifacts

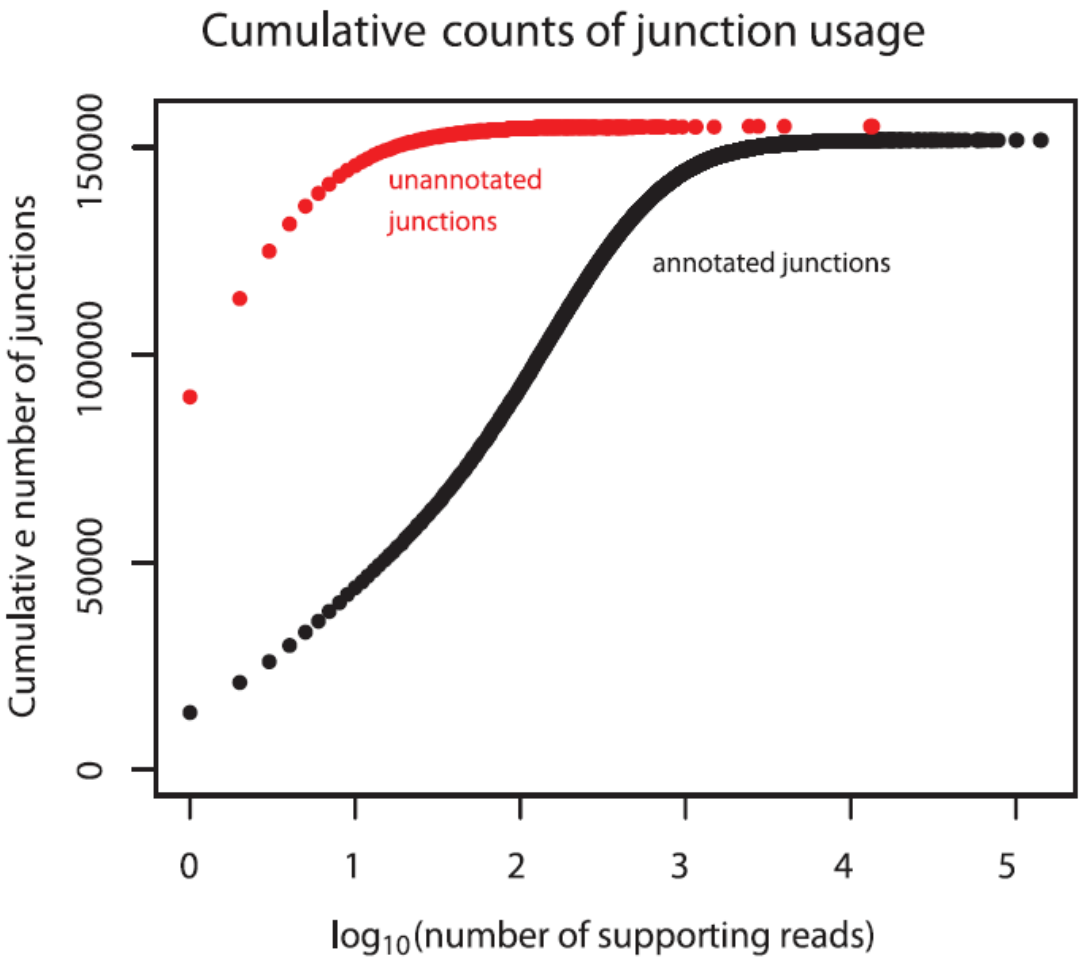
- The majority contain the canonical GT-AG splice junction

- Estimated false discovery rate (FDR): 1.5% for **306,606** splice junctions (GT-AG or GC-AG)

Properties of splice junctions



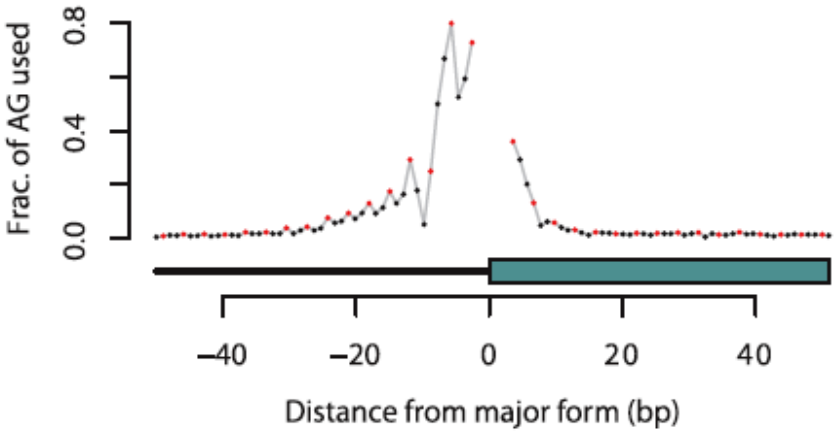
Unannotated junctions are expressed at much lower levels than annotated junctions

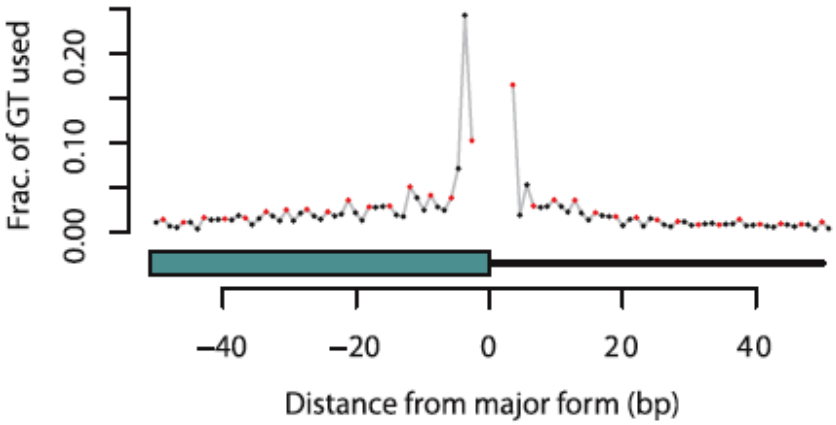


Alternative splice junctions near protein-coding junctions

- Select splice sites annotated as protein-coding
- For 3’ splice sites:
  - identify all 5’ splice sites that are
    - covered by at least 20 reads
    - spliced to at least two 3’ splice sites
    - “major” 3’ splice site captures more than 80% of the reads
- For each position, count the the AG dinucleotides used as alternative splice sites

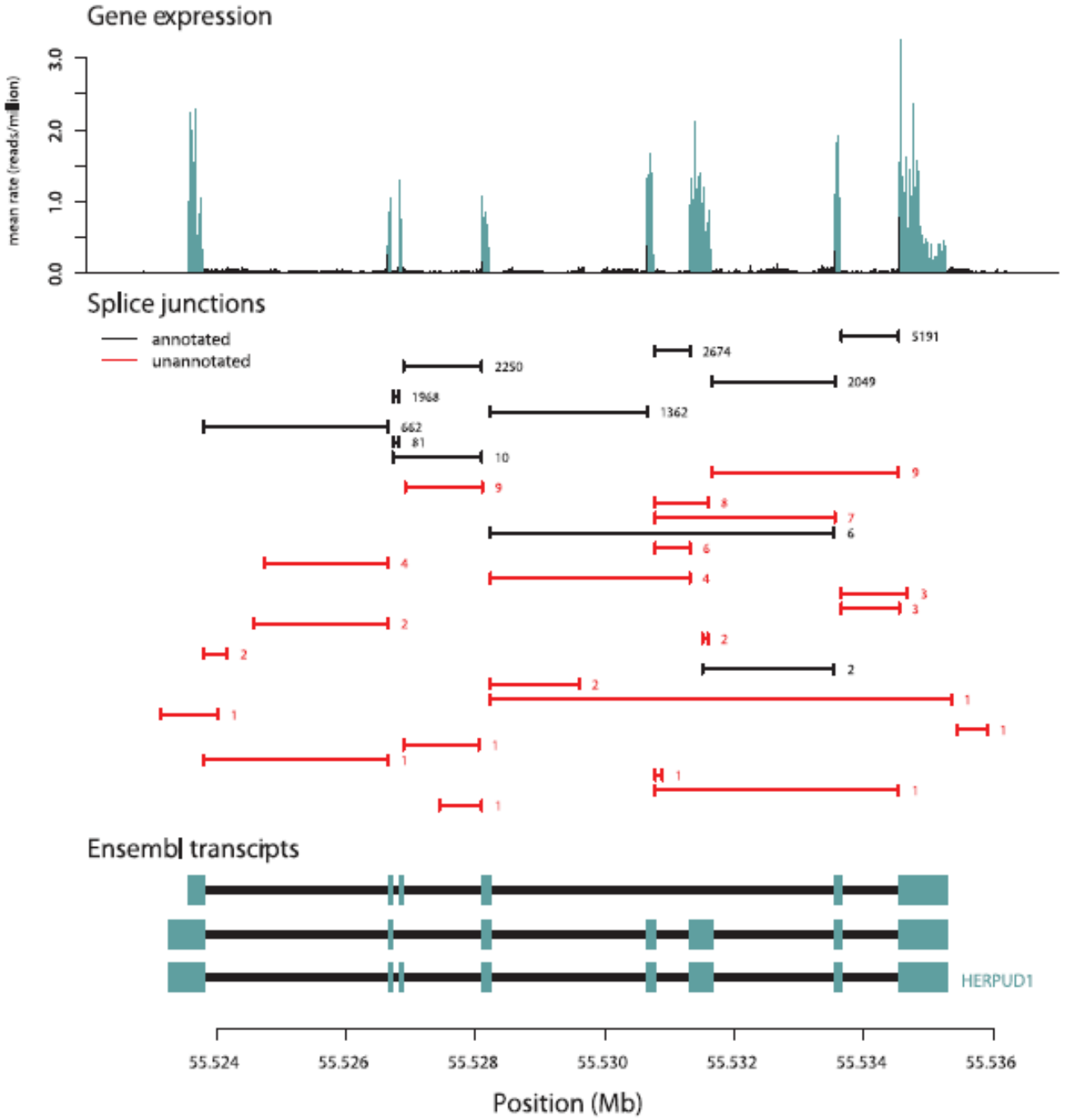
- The red points denote positions that are a multiple of three base pairs from the major splice form
- The black points denote those that are not
- **Conclusion: Nearby splice sites that maintain the coding frame are observed more often**





Many identified junctions are novel

- Of the 306,606 junctions
  - Approximately 50% are not part of known gene models (UCSC, Ensembl, Vega, RefSeq)
  - The subset of unannotated junctions accounts only for 1.7% of all junction-spanning sequencing reads

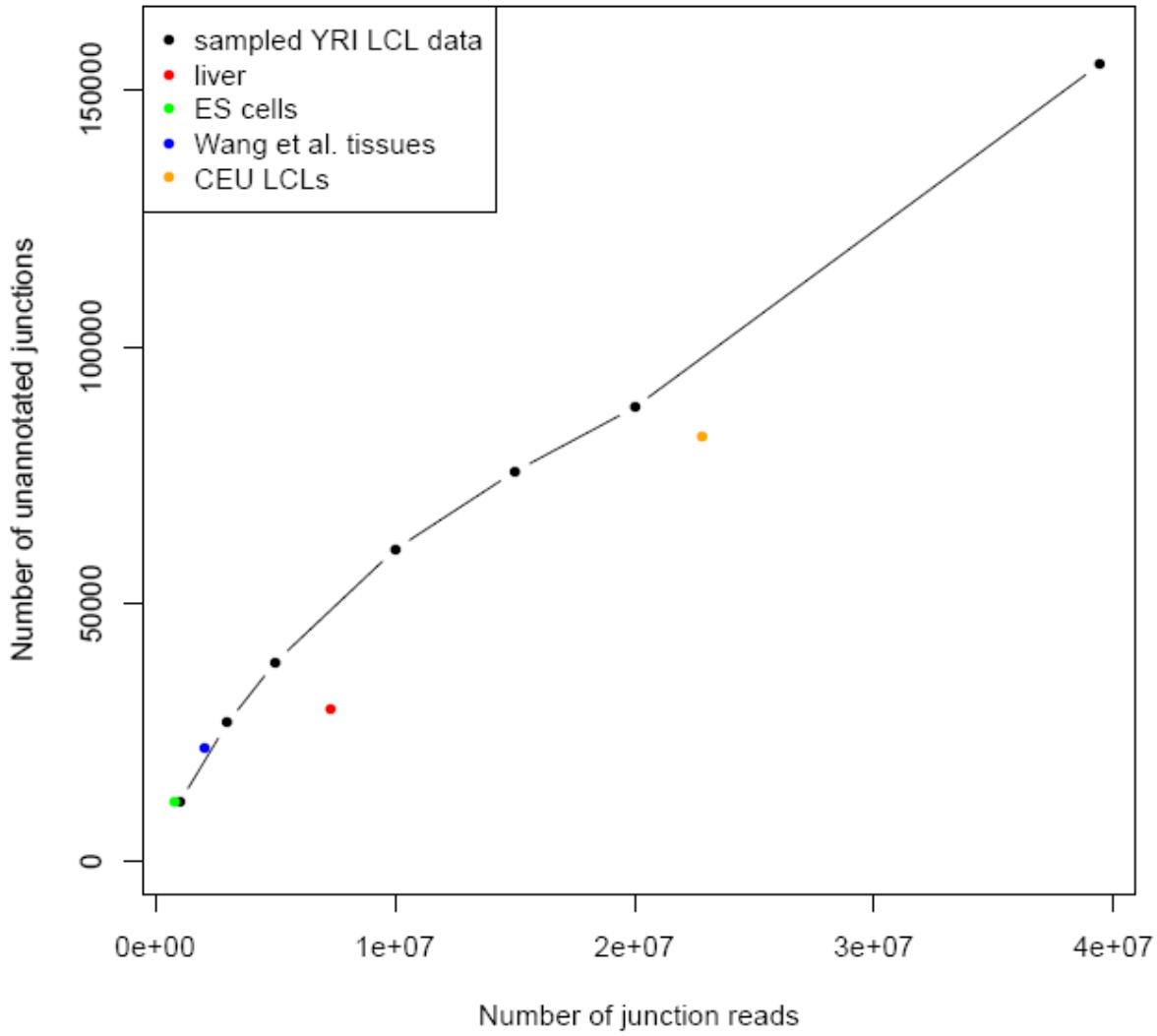


Example: HERPUP1

21 of 32 splice junction are unannotated

Only 0.5% of the reads from this gene map to the unannotated junctions

Identification of isoforms is not at saturation



Characteristics of identified junctions and splice sites

	Both ends known				
	Known junc.	**New junc.**	**New 3′ SS**	**New 5′ SS**	**Both new**
number	151,679	27,611	50,839	43,447	33,030
mean coverage	255	5	4	5	3
% obs. in other tissues	87	27	23	24	16
% near known [5′,3′]	−,−	−,−	−,33	23,−	5,8
% highly conserved [5′,3′]	76, 76	75, 77	73,10	10,72	7,6

As described in the main text, we split the observed junctions into five classes based on gene model databases. For each class, we present the number of such junctions, the average number of reads spanning each junction in that class, the percentage of the junctions observed in any tissue assayed in Wang et al. [2], the percentage of 5′ and 3′ splice sites of each junction that fall near an annotated splice site ("near" here is defined as within 50 base pairs), and the percentage of the 5′ and 3′ splice sites of each junction that show strong evidence of evolutionary conservation (defined as a mean *phyloP* score >2 [31] at the two canonical bases of the splice site).
doi:10.1371/journal.pgen.1001236.t001

Extensive unannotated splicing is present in different human populations and tissues

- Performed same analysis for different RNA-Seq data sets
  - European lymboblastoid cell lines (Montgomery et al., *Nature*, 2010)
  - Human liver samples
  - Human body map (Wang et el, *Nature*, 2008)

- Extent of unannotated splicing is comparable and broadly generalizable

- Little evidence of individual-specific splicing

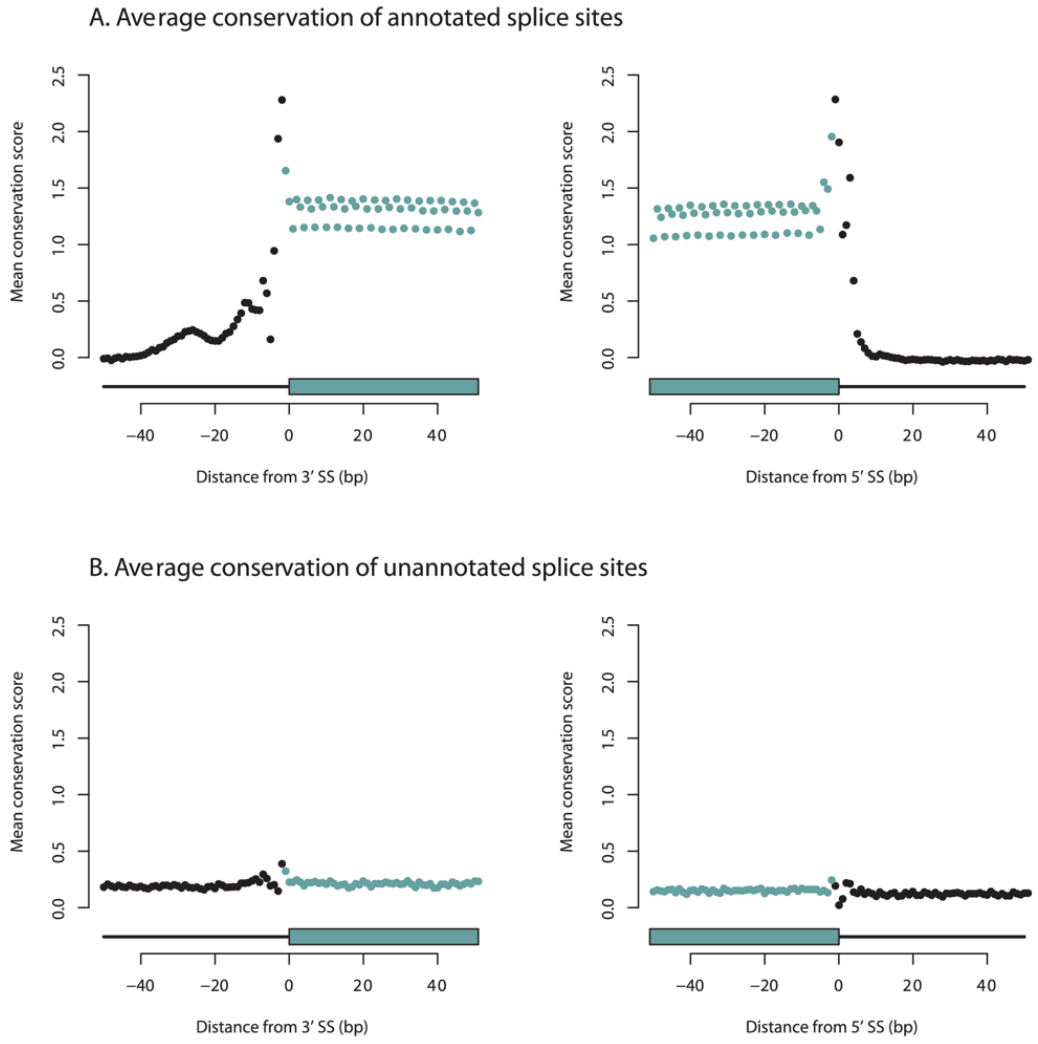
Most unannotated splice junctions show no evolutionary conservation

- If unannotated splice junctions are functionally relevant, then their sequence conservation should be comparable to that of annotated splice junctions

- Compared sequence conservation across mammals using the *phyloP* score

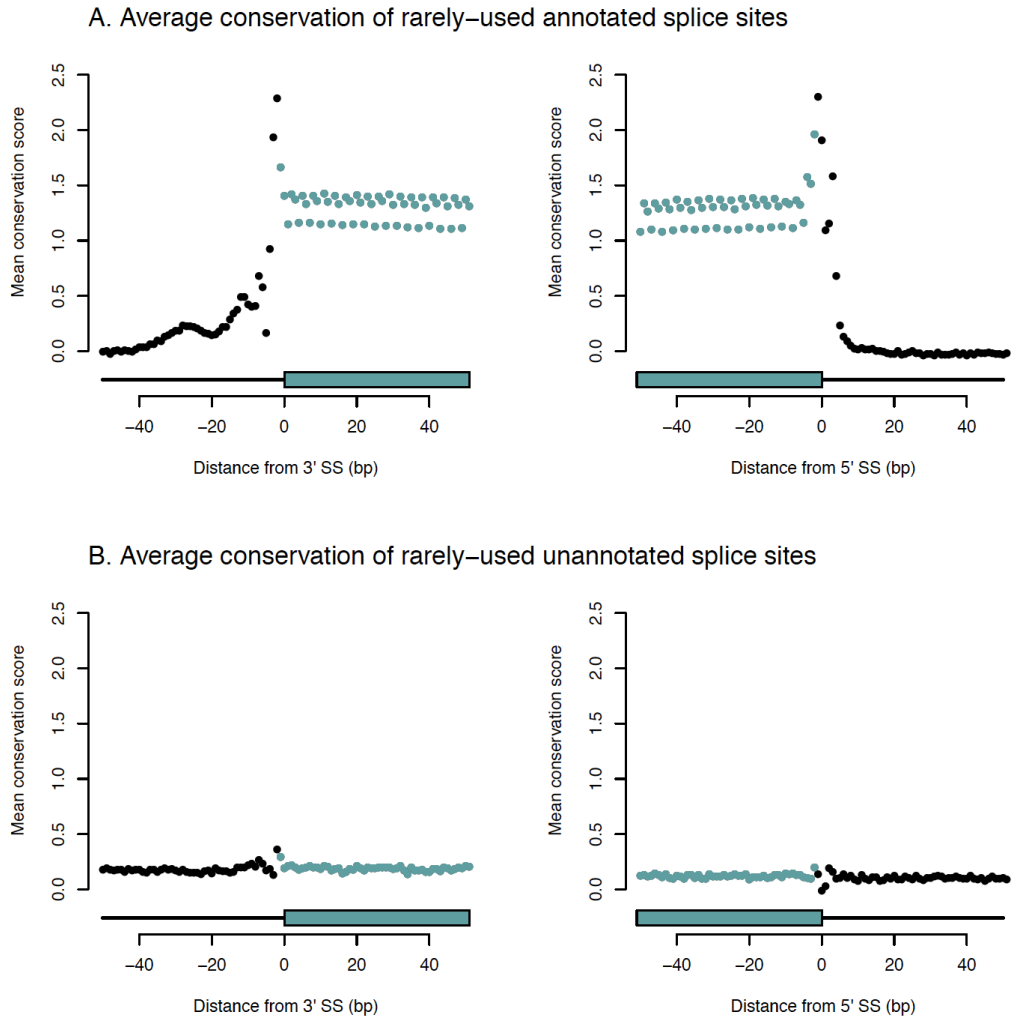
- Similar analysis was performed using *phyloP* scores from primates only
  - Rationale: exclude the possibility that splice junctions are conserved in only a subset of mammals

Unannotated splice sites show little evidence of evolutionary conservation

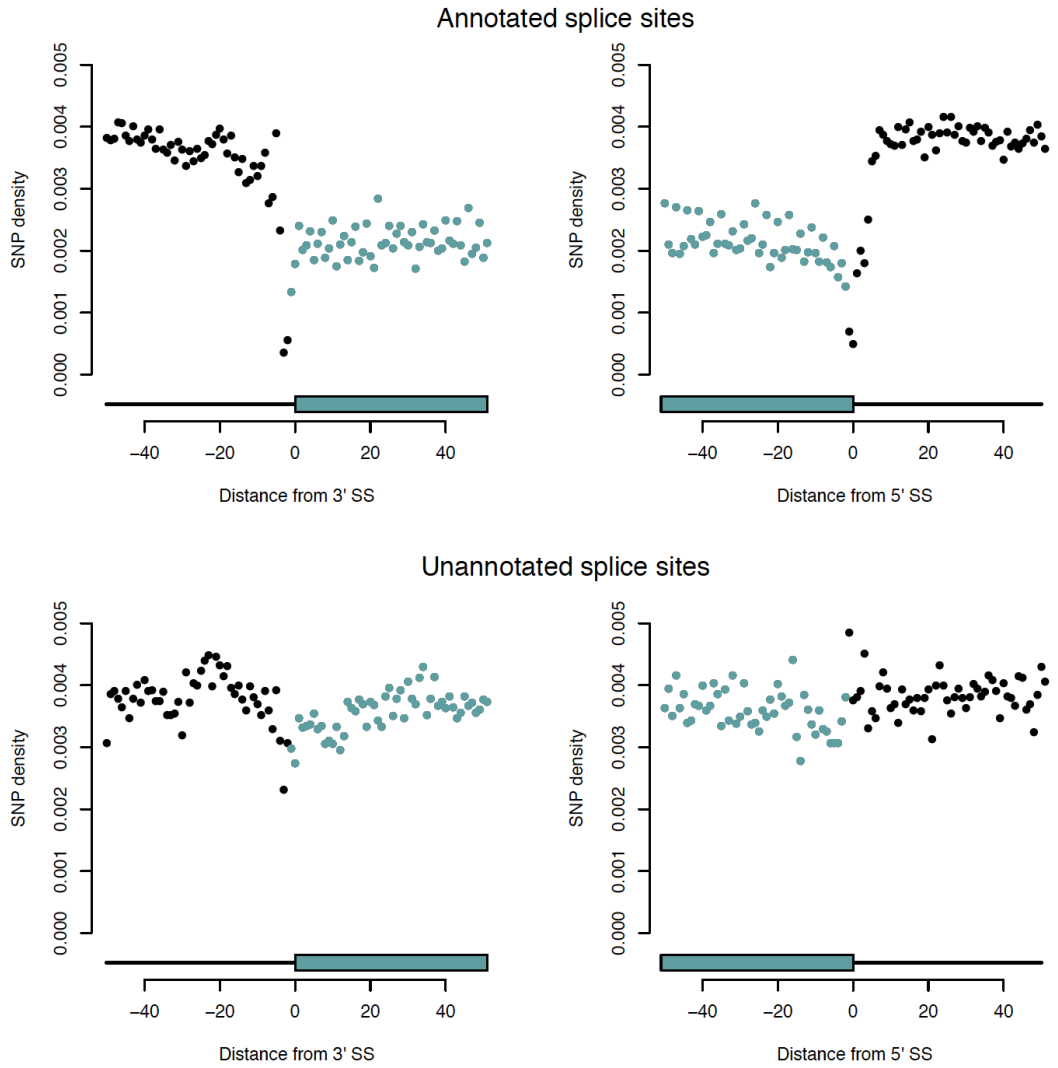


Similar results were obtained using the other RNA-Seq samples or using the *phyloP* scores from primates only

Rarely-used but annotated splice sites are highly conserved



Unannotated splice sites show no reduction in polymorphisms in splice sites

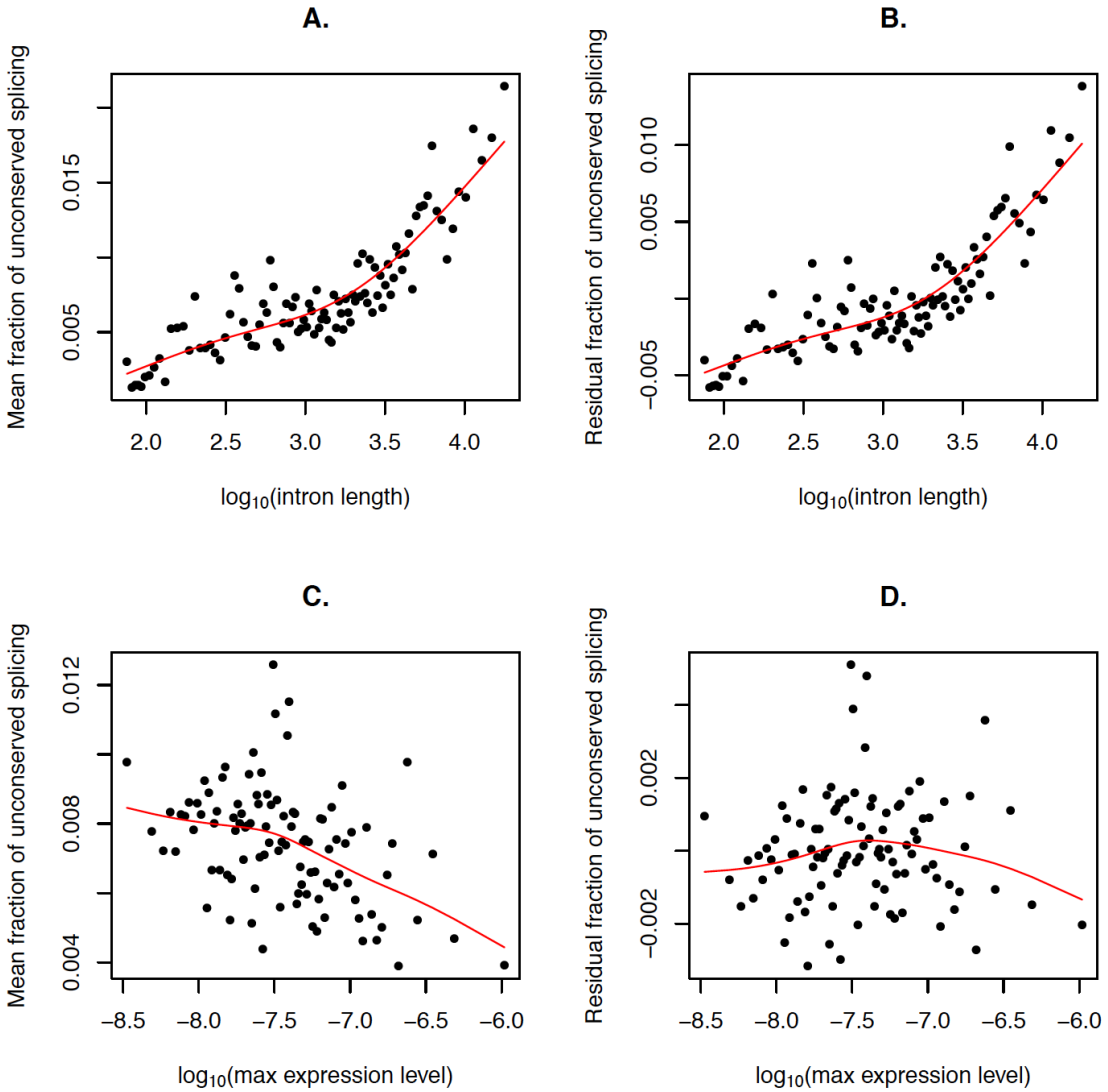


Unannotated and rarely-used splice sites originate from mis-spliced transcripts

Estimation of splicing error rates

- Identified a set of splice junctions where both splice sites where highly conserved (*phyloP* scores > 2)
- Identified a set of splice junctions where both splice sites where not conserved (*phyloP* scores < 0.5)
- How often are conserved and unconserved splice sites spliced togher?
  - Result: 0.7% of the reads
- Median number of exons for human genes: 4
- Conclusion: Approximately 2% of the transcripts from the average human gene are mis-spliced
  - Conservative estimate: mis-spliced transcripts are preferentially removed by NMD

Small introns have a lower splicing error rate



Introns are divided into 100nt bins

For each bin, calculate mean fraction of sequencing reads from either splice site to an unconserved splice site

Legend:

A: Conservation vs. intron length

B: Conservation vs. intron length after correcting for gene expression

C: Conservation vs. gene expression

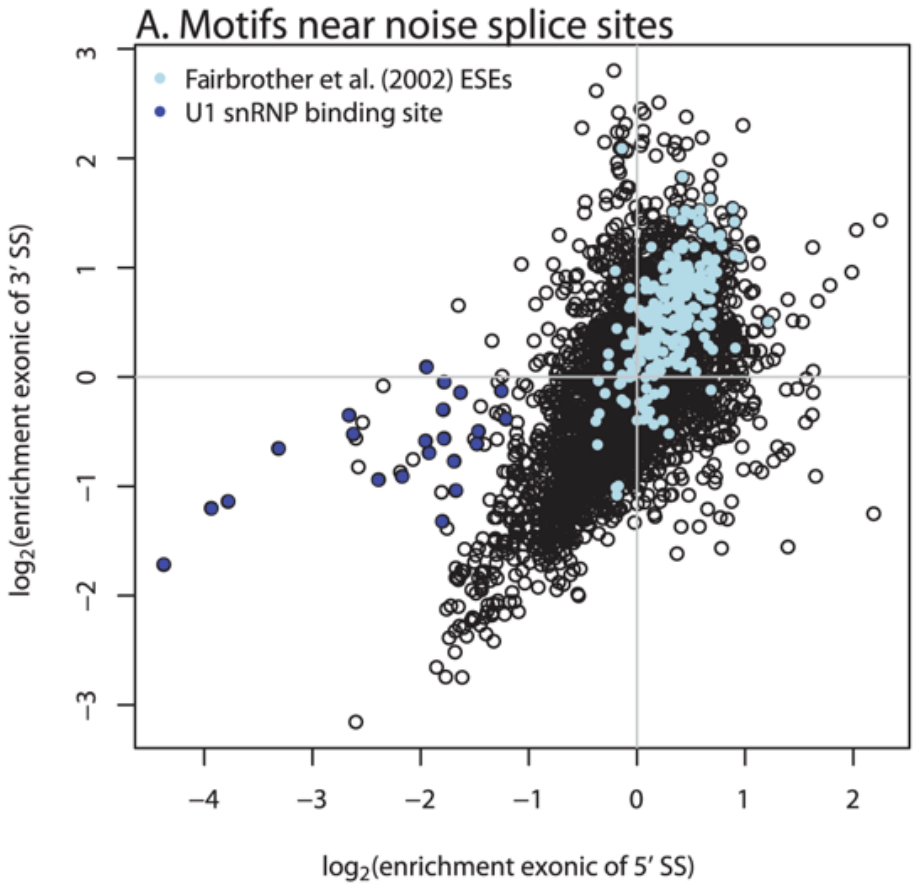
D: Conservation vs. gene expression after correcting for intron length

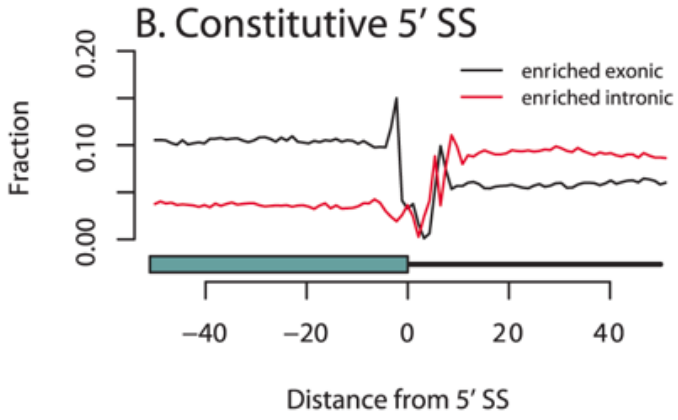
Noisy splice sites are marked by genomic features that define exons

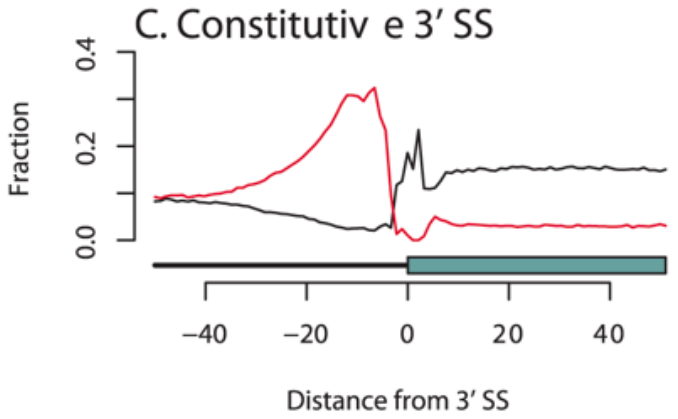
- Analysis of exonic hexamer enrichment in the vicinity of unconserved, rarely-used splice sites compared to nearby “decoy” splice sites
  - Decoy splice sites were never observed to be used in the data

- Number of significant enriched/depleted hexamers in exonic region:
  - 5’ splice site: 574
  - 3’ splice site: 728

Hexamers enriched near unconserved splice sites are relevant in exon definition







This is evidence that noisy splicing uses the same general splicing factors

Hexamers identified near noisy splice sites demarcate exon boundaries of constitutive exons

Conclusions

- Study examined alternative splicing in human cells
- Demonstrated that a large fraction of alternative transcript isoforms is evolutionary unconserved
- Noisy splicing is a result of stochastic binding of splicing factors involved in exon recognition
  - Splicing machinery occasionally misses the intended splice site
  - Implication: stochastic vs. deterministic splicing code
- Authors extrapolate that the majority of transcript isoforms is not functionally relevant