Journal Club

PETModule: a motif module based approach for enhancer target gene prediction

Changyong Zhao, Xiaoman Li & Haiyan Hu

Lilly Reich
Mentor: Shaoke Lou
Gerstein Laboratory

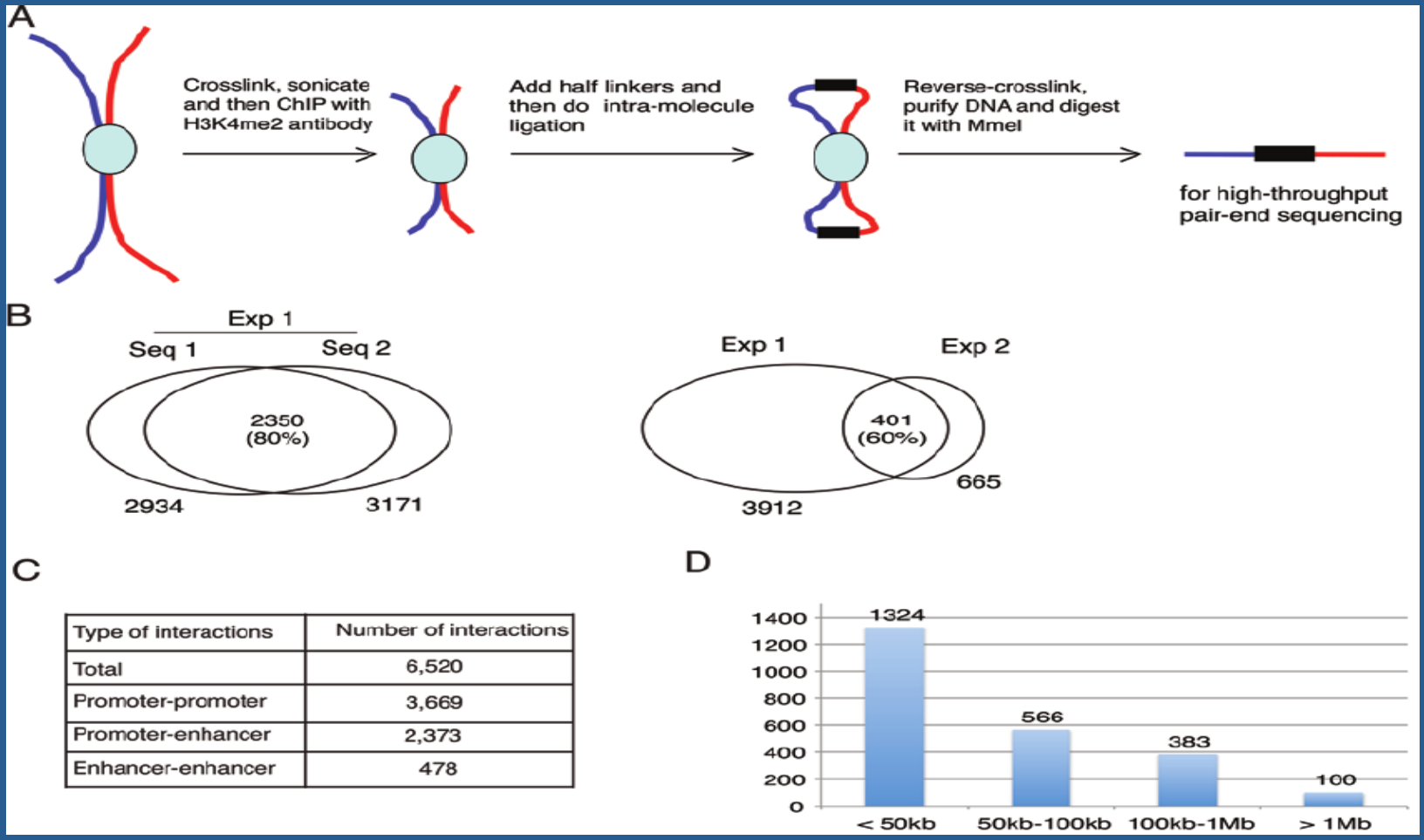
What is the enhancer target gene problem?

- Human genome contains hundreds of thousands of enhancers
- Enhancers scattered across 98% of human genome that doesn't encode proteins results in large search space (billions bp DNA)
- Location to target gene is highly variable
- General sequence of code enhancers is poorly understood
- Enhancers cannot be identified computationally with high confidence

What were the previous approaches?

- **IMPET (Integrated Methods for Predicting Enhancer Targets)**
- Pre-STIGE (Protein Epitope Signature Tags)
- Not been many supervised methods for enhancer target prediction

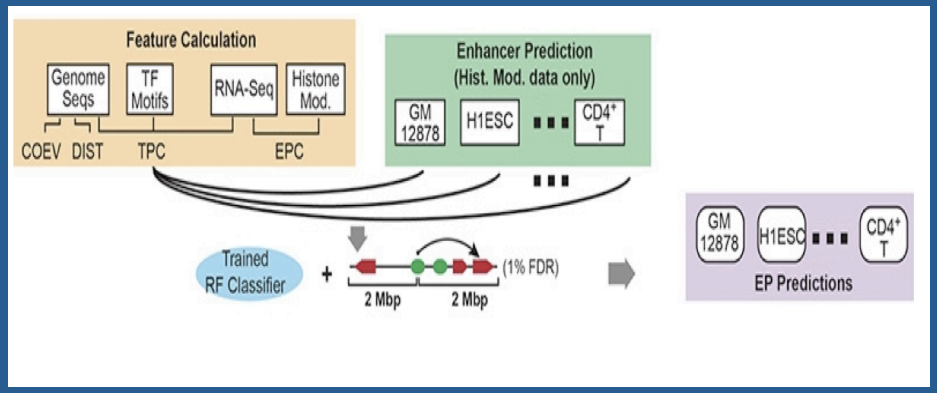
Identification of genome wide enhancer promoter interaction



Chepelev et al., 2012.

IMPET

- Identifies target promoters integrating multiple types of genomic data
- supervised method that uses chromosome conformation data
- Tests 4 features
  - Distance constraint
  - Enhancer promoter activity correlation (EPC)
  - TF target and promoter correlation (TPC)
  - Co-evolution of enhancer and target promoter (COEV)
- positive examples enhancer – promoter pairs with ChIA-PET connections in K562
- random enhancer-promoter pairs with distance follows background distribution of non-interacting genomic info in chromatin fiber



PreSTIGE

- Considers gene expression within each replicate across entire gene expression profile
- Identifies outliers across replicates
- Distant metastasis related genes from noisy expression data CD44 + CD24-/low tumor initiating cells
- Required to prepare 3 different input files before ETG pair prediction

Torkamani et al., 2009.

ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)

- Immuno-precipitation step (anti-body) to enrich for chromatin that's bound for specific protein
- Used to map interactions of many transcription factors (Li et al.,2013)
- Gives information about the (potential) role of proteins in structuring 3D genome organization

Goh et al., 2012

PETModule (Predicting enhancer target by modules)

- Predicts ETG pairs
- Trained with positive and negative ETG pairs
- Four features (distance, CSS,FSS,correlation)
- More precise than IMPET and PresTIGE
- Machine learning approaches
  - Information gain attribute evaluator (evaluates contribution of feature)
  - SVM (evaluates importance of feature)
  - LASSO(least absolute shrinkage and selection operator) (constructs linear model and coefficients to zero)
  - Random forests (assigns classification trees and assigns new object to class most trees vote for)

PETModule features

- CSS (Conserved Synteny Score)
  - Group of linked genes that are considered homologous

$$CSS(e, g) = \sum_{s=1\ldots k} \delta_s(e, g) \times \emptyset(r, s)$$

$$\delta_s(e, g) = \begin{cases} 1 & \text{if } d_s(e, g) < \Theta \text{ in species s} \\ 0 & \text{otherwise} \end{cases}$$

FSS cont.

- FSS (Functional Similarity Score)
  - GO (gene ontology) serve as a connector for understanding functional relationship between genes

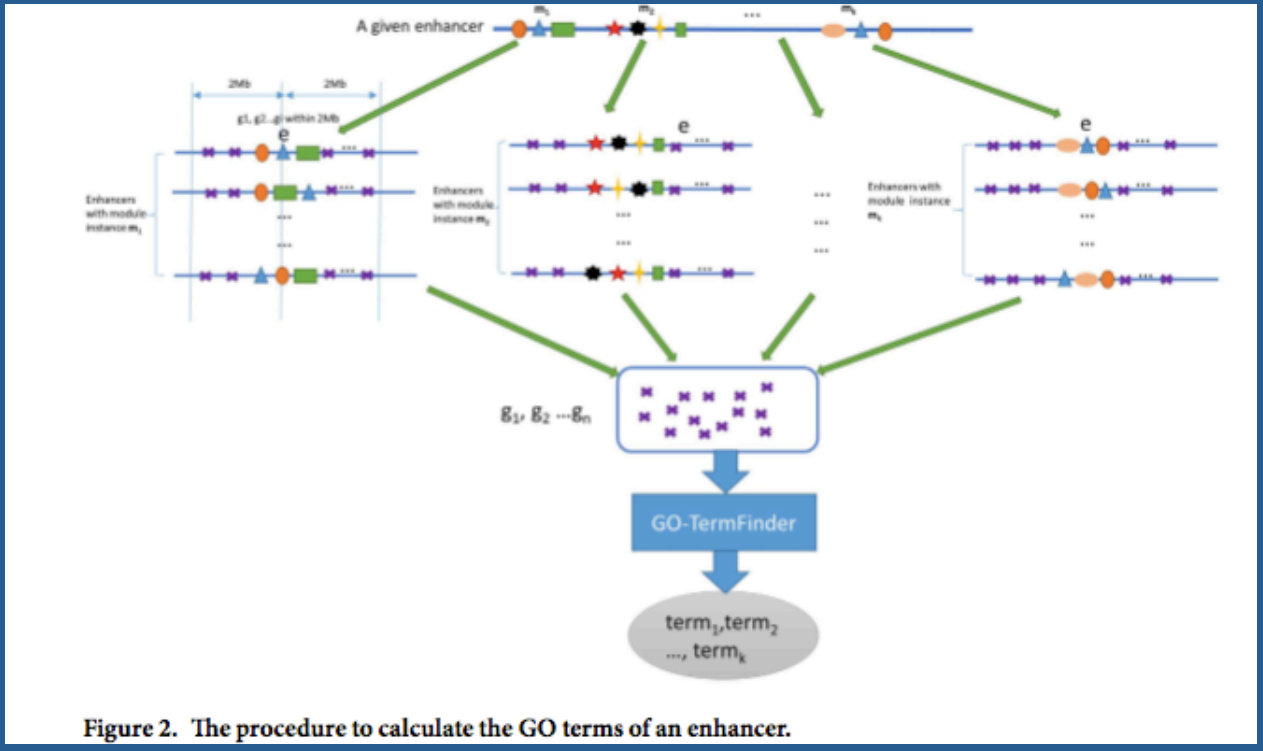


Figure 2. The procedure to calculate the GO terms of an enhancer.

Cont.

- FSS
  - IC – information content
  - MICA – most informative common ancestor
  - Calculate the GO terms in respect of e, t1, and each GO term of g, t2

$$sim(g_i, g_j) = \frac{sim(g_i \rightarrow g_j) + sim(g_i \rightarrow g_j)}{2}$$

$$sim(g_i \rightarrow g_j) = avg\left[\sum_{t \in g_i} \max_{t \in g_j} IC(MICA(t1, t2))\right]$$

PETModule prediction on three datasets

Dataset	Enhancers	Known pairs	Predicted pairs	Known pairs predicted	Recall	Precision	ROC AUC	F1 score
ChIA-PET (K562)	3300	4110	9244	1917	0.466	0.207	0.938	0.287
ChIA-PET (MCF7)	341	370	560	187	0.505	0.334	0.968	0.402
Hi-C (IMR90)	10920	19666	26467	7811	0.397	0.295	0.942	0.338
Overall	14561	24146	36271	9915	0.411	0.273	0.949	0.328

Table 1. PETModule prediction on three datasets with experimentally defined ETG pairs. The known ETG pairs here do not contain any of the positive ETG pairs used for training.

- Correlation approaches identify multiple targets of enhancer
- Correlation calculation might be affected by selected experiments and certain target genes missed

PETModule predicted ETG pairs supported by Hi-C and ChIA-PET data

Cutoff	#Enhancers with supporting Hi-C data	#Predicted ETG pairs	#Known ETG pairs	#Known ETG pairs predicted	Recall	Precision	ROC AUC	F1 score
5	10881	23454	64075	17354	0.271	0.740	0.890	0.397
10	9918	22869	32837	12031	0.366	0.526	0.914	0.432
15	8433	21145	20319	8413	0.414	0.398	0.924	0.406
20	7069	19131	14024	6054	0.431	0.316	0.928	0.365
25	5945	17025	10219	4479	0.438	0.263	0.929	0.329

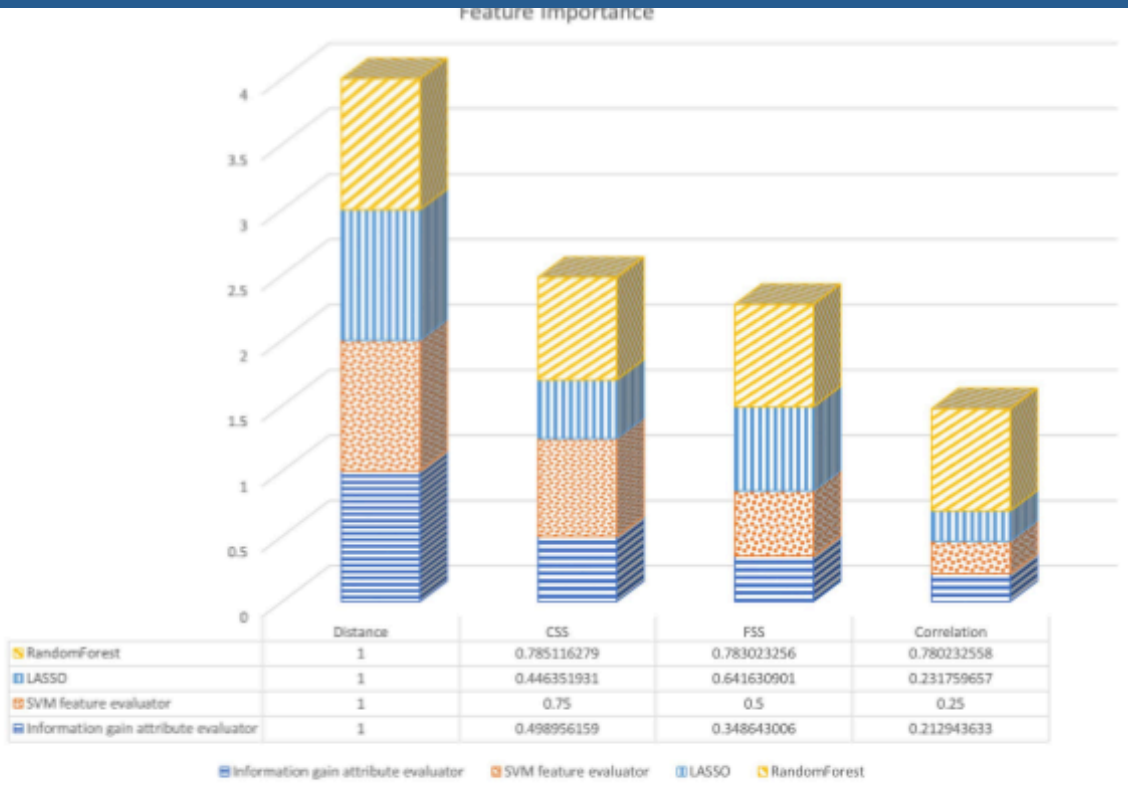
Table 2. PETModule prediction on IMR90 assessed with Hi-C contact matrices. The cutoff specifies the minimum number of supporting Hi-C reads required to define known ETG pairs. The known ETG pairs here do not contain any of the positive ETG pairs used for training.

- Large number of predicted ETG pairs meaningful and precision PETModule underestimated Table 1

ETG pairs reveals new characteristics

- Number of predicted ETG pairs doubled but precision on average decreased 8.3% when predicting ETG within 2 Mb instead of 1Mb
- 69.9% of enhancers not consecutive in genome

The importance of features ranked by four methods



	Distance	CSS	FSS	Correlation
RandomForest	1	0.785116279	0.783023256	0.780232558
LASSO	1	0.446351931	0.641630901	0.231759657
SVM feature evaluator	1	0.75	0.5	0.25
Information gain attribute evaluator	1	0.498956159	0.348643006	0.212943633

Figure 1. The importance of the four features ranked by four methods.

Predicted mouse ETG pairs supported by Hi-C and 3C data

- Applied trained model on human data without correlation feature to two mouse datasets in CH12 cell line and macrophage cell
- Accuracy of PETModule on two mouse datasets similar to human datasets
- Trained model using mouse data 5% higher recall and >9% precision
- Performance suggested difference between human ETG and mouse ETG pairs

Prediction results on two mouse cells

Prediction Model	Dataset	Enhancers	Known pairs	Predicted pairs	Known pairs predicted	Recall	Precision	ROC AUC	F1 score
Human model	CH12	14195	24516	124102	16540	0.667	0.133	0.938	0.220
	macrophage	387	387	3171	251	0.650	0.076	0.923	0.135
Mouse model	CH12	14195	24516	64512	18252	0.744	0.283	0.968	0.410
	macrophage	387	387	1468	271	0.700	0.167	0.961	0.269

Table 3. Prediction results on two mouse cells.

- Mouse PETModule trained 5% higher recall and 9% higher precision

PETModule Results

Dataset	Tools	Enhancers	Known pairs	Predicted pairs	Known pairs predicted	Recall	Precision	ROC AUC	F1 score
ChIA-PET (K562)	PETModule	694	907	2285	429	0.473	0.188	0.938	0.269
	IM-PET	694	907	1872	278	0.307	0.149	0.88	0.200
	PreSTIGE	694	907	1468	382	0.421	0.260	0.8	0.322
ChIA-PET (MCF7)	PETModule	94	107	282	61	0.570	0.216	0.968	0.314
	IM-PET	94	107	191	33	0.308	0.173	0.88	0.221
	PreSTIGE	94	107	178	62	0.579	0.348	0.8	0.435
Hi-C (IMR90)	PETModule	202	411	714	184	0.448	0.258	0.942	0.327
	IM-PET	202	411	282	75	0.182	0.266	0.89	0.216
	PreSTIGE	202	411	342	114	0.277	0.333	0.8	0.303
Overall	PETModule	990	1425	3281	674	0.473	0.205	0.949	0.286
	IM-PET	990	1425	2345	386	0.271	0.164	0.88	0.205
	PreSTIGE	990	1425	1988	558	0.392	0.281	0.8	0.327

Table 4. Comparison of PETModule with IM-PET and PreSTIGE. Only the common enhancers with predictions by three methods were considered.

Conclusion/Discussion

- Enhancers regulate targets in condition-specific way
- PETModule comparable with state-of-the-art computational methods, still mandatory to improve
- FSS score still needs improvement