Why there is more to gene evolution than protein function: splicing, dual-coding sequence and why it matters

Laurence D. Hurst
University of Bath, UK
The history of genomics is a history of unexpected and enigmatic discoveries:

- Why if you have two copies of every gene do we see many expressed in a haploid manner (imprinting, allelic exclusion)?
- Why do most knockouts have little or no phenotype?
- Why do most genes have more than one transcript?
- Why make proteins you don’t currently need?
- Why is most of our genome transcribed but not protein coding?
  - Why do we see ncRNA that is destroyed almost as soon as it is made?
Two conceptions of genome organization:

A perfect machine

A load of rubbish, but gets the job done

Is the genome a precisely engineered Swiss movement or a Mickey Mouse watch?
For example:

- Why is the genome nearly all transcribed?
  - SW) There is rich world of RNA level regulation.
  - MMW) Transcription is intrinsically noisy – open chromatin gets transcribed.
Why is it important to know?

Understanding how genes and genomes work is necessary for understanding how mutations and genomic alterations cause disease and, in turn, for diagnostics.

It can also aid rational design of genetic manipulations (transgenes, gene therapy).
In this talk I’ll examine two classical presumptions about the evolution of our genes:

1. Mutations that affect exons but don’t alter the protein (synonymous mutations) will be effectively neutral in mammals.
   - Parts of the mechanism assumed to have no meaning.

2. That to understand where purifying selection acts in proteins you need only consider the biology of the protein.
   - Parts of the mechanism where we presume we know the meaning.
Splicing is the removal of intronic sequence from a pre-mRNA

So how are exon-intron junctions recognized?
Structure of the talk:

The old view of splicing: exons are defined by information in introns
The new view of splicing: information in exons is also important in some species
The impact of splice related constraints in exons on:

1. Evolution at synonymous sites:
   1.a Is the rate of synonymous site evolution affected?
   1.b Is codon usage bias affected?

2. Evolution at non-synonymous sites
   2.a Is the rate of protein evolution affected?
   2.b Is amino acid usage affected in mammals?

Why does this matter?
In the classical model of splicing, information for intron removal is contained within the intron.
In many taxa information in exons is used to define exon-intron junctions:

- Serine-Arginine (SR) proteins bind to exonic splicing enhancers (ESE) motifs (≈6-8 nt).
- The more ESEs in an exon, the higher the chance that the exon is correctly spliced.
**ESE properties:**

- Most abundant near intron-exon boundaries
- ESEs tend to be A-rich (50%) C-poor (~10%)
- May explain a dearth of SNPs near exon-intron boundaries

*Fairbrother et al, 2004*
How does the need to specify ESEs impact:

1. *Evolution* at synonymous sites:
   1.a Is the synonymous rate of evolution affected?
   1.b Is codon usage bias affected?

2. *Evolution* at non-synonymous sites
   2.a Is the protein rate of evolution affected?
   2.b Is amino acid usage affected?
ESEs evolve $1/3$ slower than non-ESE sequence.

Depending on the definition of ESE, around 5% [range 2-9%] of all synonymous mutations are under purifying selection.
Genes with much sequence near exon-intron junctions have low Ks

Spearman rank correlation, rho = -0.15
How does the need to specify ESEs impact:

1. Evolution at synonymous sites:
   1.a Is the synonymous rate of evolution affected?
   1.b Is codon usage bias affected?

2. Evolution at non-synonymous sites
   2.a Is the protein rate of evolution affected in mammals?
   2.b Is amino acid usage affected?
Some codons commonly employed in ESEs, e.g. GAA*, are more abundant near junctions than their synonyms.

Are codons preferred in ESEs generally preferred near exon ends?

Willie and Majewski TiGs 2004
Most codons show biased usage near boundaries

Of the 59 codons with at least one synonym, 47 show significant trends in usage near intron-exon junctions of which 42 are significant after sequential Bonferonni correction
In 63 of 87 pairwise comparisons, synonymous codons relatively preferred in ESE are more preferred near boundaries.
How does the need to specify ESEs impact:

1. Evolution at synonymous sites:
   1.a Is the synonymous rate of evolution affected?
   1.b Is codon usage bias affected?

2. Evolution at non-synonymous sites
   2.a Is the protein rate of evolution affected?
   2.b Is amino acid usage affected?
Protein evolution is slowest near intron-exon boundaries.

Empty circles - all exons

Squares - exons longer than 80 codons

Filled circles - functional retrogenes
Protein evolution is dependant on the proportion of sequence near intron-exon boundaries.

When between gene heterogeneity is controlled, by comparing exon centres to exon flanks, the reduction in $K_a$ is 30-50%.
How does splicing compare to other covariates of protein evolution?

Controlling for covariance, partial correlation reveals that the proportion of sequence within 70 nucleotides of intron-exon boundaries is an equally strong factor as expression breadth in mammals.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>$R_{XY}$</th>
<th>$R_{XYI}$</th>
<th>$P_{XYZ}$</th>
<th>$R_{XZ}$</th>
<th>$R_{XZY}$</th>
<th>$P_{XZY}$</th>
<th>$R_{YZ}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ka</td>
<td>Proportion 70</td>
<td>Breadth of expression</td>
<td>-0.202</td>
<td>-0.16</td>
<td>&lt;0.0001</td>
<td>-0.21</td>
<td>-0.17</td>
<td>&lt;0.0001</td>
<td>0.23</td>
</tr>
</tbody>
</table>
How does the need to specify ESEs impact:

1. Evolution at synonymous sites:
   1.a Is the synonymous rate of evolution affected?
   1.b Is codon usage bias affected?

2. Evolution at non-synonymous sites
   2.a Is the protein rate of evolution affected?
   2.b Is amino acid usage affected?
As expected, amino acid usage tends to be highly skewed near intron-exon boundaries.

Positive Rho - amino acid disfavored
Negative Rho - amino acid preferred near boundaries.

Of 46 amino acid comparisons (5 prime & 3 prime), 34 showed significant trends near boundaries.

<table>
<thead>
<tr>
<th>AA</th>
<th>5' Rho</th>
<th>5' P</th>
<th>3' Rho</th>
<th>3' P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.866</td>
<td>1.36E-07</td>
<td>0.661</td>
<td>4.32E-05</td>
</tr>
<tr>
<td>C</td>
<td>0.095</td>
<td>0.59</td>
<td>0.140</td>
<td>0.436</td>
</tr>
<tr>
<td>D</td>
<td>-0.499</td>
<td>0.0035</td>
<td>-0.578</td>
<td>0.0005</td>
</tr>
<tr>
<td>E</td>
<td>-0.642</td>
<td>8.31E-05</td>
<td>0.125</td>
<td>0.48</td>
</tr>
<tr>
<td>F</td>
<td>-0.520</td>
<td>0.002</td>
<td>-0.757</td>
<td>1.40E-06</td>
</tr>
<tr>
<td>G</td>
<td>-0.058</td>
<td>0.75</td>
<td>0.301</td>
<td>0.0886</td>
</tr>
<tr>
<td>H</td>
<td>0.607</td>
<td>0.0002</td>
<td>-0.202</td>
<td>0.26</td>
</tr>
<tr>
<td>I</td>
<td>-0.830</td>
<td>3.54E-07</td>
<td>-0.839</td>
<td>2.88E-07</td>
</tr>
<tr>
<td>K</td>
<td>-0.881</td>
<td>6.95E-08</td>
<td>-0.936</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0.279</td>
<td>0.115</td>
<td>0.505</td>
<td>0.003</td>
</tr>
<tr>
<td>M</td>
<td>-0.628</td>
<td>0.00013</td>
<td>-0.446</td>
<td>0.00980</td>
</tr>
<tr>
<td>N</td>
<td>-0.582</td>
<td>0.0005</td>
<td>-0.590</td>
<td>0.0004</td>
</tr>
<tr>
<td>P</td>
<td>0.617</td>
<td>0.00018</td>
<td>0.660</td>
<td>4.42E-05</td>
</tr>
<tr>
<td>Q</td>
<td>0.874</td>
<td>9.77E-08</td>
<td>0.440</td>
<td>0.011</td>
</tr>
<tr>
<td>R</td>
<td>0.875</td>
<td>9.34E-08</td>
<td>0.959</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td>0.476</td>
<td>0.005</td>
<td>0.450</td>
<td>0.0091</td>
</tr>
<tr>
<td>T</td>
<td>0.723</td>
<td>4.45E-06</td>
<td>-0.257</td>
<td>0.15</td>
</tr>
<tr>
<td>V</td>
<td>-0.175</td>
<td>0.33</td>
<td>0.391</td>
<td>0.025</td>
</tr>
<tr>
<td>W</td>
<td>-0.069</td>
<td>0.71</td>
<td>-0.125</td>
<td>0.49</td>
</tr>
<tr>
<td>Y</td>
<td>-0.055</td>
<td>0.759</td>
<td>-0.376</td>
<td>0.033</td>
</tr>
<tr>
<td>I</td>
<td>-0.958</td>
<td>0</td>
<td>-0.728</td>
<td>3.67E-06</td>
</tr>
<tr>
<td>s</td>
<td>0.795</td>
<td>6.32E-07</td>
<td>0.791</td>
<td>6.75E-07</td>
</tr>
<tr>
<td>r</td>
<td>-0.696</td>
<td>1.22E-05</td>
<td>-0.840</td>
<td>2.84E-07</td>
</tr>
</tbody>
</table>
These trends are, at least in part, owing to effects at the nucleotide not the protein level:
Non-synonymous substitutions are less frequent in \textit{ESE} sequence.

Five prime domain of exons from mouse-human alignment comparison.
Amino acids whose codons are preferred in ESEs are preferred near intron-exon boundaries.

Codon enrichment is ESEs

\( \rho = -0.54, P < 0.0001 \)
Exons of protein-coding genes encode the information not just for *what* protein is made but also *how* the protein is made.

Just like tetra pak, protein anatomy reflects a mixture of functional necessity and remnants of the manufacture process.
So what?

These manufacture remnants are especially common in humans.

- This seems to relate to the fact that our genes have many and large introns...
Where exons are small islands in large deserts of intronic DNA, ESE usage near boundaries is more common. True also for 30 species in a Bayesian framework.
Larger introns cause difficulties in exon boundary recognition.

- **Experimental evidence that increasing intron size leads to splicing problems:** Insertion of exogenous DNA into introns greatly reduces splicing rate. (e.g. Cullen et al. 1982, *NAR* 10, 6177-6190; Klinz and Gallwitz, 1985, *NAR* 13, 3791-3804)

- **Functional genomic evidence:** alternatively spliced exons are associated with large flanking introns (e.g. Bell et al. 1998, *Mol. Cell Biol.* 18, 5930-5941; Fox-Walsh et al. 2005, *PNAS* 102, 16176-81).

- **Comparative genomic evidence:** exons flanking large introns are more likely to be lost evolutionarily (Kandul and Noor, *BMC Genomics*, 2009 10, 67)
In humans ESE density at exon ends increases with increasing size of the flanking intron.

More synonymous sites are under selection in ESEs next to long introns.
We might then expect that many disease-causing mutations act by disrupting splicing, including synonymous variants

Prior evidence has indicated circa 60-100 examples of synonymous SNPs that cause disease, most via splice disruption:

- e.g. FBN1 and Marfan syndrome, SMN2 Motor Neurone Disease, both associated with ESE synonymous mutations
How commonly do disease-associated mutations act via disruption of splicing?

- Extract missense and synonymous SNPs from internal exons in Clinvar database
  - =8,250 SNPs
- There is a great enrichment at exon ends
A conservative estimate suggests 24% of disease causing SNPs have their effects via splicing defects:

- Assume no mutation in exon cores causing splicing defects -> background rate of 0.00175 SNPs/bp
- End of exon in disease causing genes is $2.7 \times 10^6$ bp -> expected 4797 disease causing SNPs
- 6774 observed
- therefore an excess of $1977/8250 = 24\%$
A more liberal estimate suggests this could be more like >40%

- 25% of mutations at exon cores also disrupt splicing.
- Background rate is thus 0.0013 SNPs/bp not affecting splicing
- Total length internal exons x background, predicts 4639 mutations that don’t have effects via splicing
- \( \frac{8250 - 4639}{8250} = 43\% \) of disease causing SNPs have effects via splice disruption, mostly via ESE disruption, not splice sites.
Close analysis of some exons agree, e.g.:

- 61% of random point mutations in exon 6 of FAS disrupt splicing: *Nature Comms* 2016 10.1038/ncomms11558
- 77% of random point mutations in exon 10 of MLH1: *PlosG* 2016 12 e1005756
- In all cases synonymous and non-synonymous sites were affected
Disease-causing SNPs associate with “fragile exons”

- The exons most likely to bear disease causing SNPs are associated with low ESE density:
  - Flanked by small introns
  - Towards 3' end of genes (but not of CDS)
  - Associated with non AGgt splice sites
  - Disease causing SNP density is higher where ESE density is lower
- Hypothesis: some exons more “fragile” than others?
Conclusions:

- Owing to our large and numerous introns, our genes have a problem splicing.
- Reinforcement with ESEs partially solves this problem. This affects:
  - codon usage and rates of synonymous evolution
  - amino acid usage and rates of protein evolution
  - 1/4-1/2 of disease-causing mutations have effects via splice modification
- Don’t ignore synonymous SNPs as potential disease candidates
- Don’t suppose missense mutations act via one amino acid change
Looking ahead:

- Can we use this information to better predict which synonymous SNPs might be disease causing?

- Can we use this information to improve largely intronless transgenes (e.g. for gene therapy)?
Acknowledgements: the people who did the work

Joanna Parmely
Toby Warnecke
XianMing Wu
Jean-Vincent Chamary
Andreas Schueler