Funseq and NCVARG

Lou Shaoke

Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

June 30, 2016

Yale

update about funseq2

New whole genome calculation based on hg19 using Funseq2.1.6
Somatic mode
All possible mutations for each positive: $A->C|G|T$
Query with a bed file using tabix, with some simple scripts
Available at: /net/gerstein/skltmp/funseq2.1.6/
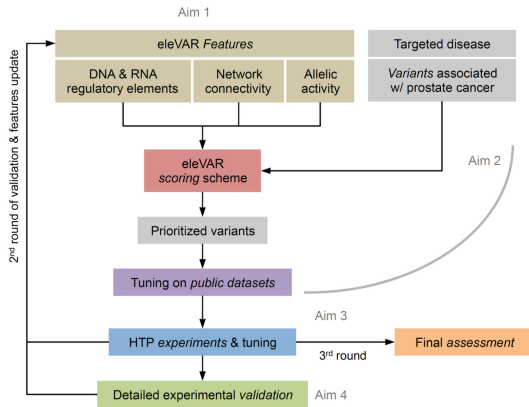Also updated: funseq2 and funseq3 website.

NCVARG workflow

Year 1
- identify a uniform list of disease *variants*
- tuning & validation flow (1st round):

  • list of *features* & weighting scheme refinement
  • variant *scoring*
  • tuning of parameters on *public datasets* (HGMD, ClinVar, CRE-seq, 2HIT)
  • HTP *experiments* & tuning (Clone-seq and luciferase reporter assays)
  • detailed experimental *validation* (TaqMan assays, qPCR, CRISPR/CAS)

Year 2
- add new features
- continue tuning flow (2nd round)

Year 3
- continue tuning flow (3rd round)
- final *assessment*

Aim 1

eleVAR *Features*			Targeted disease
DNA & RNA regulatory elements	Network connectivity	Allelic activity	*Variants* associated w/ prostate cancer

2nd round of validation & features update

Aim 2

eleVAR *scoring* scheme

Prioritized variants

Tuning on *public datasets*

Aim 3

HTP *experiments* & tuning → 3rd round → Final *assessment*

Detailed experimental *validation*   Aim 4
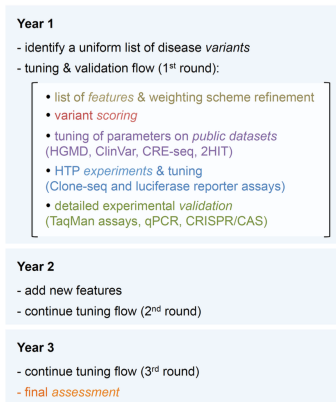
Fig. 5: Timeline & overall grant workflow.

Dataset and features emphasized in NCVARG

- RNA features and information
  CLIP-Seq binding Peaks
  ncRNA structure deleterious effect
- Allelic analysis: ASE, ASB etc
- Distal regulatory element
- tissue specifity

enhancer gene linkage

Three source: FunSeq built-in, MIT LDA, KevinYip.

Source	Description	Tissue-specific	Number
Funseq built-in	Correlation between gene expression and Histone mark and DNA methylation	No	251469
MIT LDA	LDA mix-membership algorithm	Yes	709700 (56)
KevinYip	correlate histone mark with expression and learn using LASSO	Yes	1093488 (56)
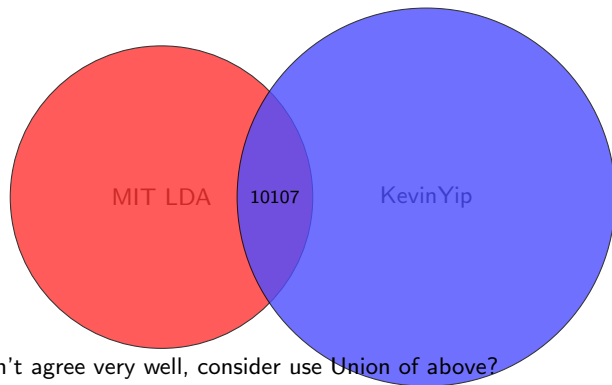
Doesn't agree very well, consider use Union of above?

Proposed experimental update

Experimental Eval ▪   Un-validated, sorted by $|-(RS_v^{(current)} - RS_v^{(previous)})|$
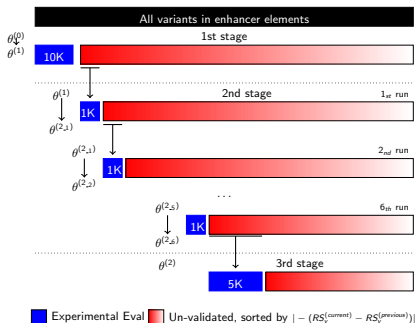
Stage 1: randomly select 2,000 active enhancers in K562 cells

$$logit(P(y_v = 1)) = -k \times (RS_v - \alpha) = -k \times (\sum_m \theta_m * s_{v,m} - \alpha)$$

Stage 2: clone 200 untested enhancers and choose 1,000 variants
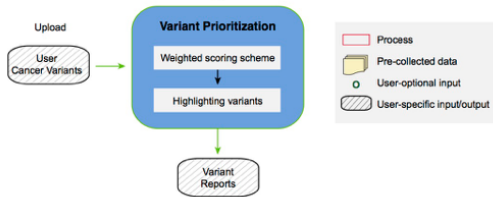
$$|log(OR_v)| = |logit(P(y_v = 1|\theta^{(current)})) - logit(P(y_v = 1|\theta^{(previous)}))|$$
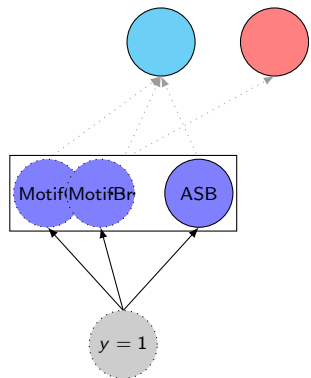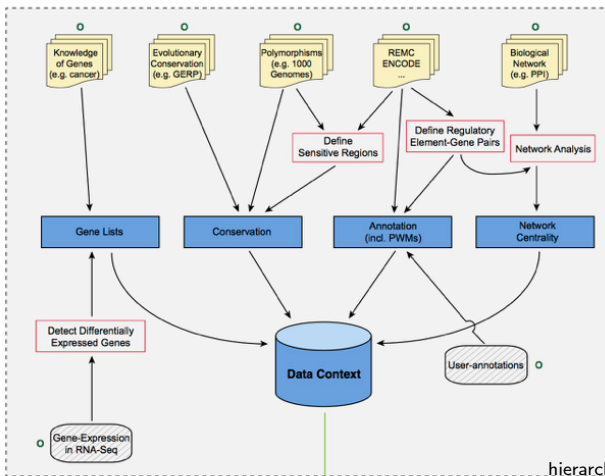
Stage 3: select 1,000 variants (400 with predicted high impact, 200 with medium impact, and 400 with low impact) on previously cloned enhancers and 4,000 variants

Enhancer binding and experimental validation results == variant impact? NO!!!
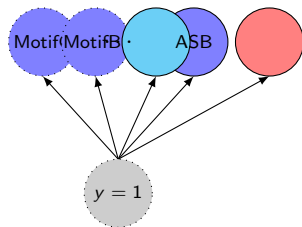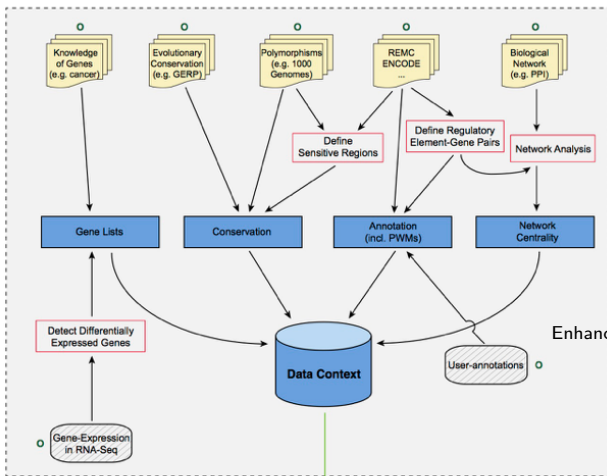
Enhancer activity is not equivalent to impact

hierarchical structure learning

make a latent enhancer/TF binding var