The Evolution of Junk

Revising the proportion of the genome undergoing evolutionary constraint through RNA structure

Presented by
Martin A. Smith

Supervised by
John S. Mattick

e-mail: martin.smith@imb.uq.edu.au
Evidence that “Junk” DNA is underrated

<table>
<thead>
<tr>
<th>Genes</th>
<th>~19 300</th>
<th>~20 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome (Mb)</td>
<td>~100</td>
<td>~3 100</td>
</tr>
<tr>
<td>% Coding</td>
<td>26</td>
<td>1.2</td>
</tr>
</tbody>
</table>
...and that ncRNAs are prevalent
How to detect functional transcripts?

1. $\partial$-expression

Dinger et al. Genome Research 2008

Mercer et al. PNAS 2007
How to detect functional transcripts?

2. Sequence conservation
How to detect functional transcripts?

3. Discrepancies of structural conservation
What proportion of mammalian genomes bear evidence of functional selection through RNA structure?

Discover new genomic regions harboring conserved RNA structures

- Evolutionary conservation
- Locus-specific abundance

Functional characterization of long ncRNAs and transcriptome annotation

- Parsing structural predictions
- Ontological enrichment of structural motifs
- Association with SNPs, DNA triplexes & Pyknons
Some previous work

Structured RNAs in the ENCODE selected regions of the human genome
Stefan Washietl, Jakob S. Pedersen, Jan O. Korbel, et al.
*Genome Res.* 2007 17: 852-864

Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions
Elfar Torarinsson, Zizhen Yao, Eric D. Wiklund, et al.
*Genome Res.* 2007 18: 000

Caveats

- Sketchy background models (mono-nt shuffling)
- Constrained input alignments (# and quality of seqs)
Can targeting specific ranges of sequence conservation improve the resolution of conserved structure prediction?

Focus on informative sequences: compensatory mutation “sweet-spots”

Compare the efficiency of existing tools with different input parameters
Analysis Pipeline

Multiple Genome Alignment

Split Alignment Blocks into Overlapping Windows

Limit Sequence Conservation and Gap Content

Structural Prediction

Comparative Variables

[Enredo-Pecan-Ortheus] vs. TBA/Multiz

Window Size & Overlap [150, 75]

Min and Max Similarity [65-95%]
Mean Pairwise Identity/Distance

Prediction Algorithms
[SISSIz] vs ALIFOLDz vs EVOFOLD
Pre-processing input data
Pre-processing input data

- Remove similar sequences
- Remove divergent & indel-rich sequences
Pre-processing input data

Remaining sequences fall within optimal range for comparative analysis of compensatory mutations

Reduce background noise, time & memory constraints
How many species are maintained post-filtering?
How many species are maintained post-filtering?

Abundance of species in filtered alignment blocks [65%-95%]
Do existing tools predict known RFAM structures?

SISSIz prediction scores for 266 RFAM-overlapping filtered alignment blocks

35% $P < 0.1$
Do high-confidence predictions present specific characteristics?

Impact of evolutionary depth (amount of species) on structural predictions

Species in alignment

Z-score

- P>0.9
- P<0.9
Do high-confidence predictions present specific characteristics?

Impact of evolutionary depth (amount of species) on structural predictions
Additional Strategies for detecting evolutionary conserved structures

- Syntenic Multiple Genome Alignment
- Split Alignment Blocks into Overlapping Windows
- Limit Sequence Conservation and Gap Content
  - Structural Prediction
  - Structural Alignment
  - Ad-Hoc Scoring
- Spliced IncRNAs
Acknowledgments

**IMB**
Prof. John S. Mattick
Dr Marcel Dinger
Dr Timothy Mercer
Dr Michael Pheasant
Dr Ryan Taft
Dr Cas Simons
Paulo Amaral
Fabian Buske
Selene Fernandez-Valverde
Satu Nahkuri
and all other group members

**Uni. Leipzig**
Prof. Peter F. Stadler
Ms Petra Pregel

**Uni. Strasbourg**
Prof. Eric Westhof

**Funding**
UQRS/IPRS

**Australian Government**

---

**THE UNIVERSITY OF QUEENSLAND**

**IMB Institute for Molecular Bioscience**