Funseq2 mod for paper E

Lou Shaoke

Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

July 25, 2016

Yale

Objective

Objective

- Not change the Funseq score schema
- Use PCAWG annotation standard.
- expand Funseq2 to explain relative importance for promoter, enhancer;

Components of Funseq NC score

All SNV: coding + non-coding
Non-coding schema

Feature	W function	Components
MOTIFG	0.95978+0.00462 * value	
MOTIFBR	0.7242+0.1583 * value	
HUB	exp(-2.903 + 2.899 * value)	Gene network degree rank
ANNO	0.155385998694222	TFP,TFM, DHS, pGene,lincRNA, mirRNA, Enhancer etc
SEN	0.969106841199359	
USEN	0.997235765635279	
UCONS	0.999746528403	
HOT	0.797248596038818	
GENE	0.0114791273831854	Intron, UTR, Promoter mutex DRM-enhancer
GERP	0.622834294640819*(1/(1+exp(-40*(value-1.85))))	
RECUR	1	

Promoter and Enhancer

Promoter annotation is part of 'GENE' and also part of 'HUB'. Enhancer annotation is part of 'ANNO'. Enhancer in DRM is a subset of Enhancer from ENCODE annotation.

Promoter and Enhancer

Weight calibration (1-entropy):
GENE (discret): $n = n_{intron} + n_{utr} + n_{promoterorenhancer}$
ANNO (discret): TFP,TFM, DHS, pGene,lincRNA, mirRNA, Enhancer etc from ENCODE

Promoter and Enhancer

Weight calibration (1-entropy):
GENE (discret): $n = n_{intron} + n_{utr} + n_{promoterorenhancer}$
ANNO (discret): TFP,TFM, DHS, pGene,lincRNA, mirRNA, Enhancer etc from ENCODE

Score calcuation:
Knowledge based priority for 'GENE' when calculating Funseq score: GENE, but not HUB, not MOTIFG, which means for SNV has HUB score and MOTIFG score will be calculated a 'GENE' score.
Similarly, 'ANNO': ANNO, Not SEN, Not MOTIFBR, Not HOT.

Promoter and Enhancer

Weight calibration (1-entropy):
GENE (discret): $n = n_{intron} + n_{utr} + n_{promoterorenhancer}$
ANNO (discret): TFP,TFM, DHS, pGene,lincRNA, mirRNA, Enhancer etc from ENCODE

Score calcuation:
Knowledge based priority for 'GENE' when calculating Funseq score: GENE, but not HUB, not MOTIFG, which means for SNV has HUB score and MOTIFG score will be calculated a 'GENE' score.
Similarly, 'ANNO': ANNO, Not SEN, Not MOTIFBR, Not HOT.

The impact might not be reflected in funseq score, even when you see the feature in output;
$W_{n_{intron}+n_{utr}+n_{promoterorenhancer}}$ => GENE score (intron or utr or promoter or enhancer)
$W_{n_{intron}|n_{utr}|n_{promoterorenhancer}}$ increase => GENE-subgroup score increase
The simplest way is: #SNV have GENE annotated feature, $\#Promter/\#SNV$, $\#DRM/\#SNV$ to calculate the contribution.

Conclusion

re-calibrate new weight based on PCAWG data
Yes:Use PCAWG data, re-calibrate new weight
Question: Do we need replace all the annotations: promoter, cds, utr, enhancer, lincRNA, mirRNA etc, or just cds, promoter, utr, enhancer.

Get importance of promoter/enhancer score after recalibration
Yes: Use similar way to get promoter, enhancer relative importance weight or proportion method

Warning: Funseq score is complex, which is involved in many empirircal rules.

Discussion