


Submit About Contact Journal Club Subscribe Institution: Guangzhou Medical University Log in Log out

PNAS Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author, or DOI  Advanced Search

Home Articles Front Matter News Podcasts Authors

NEW RESEARCH IN Physical Sciences Social Sciences Biological Sciences



## Global view of enhancer–promoter interactome in human cells



Bing He, Changya Chen, Li Teng, and Kai Tan



PNAS May 27, 2014 111 (21) E2191–E2199; first published May 12, 2014 <https://doi.org/10.1073/pnas.1320308111>


Edited by Xiaole Shirley Liu, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA, and accepted by the Editorial Board April 17, 2014 (received for review October 28, 2013)

Article Figures & SI Info & Metrics PDF

 Article Alerts  Share

 Email Article  Mendeley

 Citation Tools  Request Permissions

 Current Issue

## Paper Presentation and Analysis

Shaurya Jauhari, Mora Lab.

April 30<sup>th</sup> 2019

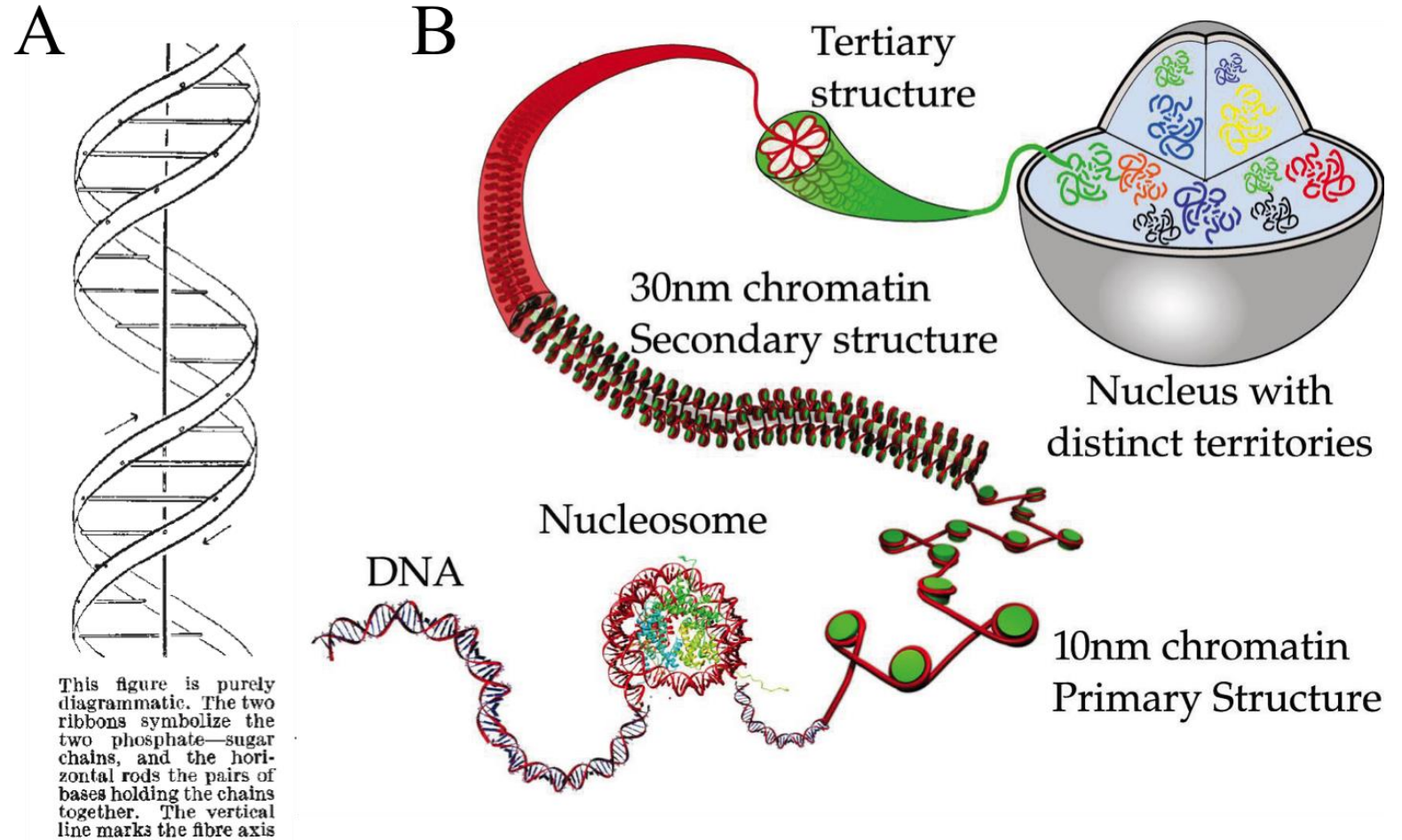
# Roadmap

- Underlying Concepts
- CSI-ANN Algorithm
- IM-PET Algorithm
- Take Home Points

# Bedrock Concepts

- Cells are the basic building blocks of life.
- Genes are the unit of heredity.
- Different cell types (hence tissues) manifest exclusive gene activities.
- How a cell body and function differs from others:
  - Which genes
  - What expression
- Gene Regulation structures gene's (and thereby cellular) function.
- Genome is the orchestrator and Activators, Enhancers, Inhibitors, Promoters are protagonists.
- Enhancers and Promoters are genomic regions (DNA sequences), while Activators, Inhibitors, other TFs are proteins that bind to the DNA.

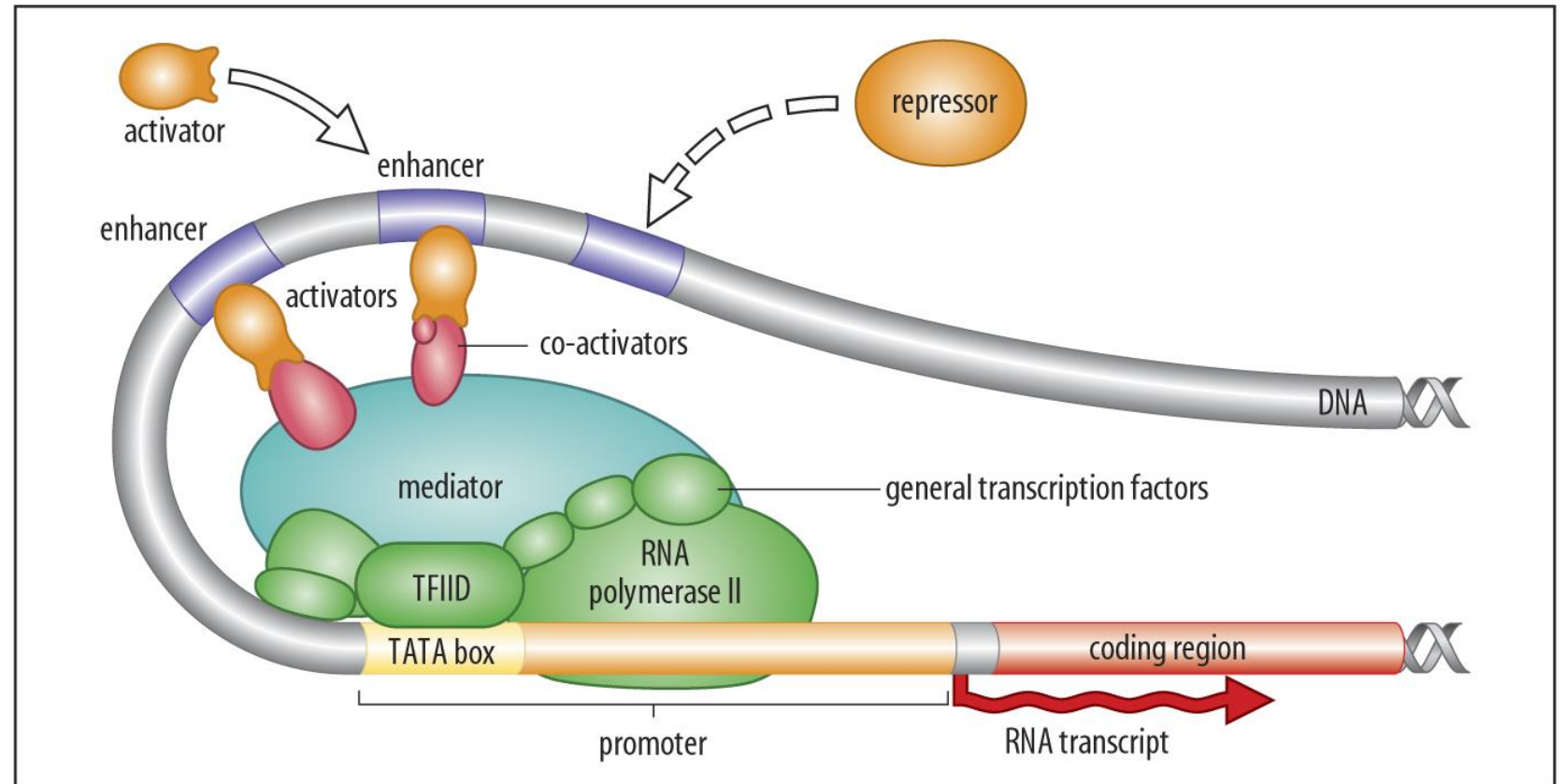
# Genome Organization



**A:** Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356), 737–738. <https://doi.org/10.1038/171737a0>

**B:** Genome Organization

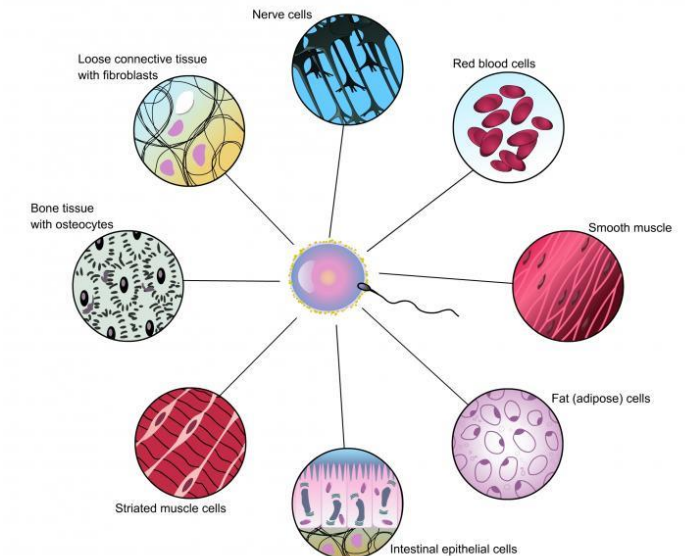
# Gene Regulation Landscape



# Cis-Regulatory Elements | Distinct Profiles

Parameters	Promoter	Enhancer	Insulator
Tentative Count	Tens of thousands	Millions	Thousands?
DNase I Hypersensitivity	-	-	-
TF Binding	RNAP, GTFs	TFs, Co-factors	CTCF, cohesion complex
Histone Modification	H3K4me3, H3K27ac, H3K9ac	H3K4me1, H3K27ac	Nucleosome Free

- Gene Expression is responsible for lineage specific programs.



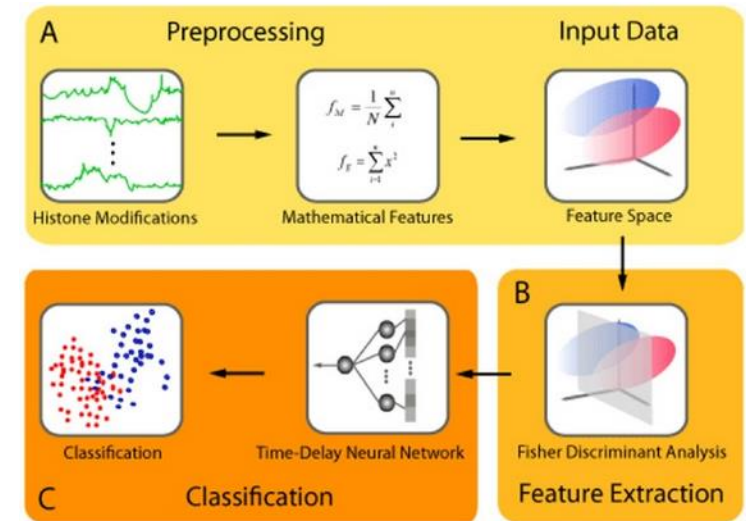
# Highlights

- Nearest promoter is assigned to the enhancer sequence.
- Authors propose Integrated Method for Predicting Enhancer Targets (IM-PET)
  - Enhancer targets ~ Promoters
  - Predicted Enhancers (CSI-ANN) -> Target Promoters
- Stochastic approach | Statistical Predictor
- Four features (inferences from 2000 real and non-interacting EP pairs from published ChIA-PET data for K562 and MCF-7 cells)
- Promoter: 2.5 Kbp (2 Kbp upstream and 0.5 Kbp downstream TSS)
- Enhancer: 2 Kbp



# CSI-ANN

- Input: Histone Modification/ChIP-Seq Data
  - SNP/ Methylation Data
  - HeLa cell ENCODE data, Human CD4+ cell data
- Background: Random genomic loci, 10X
- Target: Prediction of enhancers





# Metadata

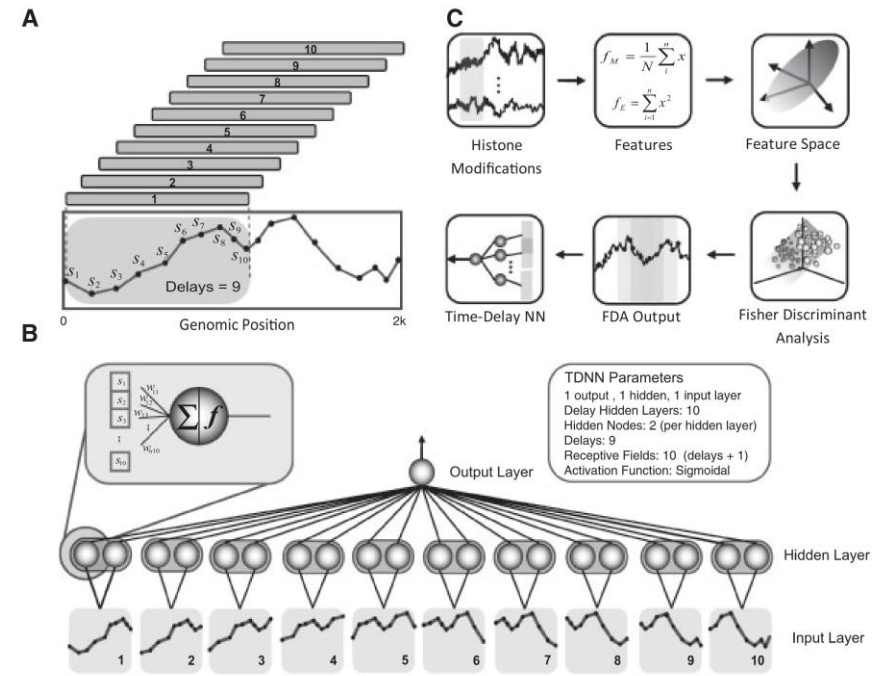
- HeLa Cell data (ENCODE)
  - 6 histone modification profiles (ChIP-ChiP data)
  - 74 training enhancers, 740 background sequences.
- CD4+ T cell data
  - P300 binding peaks- (i) < 1 kb (narrow)  
(ii) 2.5 Kb away from RefSeq TSS  
(iii) Overlapping signatures from the PReMod database  
(iv) 213 enhancers

\*Refseq is the sequence database from NCBI  
(<https://www.ncbi.nlm.nih.gov/refseq/>)

\*PReMod is a database of genome-wide cis-regulatory, predicted modules for human and mouse genomes.  
(<http://amp.pharm.mssm.edu/biotoolbay/tool/PReMod>)

\* P300 is a noted enhancer marker.

# CSI-ANN Framework



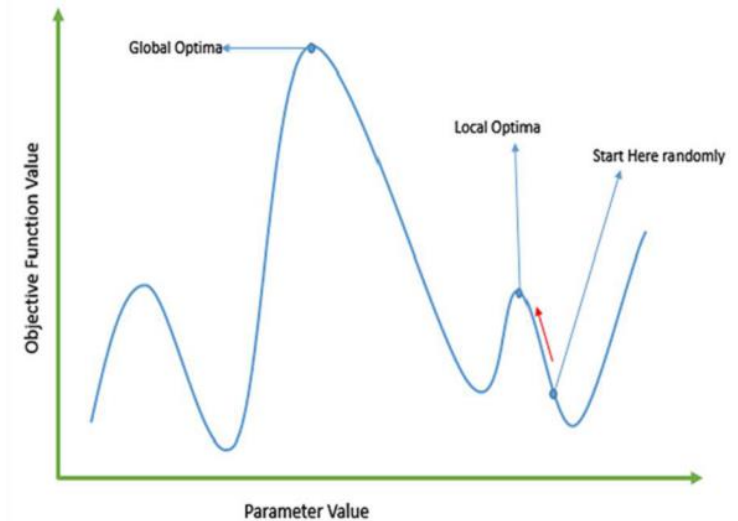
**Fig. 2.** Overview of the CSI-ANN framework. **(A)** An input data window with shifted sub-windows. **(B)** TDNN. A TDNN with one hidden layer and one output layer. Each time-copy layer (from 1 to 10) receive a portion (10 points or 1000 bp segment) of the entire 2 Kb window (20 points in total); each sliding window is denoted by a numbered light-gray bar. Each bar was slid one point for each time-copy (set of weights is the same for all time copies). **(C)** Flow chart of the CSI-ANN framework.

# CSI-ANN: Methodological Premises

- **Z-Score Normalization**
  - $X_{new} = \frac{x - \mu}{\sigma}$ , where  $\mu$  and  $\sigma$  are mean and standard deviation respectively.
- **Feature Value Calculation**
  - *Mean* |  $y_m^j = \frac{1}{n} \sum_i^n x_j(i)$
  - *Energy* |  $y_e^j = \sum_i^n x_j^2(i)$ , where  $x_j(i)$  is the preprocessed signal for histone modification type  $j$  at position  $i$ .
- **Bhattacharya Distance** (*gauges separation of two probability distributions*)
  - $D_B(p, q) = -\ln(BC(p, q))$ , where  $p$  and  $q$  are probability distributions, and
  - $BC(p, q) = \sum_{x \in X} \sqrt{p(x) * q(x)}$ , is the Bhattacharya coefficient.
- **Fisher Discriminant Analysis**
  - Strategy for dimensionality reduction
  - Identifying features that best separate two classes.
  - $S = \frac{\sigma_{between}^2}{\sigma_{within}^2}$ , where  $S$  is Fisher discriminant ratio, and  $\sigma^2$  is the variance.
- **Feature Values -> Z-transformation -> Fisher Discriminant Analysis**

# CSI-ANN: Methodological Premises (Contd.)

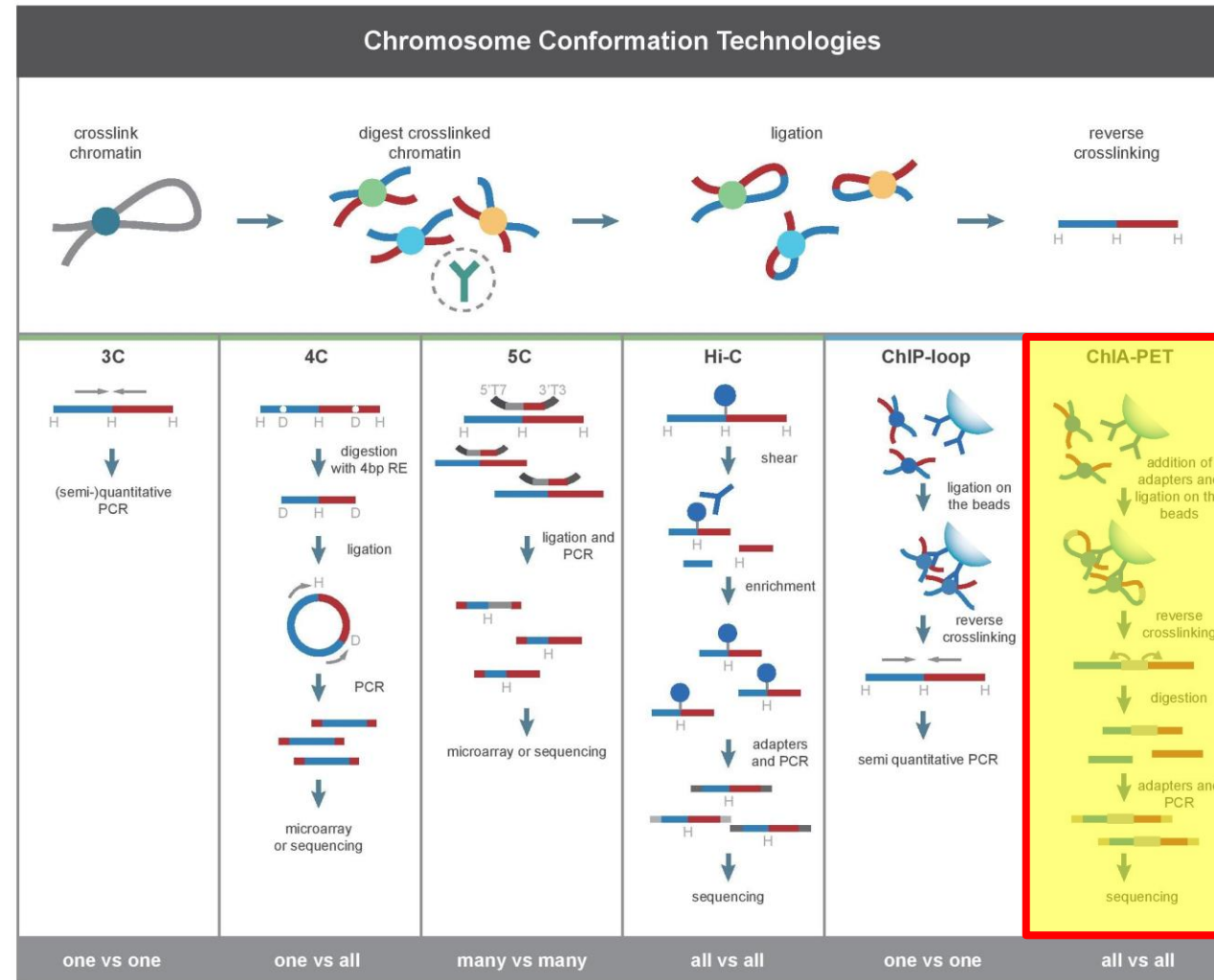
- **Particle Swarm Optimization**
  - Search optimization algorithm that was used to train time-delay neural network.
  - **Swarm:** a collection/ group/ cohort, **Particle:** an agent, individual in the swarm.
    - In the current context, particle is a solution, swarm is a subset of solutions in the solution space (n-dimensional).
  - **Local and Global optimum.**
  - Each particle is initialized with a  $x_i$  (personal best),  $y_i$  (global best), and  $v_i$  (velocity | direction)
  - The update equations:  $v_i(t+1) = wv_i(t) + c_1r_1(t)(y_i - x_i(t)) + c_2r_2(t)(\hat{y}_i - x_i(t))$   
 $x_i(t+1) = x_i(t) + v_i(t+1)$
  - The constraints:  $0 < c_1, c_2 < 2$ , where  $c_1, c_2$  are acceleration coefficients  
 $r_1(t)$  and  $r_2(t)$  are random values, with uniform distribution  $r(t) \sim U(0,1)$   
 $w$  is the weight coefficient that balances local and global bests.



# IM-PET | Essentials

- Using CSI- ANN, 208,342 enhancers in total predicted, averaging 17,362 enhancers per cell type.
- Identified 161,999 active promoters in these cell types using RNA-Seq data and GENCODE annotation of transcripts.
- Predicted 441,879 unique EP pairs across the 12 cell types, averaging 36,823 interactions per cell type.

# ChIA-PET



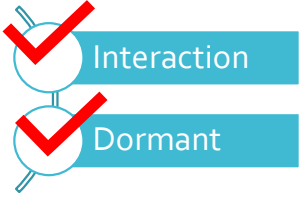
Liet al.:Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *BMC Genomics* 2014 15(Suppl 12):S11.

# Features | Scores

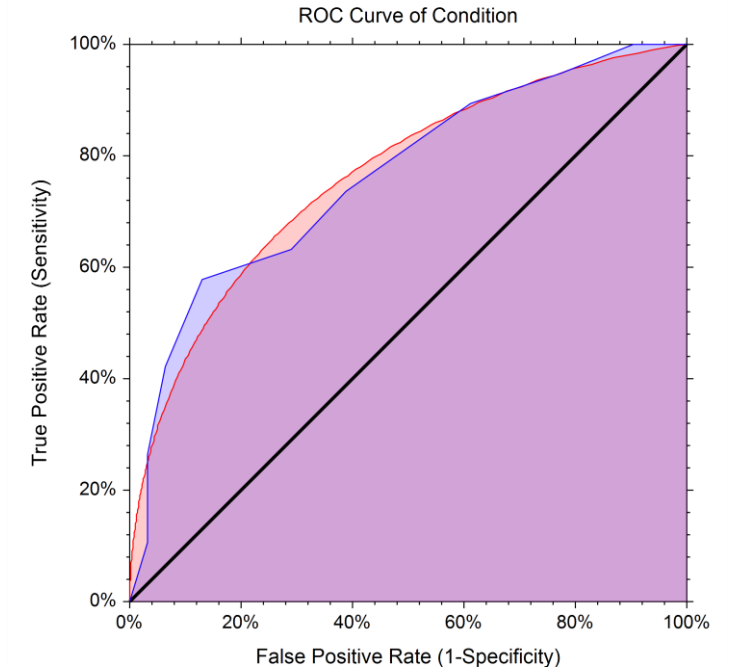
1. **E**nhancer and target **P**romoter activity profile **C**orrelation (EPC)
  1. FPKM values from RNA-seq data, for *promoters*.
  2. CSI-ANN algorithm, for *enhancers*.
2. **T**ranscription Factor and target **P**romoter **C**orrelation (TPC)
  1. Considers regulatory DNA sequences and TFs both
3. **COE**Volution of Enhancer and target Promoter (COEV)
  1. Sequence Similarity
  2. Conserved Synteny
4. **DIS**tant constraint between Enhancer and target Promoter (DIS)
  1. EP pairs are more like to occur over shorter distances- empirical data.
5. Preference of enhancers for certain classes of promoters
6. Existence of tethering elements in promoters that capture enhancers ( any mediator proteins)

# Performance Evaluation | IM-PET

- Random Forest
  - Classification Scheme of Machine Learning
  - Build over the concepts of Decision Trees, Bootstrapping, and Aggregation.

EPC	TPC	COEV	DIS	Class	
-	-	-	-	-	
...	...	...	...	...	
-	-	-	-	-	

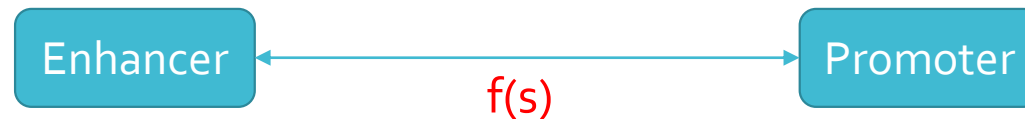
- *A predicted EP pair is considered to be true positive if the center of the enhancer in the predicted pair falls within one of the genomic regions of the gold-standard pair and the TSS in the predicted pair falls within the other genomic region of the gold-standard pair.*
- *Gold standard data is from any training or external data source.*





# IM-PET | Essentials (Contd.)

- *Positive training set (Interacting EP pairs)*
  - *Step 1:* Chromatin Interactions (K562, MCF-7 Human cells)
  - *Step 2:* CSI-ANN + 3 histone modifications (H3K4me1, H3K4me3, H3K27ac)
  - 2234 Enhancers!
- *Negative training set (Non-interacting EP pairs)*
  - Contact frequency (non-interacting genomic loci) | sites-separation distance
    - $f(s) = k \times s^{-\frac{3}{2}} \times e^{\frac{-1400}{s^2}}$ , where s denotes the sites-separation distance and k reflects the efficiency of the cross-linking reaction

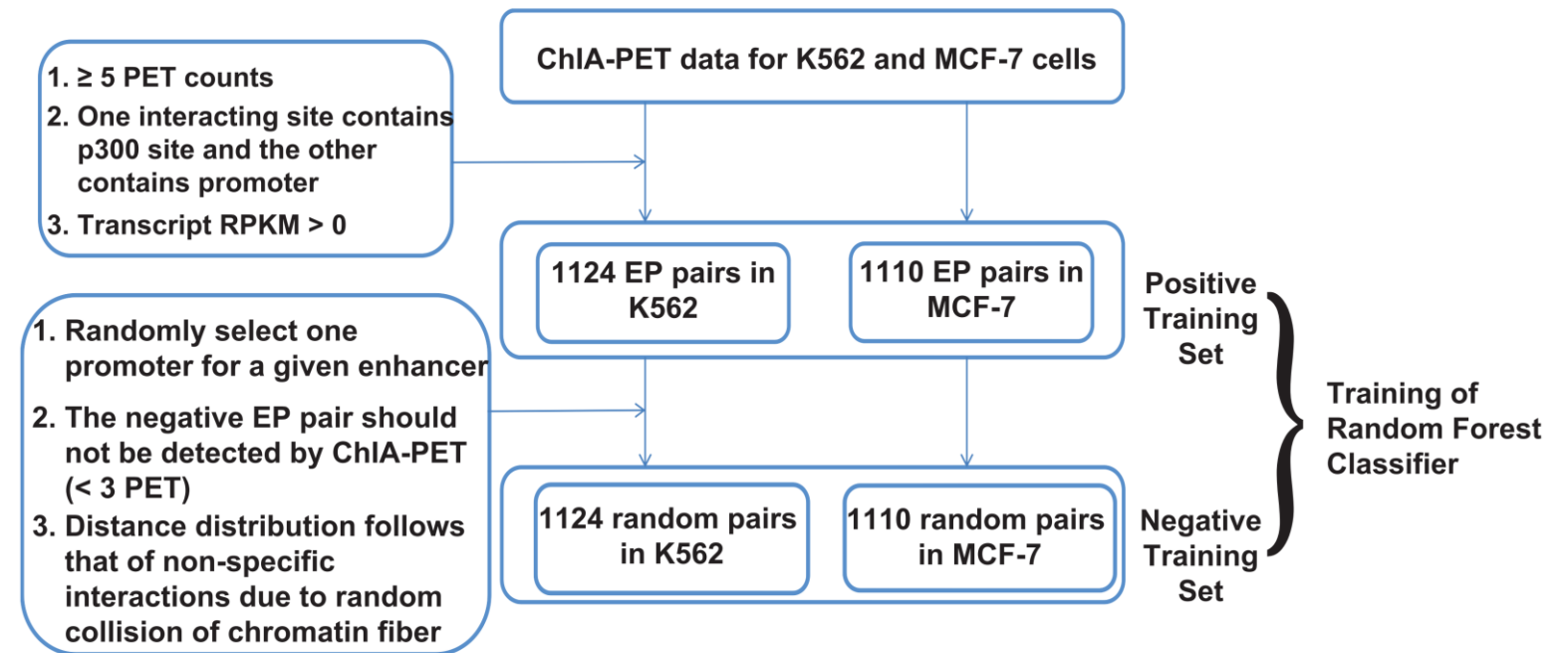


- *Step 1:* Select closest promoter from the enhancer based on  $f(s)$ .
- *Step 2:* If the promoter is undetected by ChIA-PET(<3 PET), add it to the list of possible EP pairs.
- *Step 3:* If false, find the closest promoter to the result from Step 2.
- 2234 EP (non-interacting) pairs!

# IM-PET | Essentials (Contd.)

## Supplementary Figures

**Fig. S1. Flow chart for the selection of training set of EP pairs and training of the Random Forest classifier.**

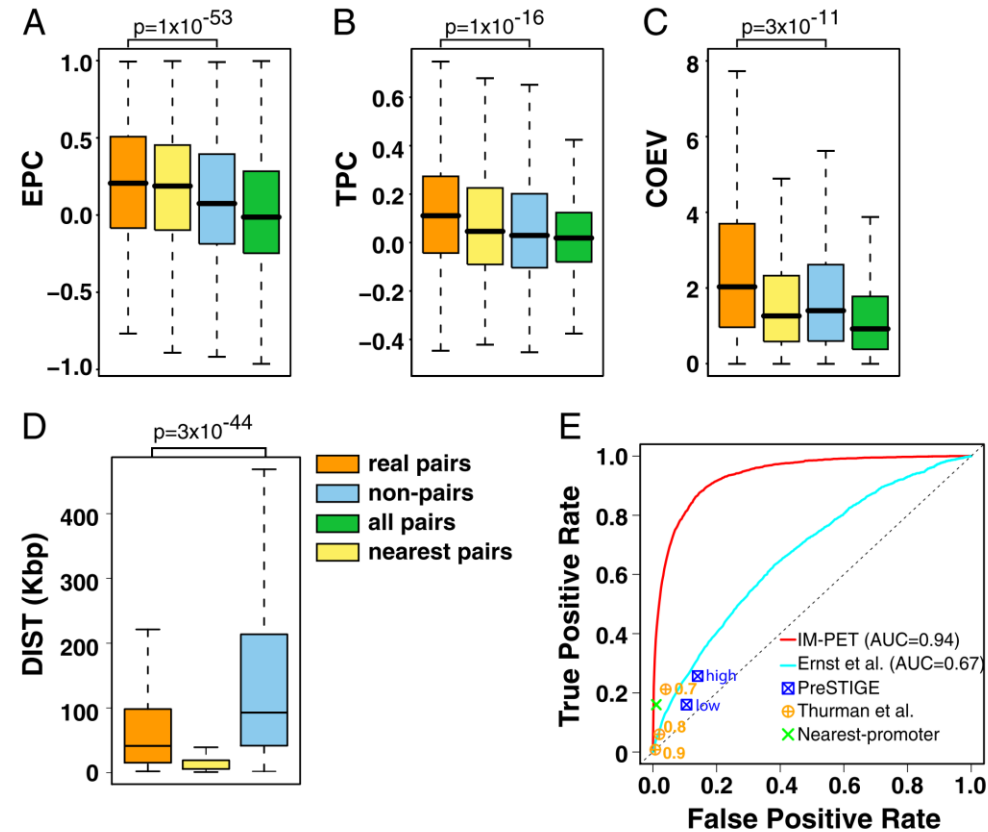


# Meta-analysis (EP pairs prediction)

IM-PET	Nearest Promoter	PreSTIGE	Ernst et al.	Thurman et al.
Random forest classifier with 4 features as defined.	-	pairing cell type-specific H <sub>3</sub> K <sub>4</sub> me <sub>1</sub> signals with genes that are specifically expressed in each cell type	histone modification profile correlation between nearest candidate pairs with 125-kbp distance	DNase I hypersensitive site (DHS) correlation of all candidate pairs within 500-kbp distance

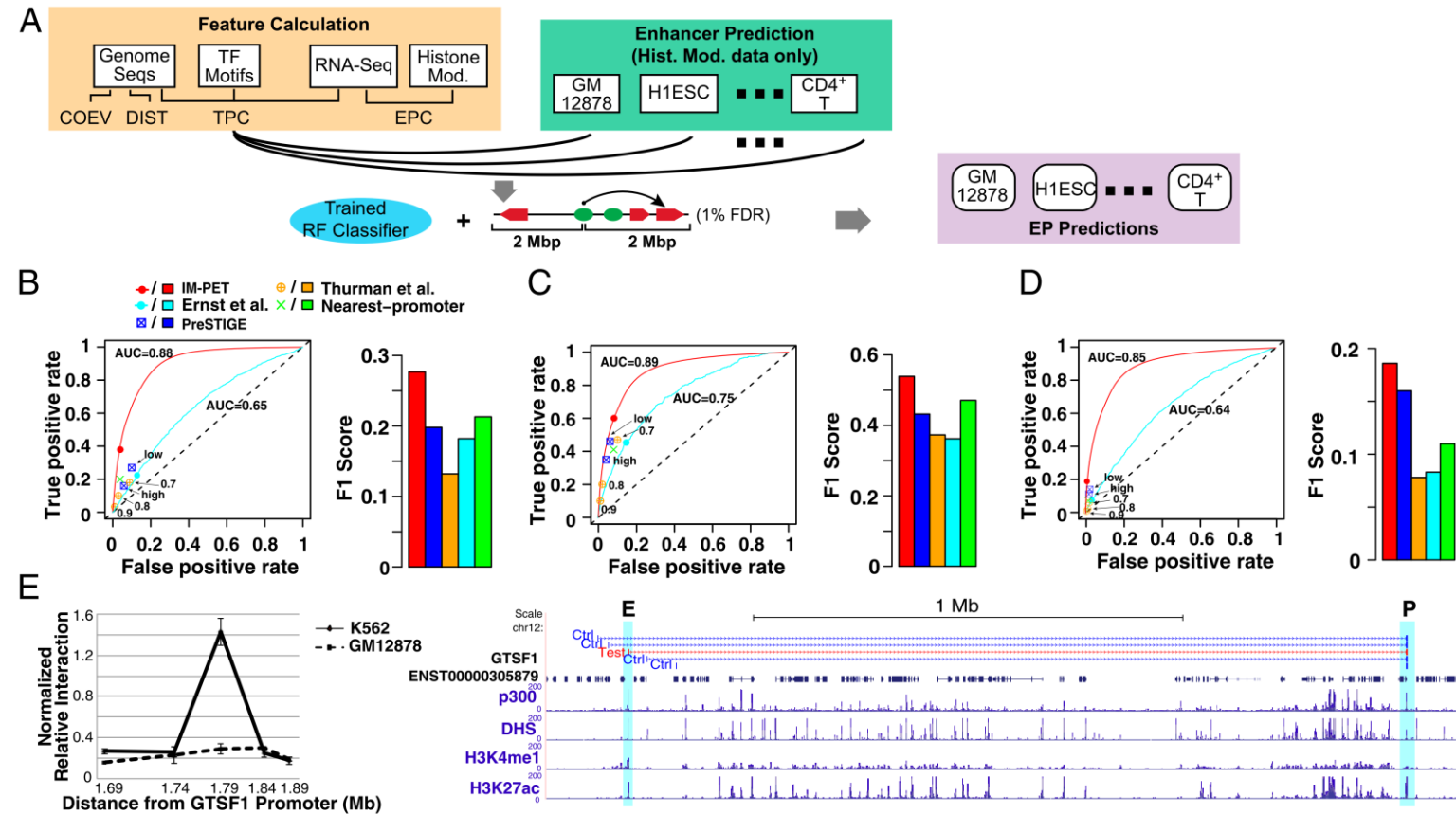
- AUC – **94%**; 27% higher than Ernst et al.
- Validation of Results (Verifying *Overfitting*)
  - Cross- validation of 5-fold
  - Logistic Regression, Support Vector Machine
  - Other Eukaryotes (*Drosophila melanogaster*)

# Features | Results Illustration



**Fig. 1.** Discriminative features and performance evaluation by cross-validation. (A) Enhancer and target promoter activity profile correlation (EPC); (B) TF and target promoter expression correlation (TPC); (C) coevolution of enhancer and target promoter (COEV); (D) distance constraint between enhancer and target promoter (DIS); “real pairs,” pairs selected using K562 and MCF-7 ChIA-PET data; “non-pairs,” noninteracting pairs according to ChIA-PET data; “all pairs,” EP pairs formed by extracting all promoters within 2 Mbp of an enhancer. “nearest pair,” EP pair in which the promoter is closest to the enhancer among all promoters in the genome. *P* values are based on one-sided Student *t* test. *n* = 2,234 for all tests. (E) ROC curve. Numbers next to circles indicate thresholds for predicting EP pairs using the Thurman et al. method. PreSTIGE made two sets of predictions: high- and low-confidence sets.

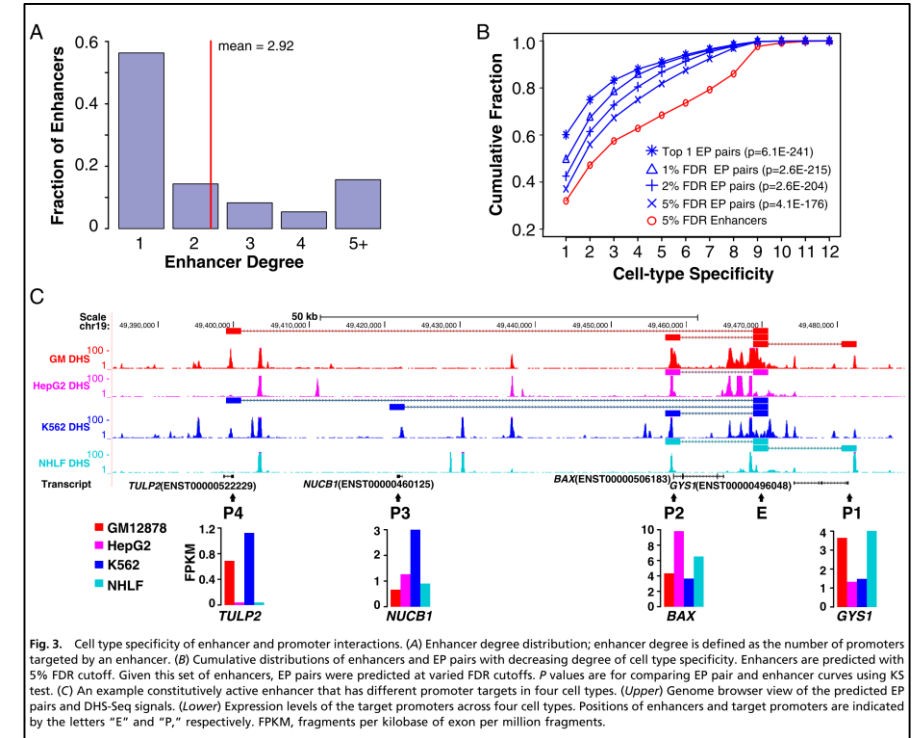
# IM-PET | Validation



**Fig. 2.** Genome-wide prediction and quality assessment of EP pairs in 12 cell types. (A) Schematic diagram for making genome-wide EP predictions using the IM-PET method. ROC curve and F1 score using additional ChIA-PET EP pairs (B), Hi-C EP pairs (C), and eQTL-gene pairs (D) as the gold-standard sets, respectively. F1 score is the harmonic mean of precision and recall. Numbers next to circles indicate thresholds for predicting EP pairs using the Thurman et al. method. PreSTIGE made two sets of predictions: high- and low-confidence sets. (E) 3C-qPCR validation of a predicted EP pair involving the transcript ENST00000305879 of the gene *GTSF1*. The EP pair is predicted in K562 but not in GM12878. The following tracks are shown from Top to Bottom: 3C-qPCR primer positions for negative controls (blue) and test (red) interactions; Refseq gene and transcript IDs (black) of the locus being tested; p300 ChIP-Seq peak; DHS ChIP-Seq peak; H3K4me1 ChIP-Seq peak; H3K27me3 ChIP-Seq peak. E, enhancer; P, promoter.

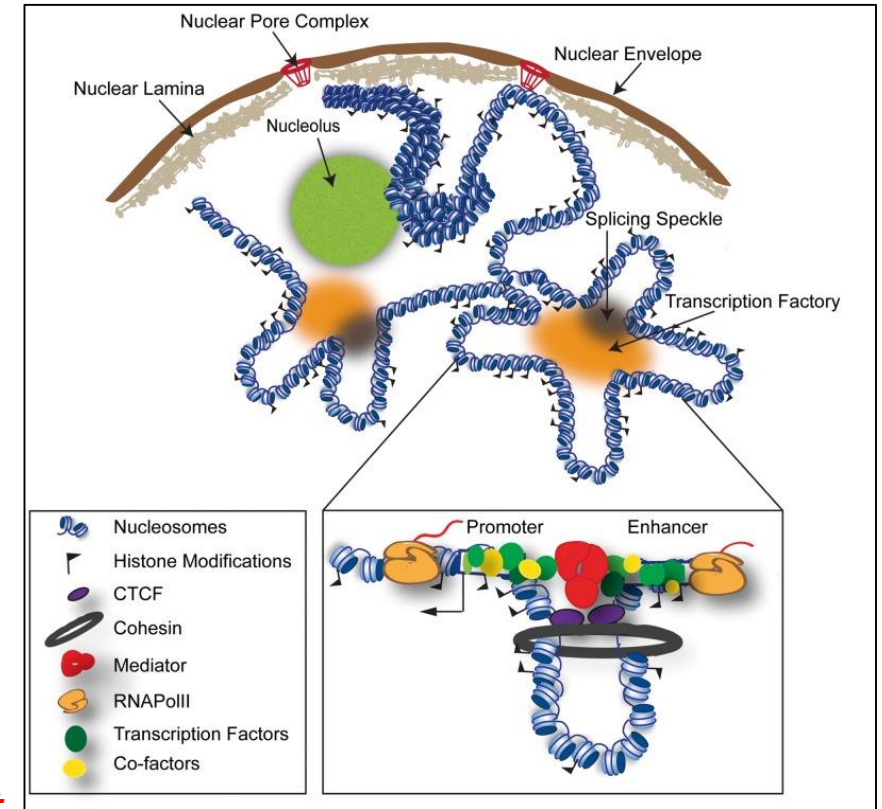
# IM-PET | Other revelations

- *EP Interactions Have Higher Cell Type Specificity than Enhancers.*
  - **2.92 promoters for every enhancer.**
- *Promoters with High Expression Specificity Are Regulated by Multiple Enhancers That Have Lower Conservation Levels.*
  - Degree of a promoter and its expression specificity are highly correlated.
  - GO term analysis indicates that promoters controlled by three or more enhancers are more enriched in cell type-specific terms.
  - significant negative correlation between enhancer sequence conservation and the target promoter degree.



# IM-PET | Other revelations (Contd.)

- *Cohesin Mediates Chromatin Loop Formation and Regulates Cell Type-Specific Gene Expression in the Absence of CTCF.*
  - ChIP-Seq of Cohesin-CTCF sites in the DNA that overlapped EP pairs.
  - Cohesin sites significantly overlapped the predicted EP pairs; and not CTCF sites.
  - Enhancers, Promoters at such sites have higher cell-type specificity.
  - TFs overlapping cohesin (not CTCF) sites show significantly higher expression specificity than those overlapping cohesin (and CTCF) sites.
  - **cohesin can mediate chromatin looping without the involvement of CTCF**



<https://doi.org/10.1093/bib/bbv097>

# Take Home Points

- EP interactions are highly expression/cell-specific
- Distal interactions are mostly established
- Training data is usually very sparse (few 1000 positives) and features are often cell-type specific.



# Thank You

- This presentation is available at <https://github.com/shauryajauhari/Presentations>