STAT1 regulates microRNA transcription in interferon γ – stimulated HeLa cells

Guohua Wang[1], Yadong Wang[1], Denan Zhang[1], Mingxiang Teng[1,2], Lang Li[2], and Yunlong Liu[2]

Harbin Institute of Technology, Harbin, Heilongjiang, China
Indiana University School of Medicine, Indianapolis, IN, USA

THE CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
THE CENTER FOR MEDICAL GENOMICS

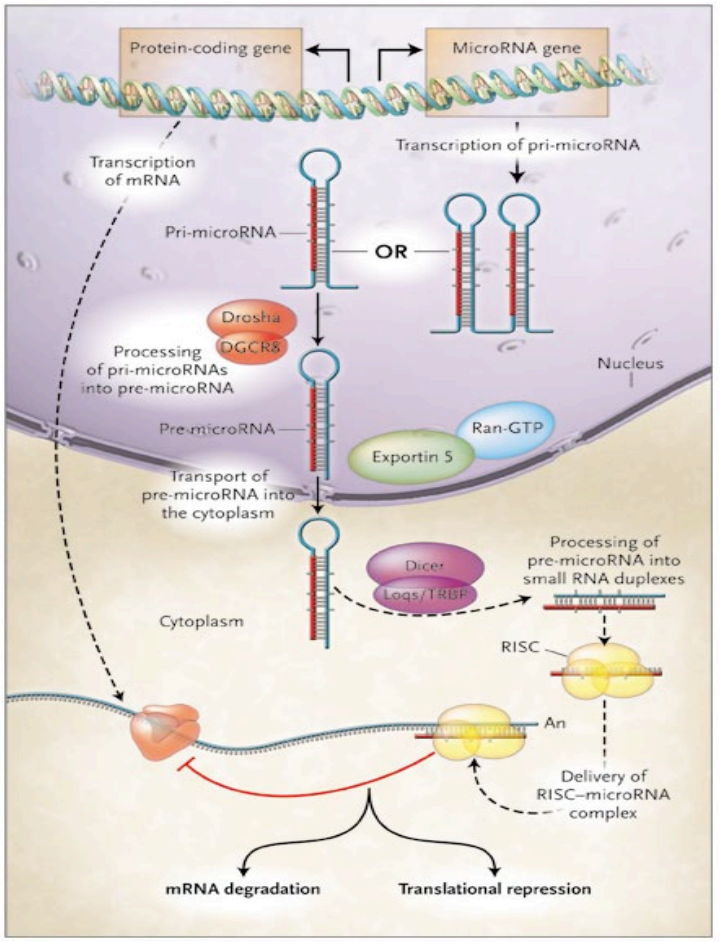
Outline

- Introduction – microRNA
- Motivation
  - A failed attempt to construct microRNA-mediated regulatory network
- Strategy and Modeling
- Results
- Conclusion
- Acknowledgement

microRNA regulation

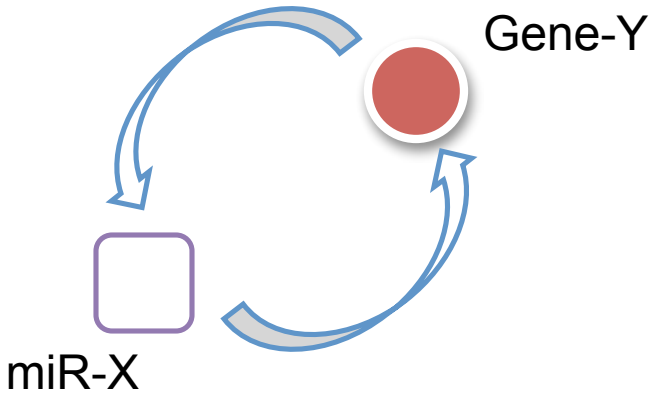
Small non-coding RNA

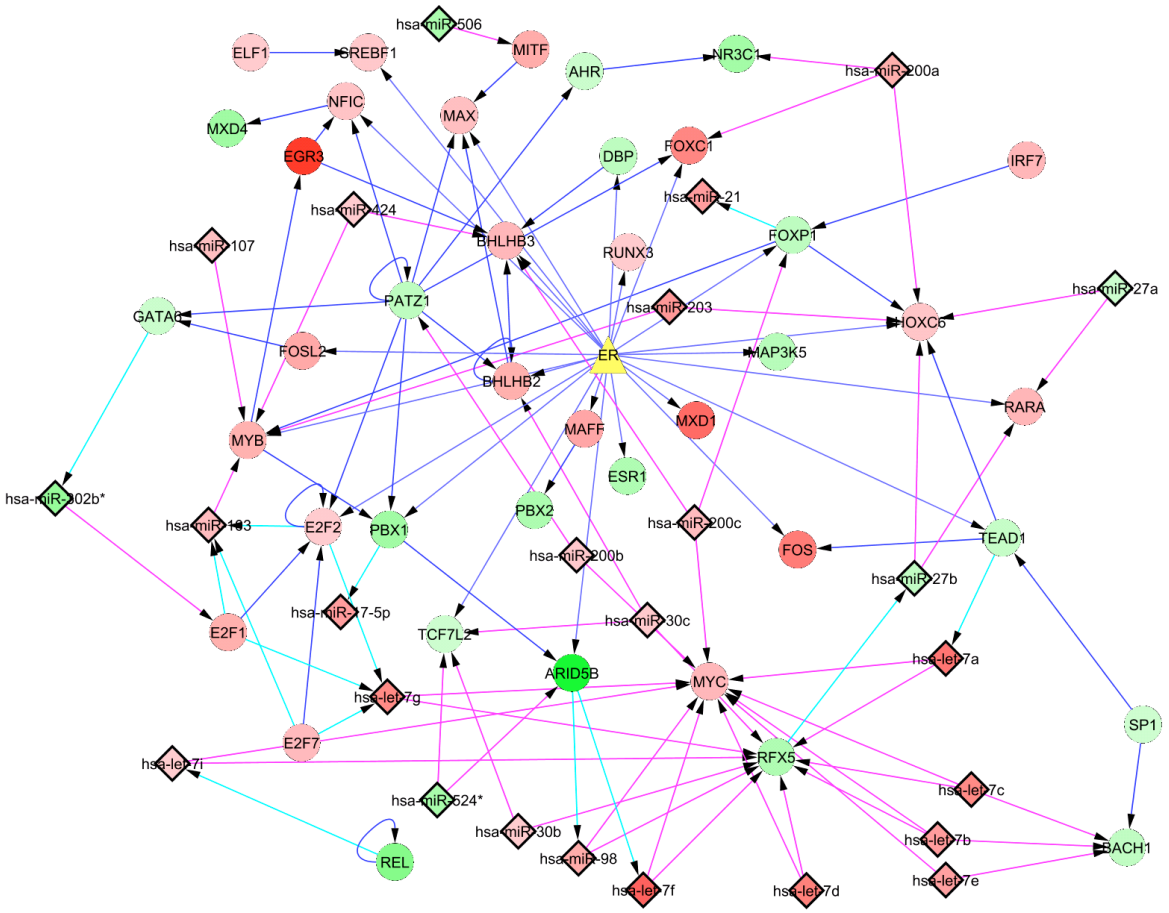
- Function: **silencing genes** through post-transcriptional regulation
- Single strand RNA (ssRNA);
- 19-25 nucleotides (~22 nucleotides);
- Endogenous
- Accounting for more than 1% of the genome (>200 members per species, >800 in human);
- >2/3 of human genes are microRNA target;



Motivation – A failed attempt

Initial goal: identify microRNA regulatory network using microRNA microarray and gene expression data





In collaboration with Dr. Nakshatri
Manuscript submitted

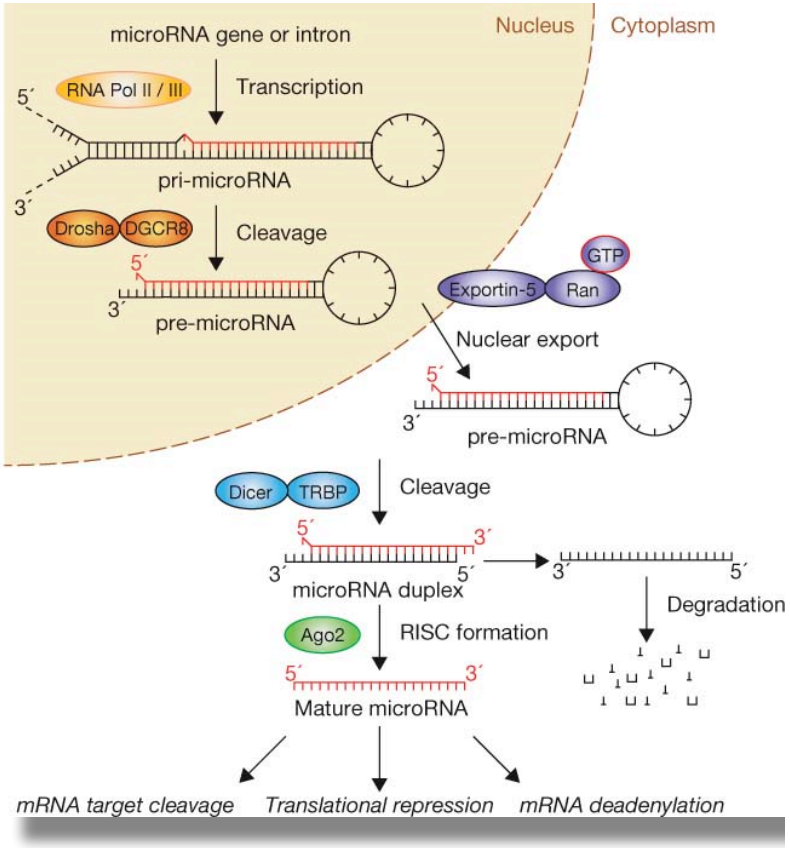
Motivation

microRNA biogenesis

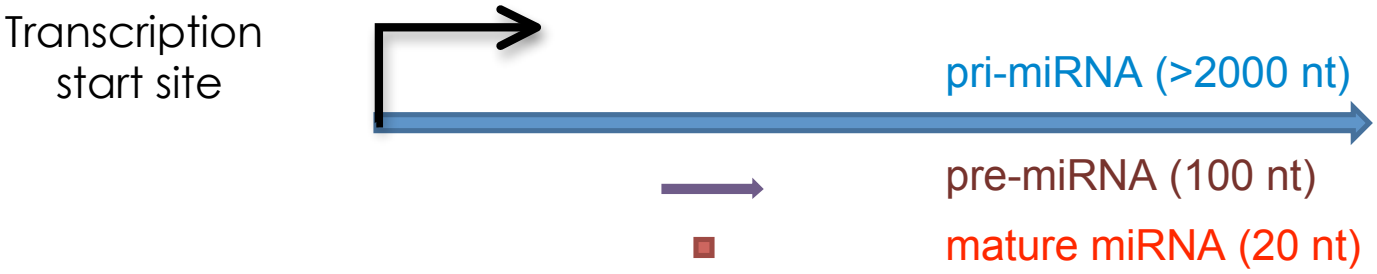
- Three forms:
  - Pri-miRNA (several thousands nt)
  - Pre-miRNA (60-100 nt)
  - Mature miRNA (22 nt)

- Most microRNAs are Pol II trascribed

- We only have pre- and mature miRNA annotation



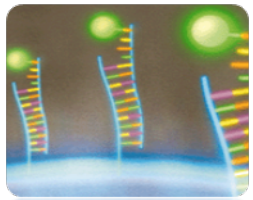
- Can we identify regulatory regions and transcription start site based on Pol II binding patterns?



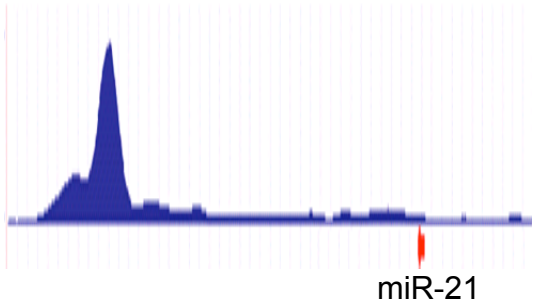
Strategy

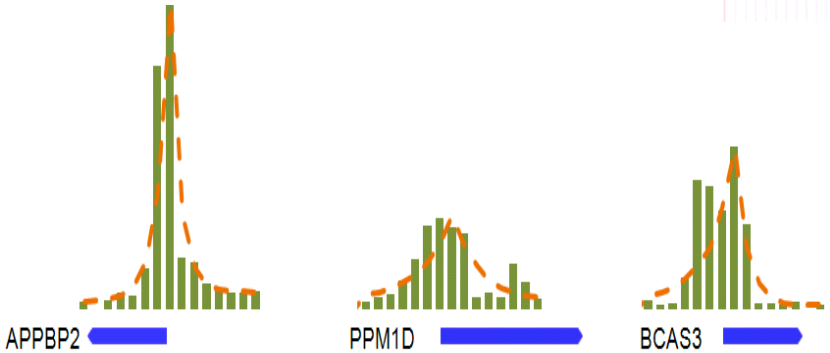
Identify Pol II binding pattern from expressed genes, and apply this pattern to identify microRNA promoters.

ChIP-seq data on Pol II



Similar pattern for miRNA?





Pol II binding feature surrounding highly expressed genes

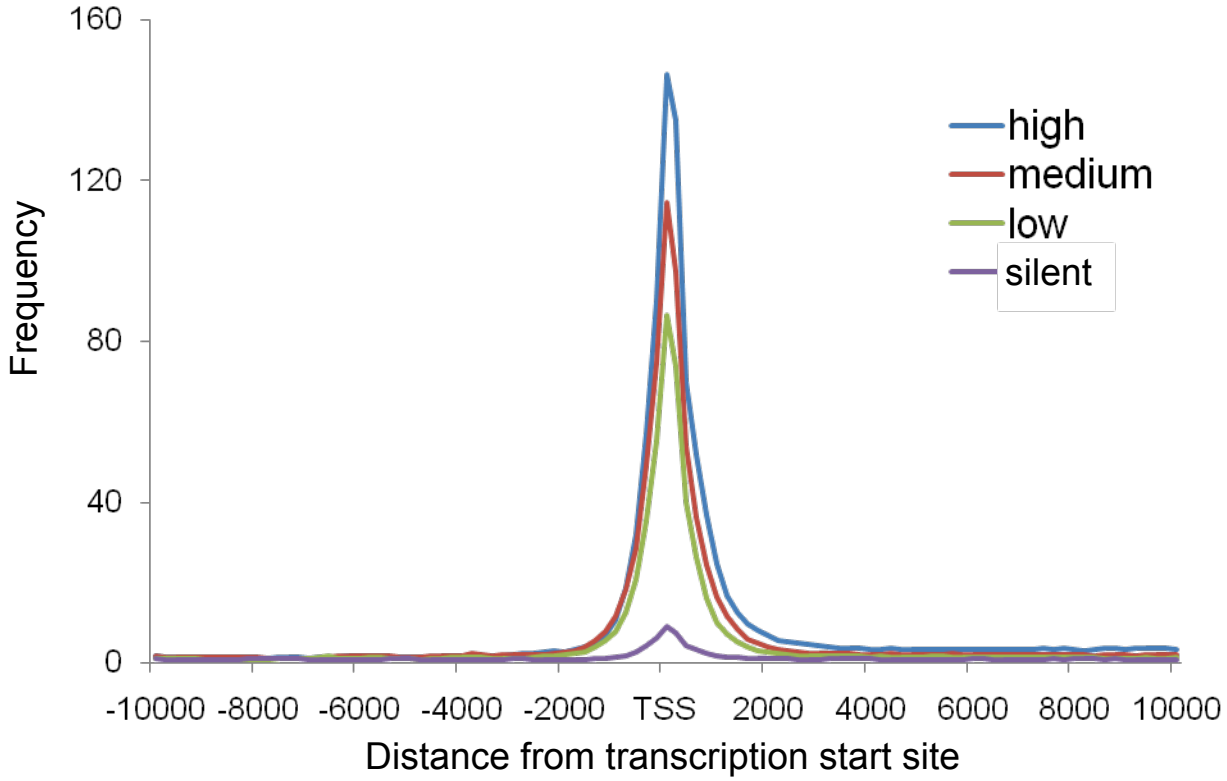
Learn from coding genes

Goal:

- Confirm whether Pol II binding follows certain pattern for high- and low- expressed genes;
- If there is a clear pattern, what does it look like.

Strategy:

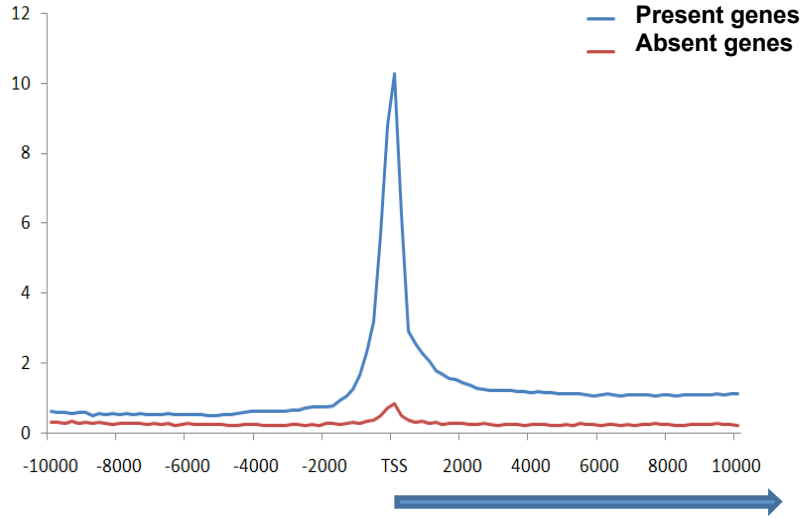
- Microarray data in HeLa cell (GSE 4483, Mense et al. Physiol Genomics 2006)



A reasonable model

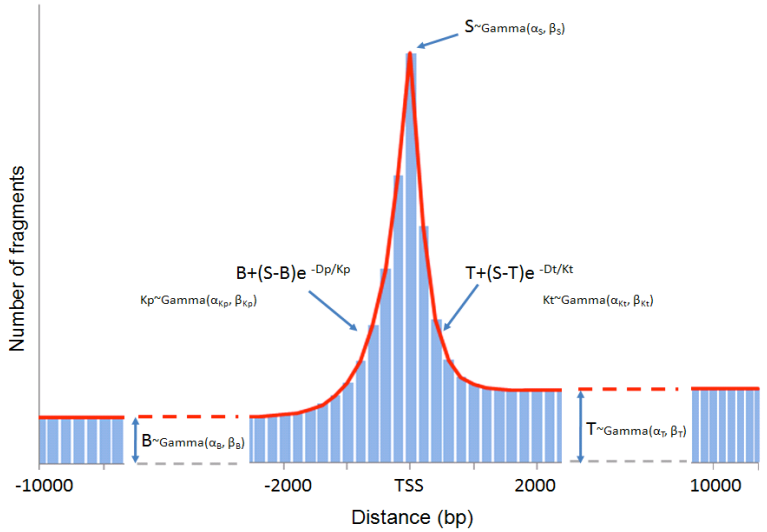
Key features:

- A spike around TSS (transcription start site);
- In both promoter region and transcript region, the Pol II signals gradually go down;
- The signals in transcript region is higher than in the background region.



Things considered in the equation:

- Five important parameters for each gene:
  - Signal Intensity in/on
    - S: Transcription Start Site (TSS);
    - B: Background region;
    - T: Transcript region;
  - Decay rate
    - Kp: Promoter region
    - Kt: Transcript region
- Each parameter follows *gamma* distribution genome-wide.



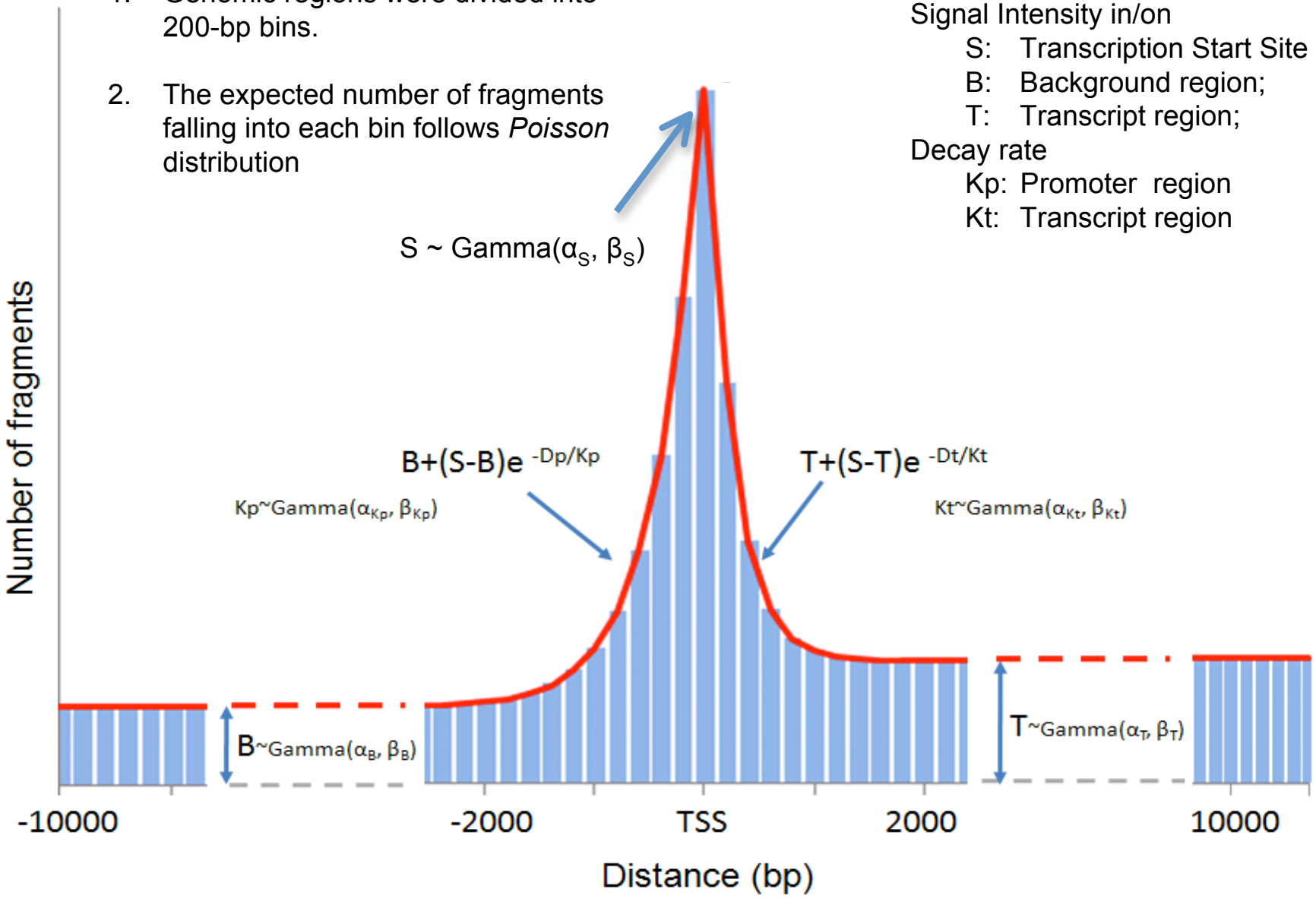
A closer look at the model

1. Genomic regions were divided into 200-bp bins.

2. The expected number of fragments falling into each bin follows *Poisson* distribution

For each gene, 5 parameters decide the expected # of fragments (λ in Poisson)

- Signal Intensity in/on
  - S: Transcription Start Site (TSS);
  - B: Background region;
  - T: Transcript region;
- Decay rate
  - Kp: Promoter region
  - Kt: Transcript region

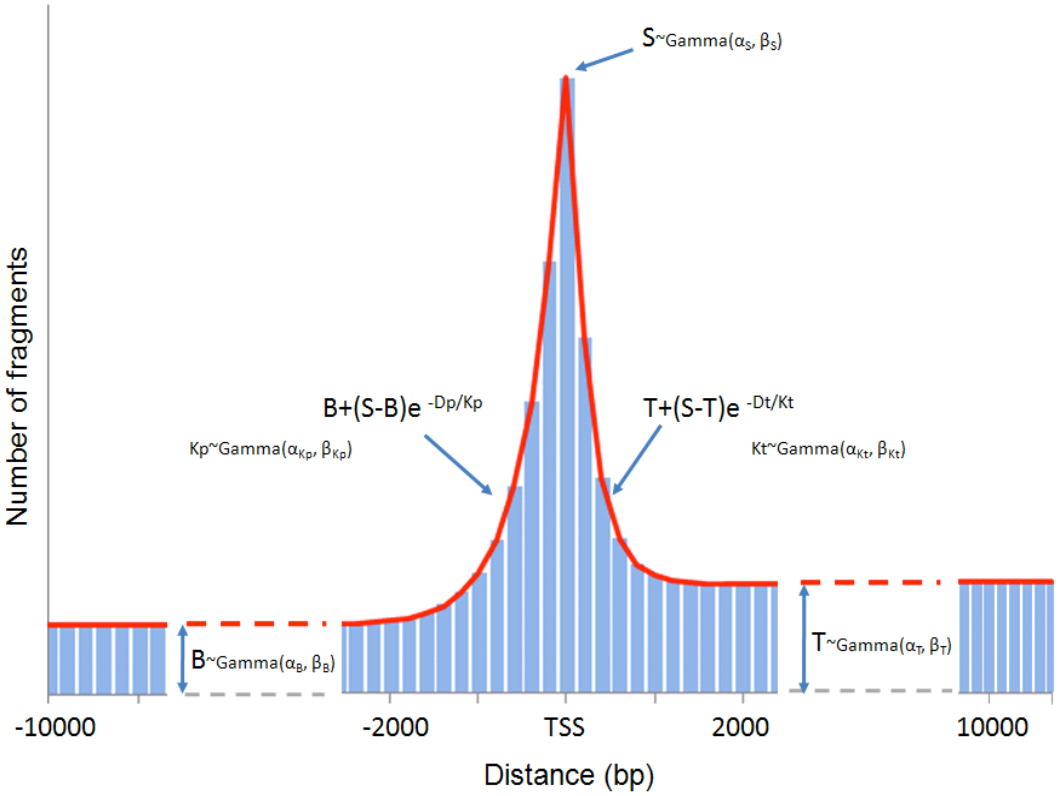


Modeling

X: observed Pol II fragments in each bin

Y: {B, T, S, Kp, Kt} – missing data

θ : $\{\alpha_B, \beta_B, \alpha_T, \beta_T, \alpha_S, \beta_S, \alpha_{Kt}, \beta_{Kt}, \alpha_{Kp}, \beta_{Kp}\}$



$$\Pr(X, Y \mid \theta) = \Pr(X \mid Y, \theta)\Pr(Y \mid \theta)$$

$$= \prod_{i=1}^{N_{GENE}} \left(\underbrace{\prod_{j=1}^{50} \frac{e^{-\lambda_{P_{ij}}} \lambda_{P_{ij}}{}^{X_{P_{ij}}}}{X_{P_{ij}}!}}_{\text{upstream bins}} \cdot \underbrace{\prod_{j=1}^{50} \frac{e^{-\lambda_{T_{ij}}} \lambda_{T_{ij}}{}^{X_{T_{ij}}}}{X_{T_{ij}}!}}_{\text{downstream bins}} \cdot \underbrace{\frac{e^{-\lambda_{TSS_i}} \lambda_{TSS_i}{}^{X_{TSS_i}}}{X_{TSS_i}!}}_{\text{TSS bins}} \cdot \underbrace{\frac{K_{Pi}{}^{a_1-1} e^{-K_{Pi}/\beta_1}}{\Gamma(\alpha_1)\beta_1^{\alpha_1}} \cdot \frac{K_{Ti}{}^{a_2-1} e^{-K_{Ti}/\beta_2}}{\Gamma(\alpha_2)\beta_2^{\alpha_2}} \cdot \frac{B_i{}^{a_3-1} e^{-B_i/\beta_3}}{\Gamma(\alpha_3)\beta_3^{\alpha_3}} \cdot \frac{E_i{}^{a_4-1} e^{-E_i/\beta_4}}{\Gamma(\alpha_4)\beta_4^{\alpha_4}} \cdot \frac{S_i{}^{a_5-1} e^{-S_i/\beta_5}}{\Gamma(\alpha_5)\beta_5^{\alpha_5}}}_{\Pr(Y|\theta)} \right)$$

Are these models working?

Goal: evaluating whether the proposed model (Pol II binding pattern around TSS based on ChIP-seq) can identify coding genes that are actively transcribed (array-derived)

Gold standard (using genes with clear annotation – no other genes within 10,000-bp):

•Positive set:
- 4,120 genes (Present on array)

•Negative set:
- 2,682 genes (Absent on array)

Procedure

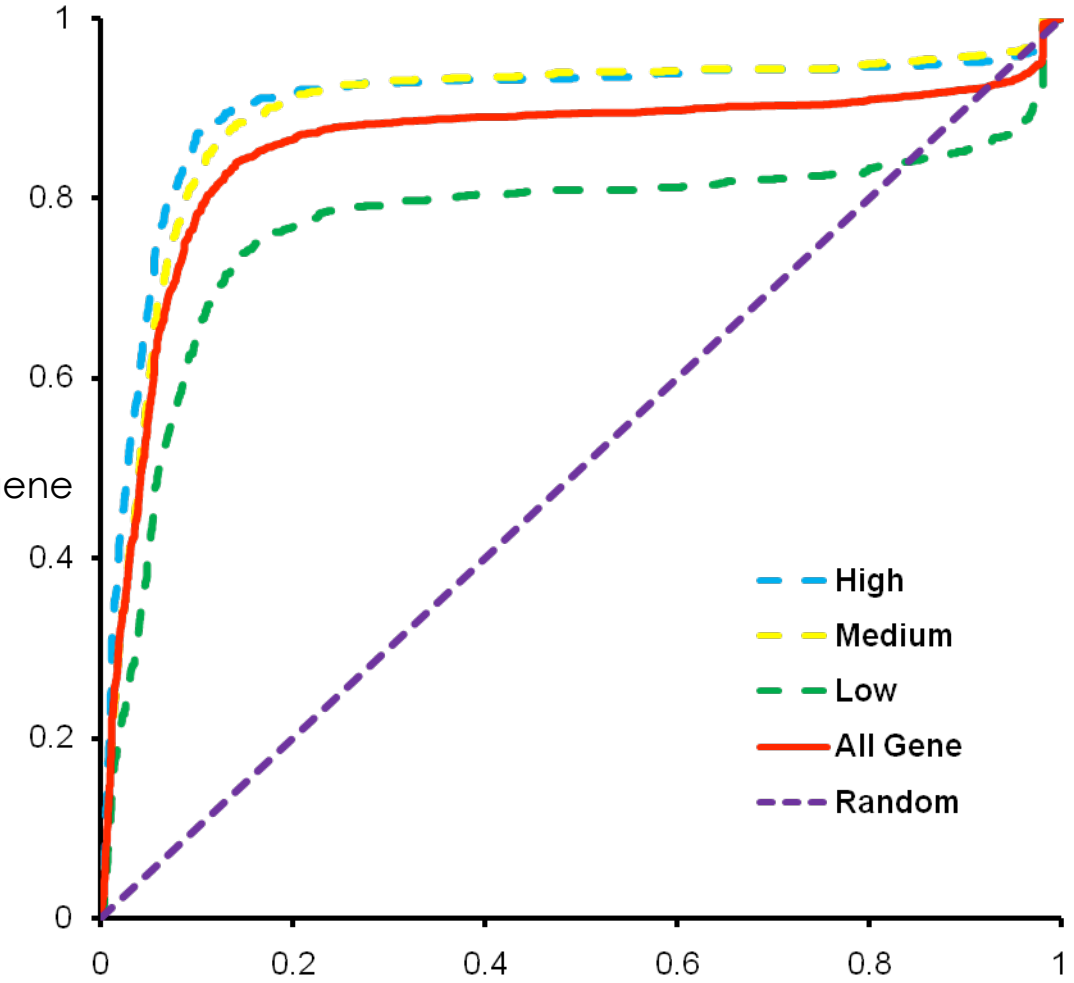
•1/4 were used to train parameters

•3/4 were used to make the call

•ROC curve are generated based upon correct prediction on gene expression.

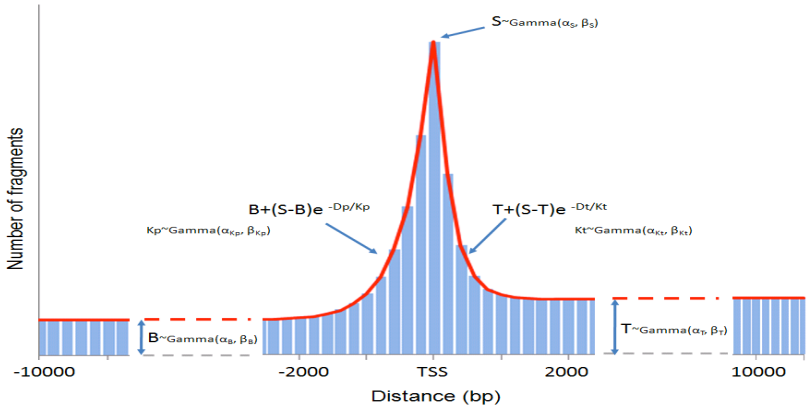
Conclusion

•The model worked well, at least for protein coding genes

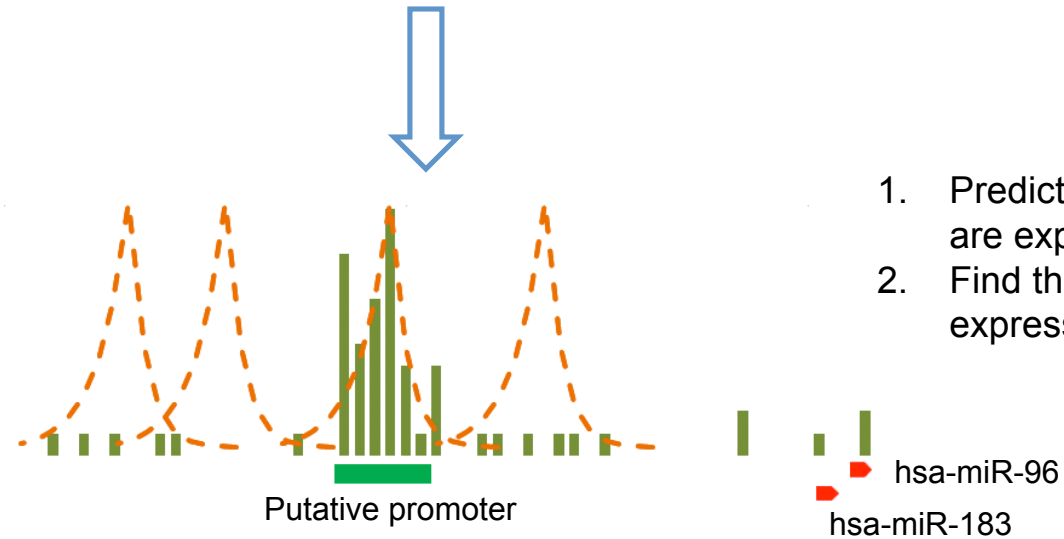


Promoter identification for miRNAs

- Strategy



Parameters identified for coding genes



1. Predict whether certain miRNAs are expressed
2. Find the promoter regions for expressed miRNAs

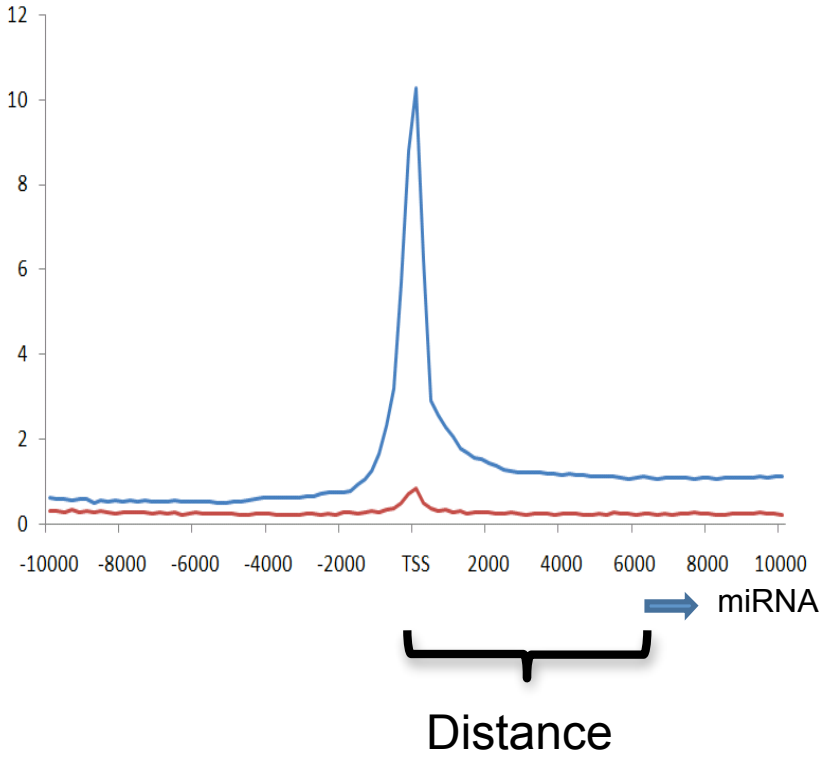
Results

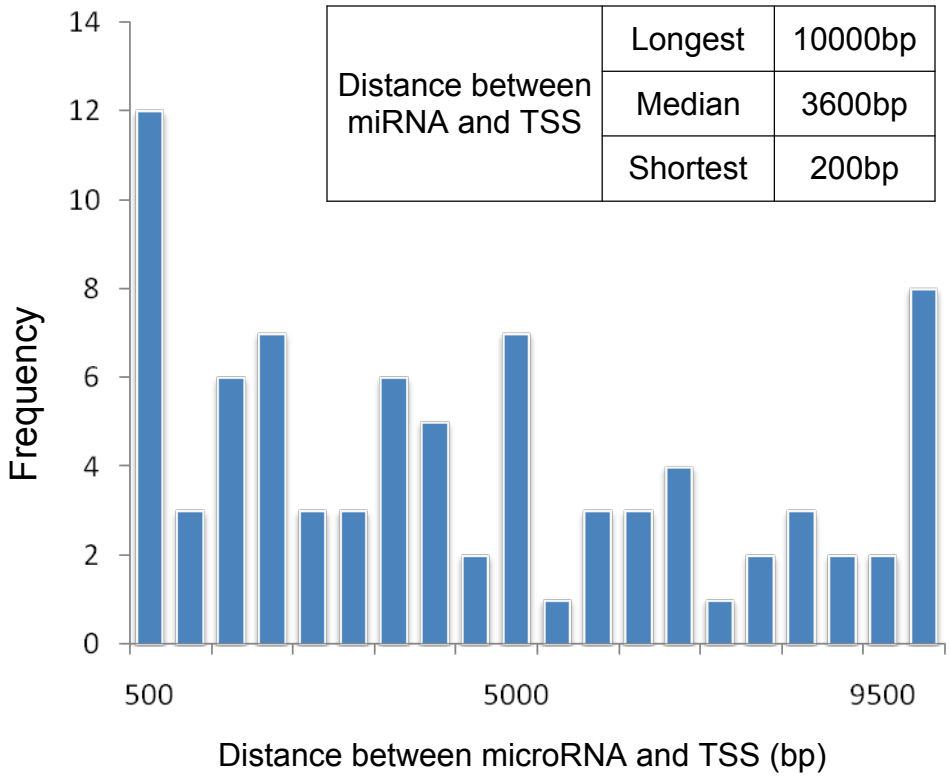
Result summary

- Focusing on 419 annotated intergenic microRNAs;
- Computational model identified 83 microRNAs predicted to be transcribed in HeLa cells (FDR ≤ 20%);
- This is consistent with RNA-seq experiments on another project.

Distance between microRNA and TSS

- Distance between annotated microRNA and transcriptions start site varies



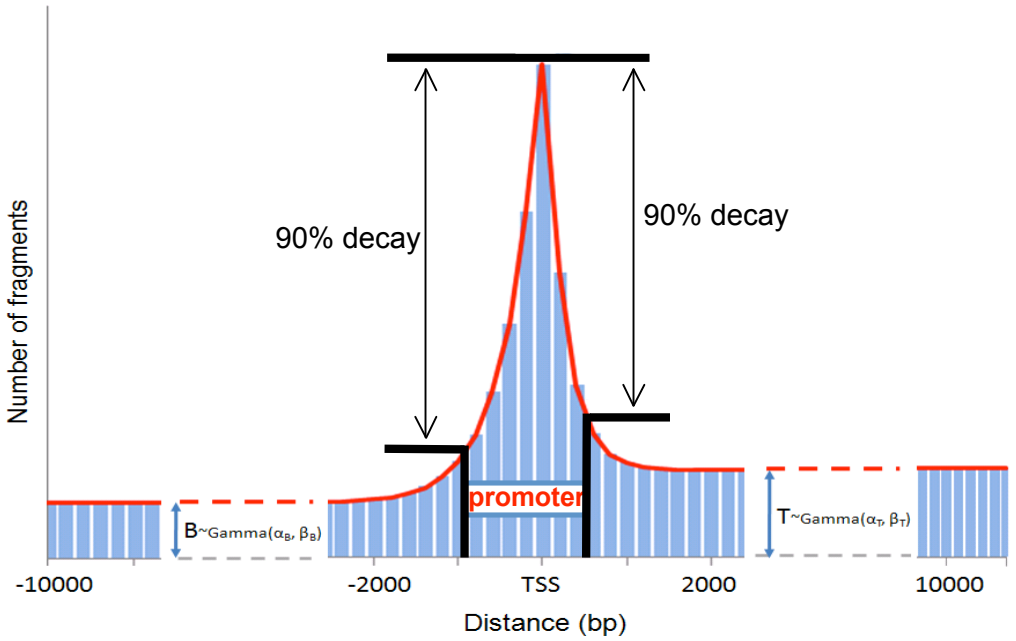


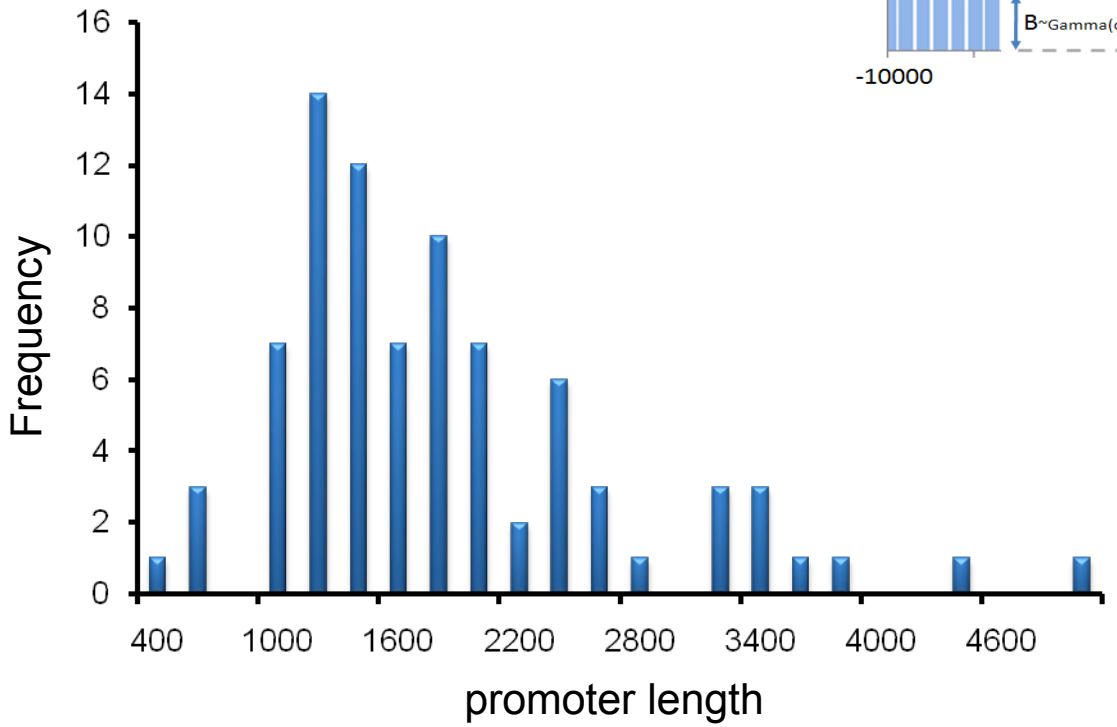
Distance between miRNA and TSS	Longest	10000bp
	Median	3600bp
	Shortest	200bp

Promoter length

- Promoter length varies.

Promoter length	Longest	4,989-bp
	Median	1,476-bp
	Shortest	397-bp

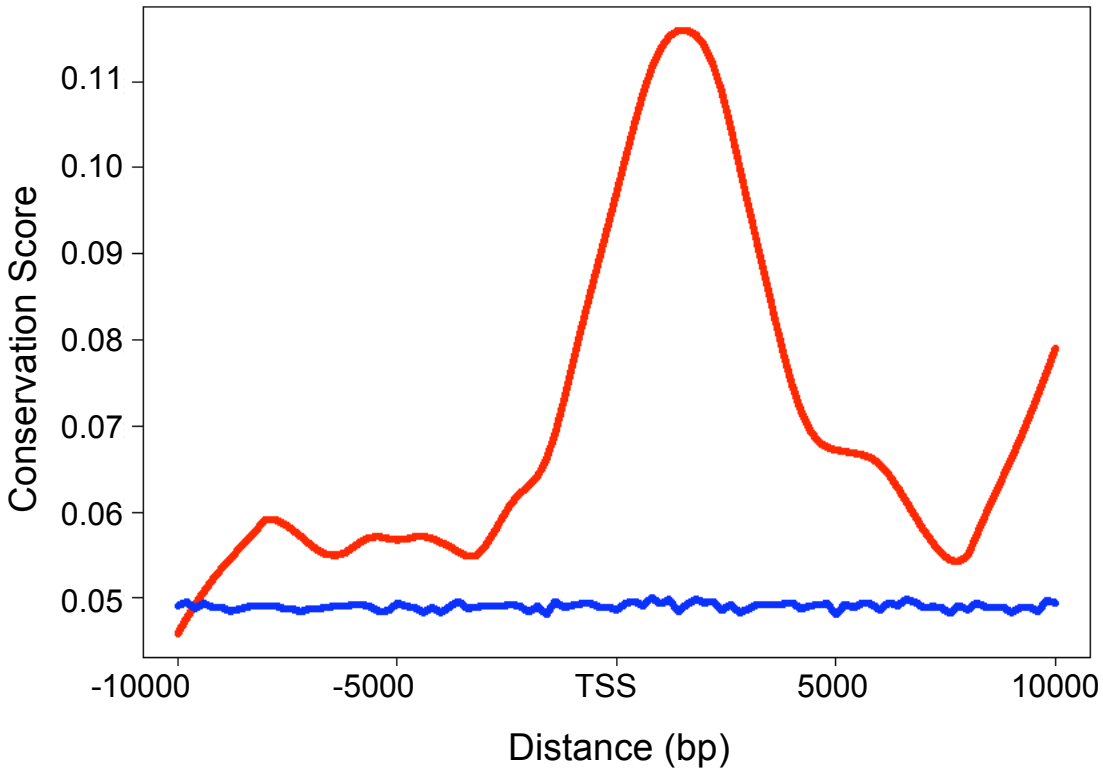




Validation

Validation I – conservation

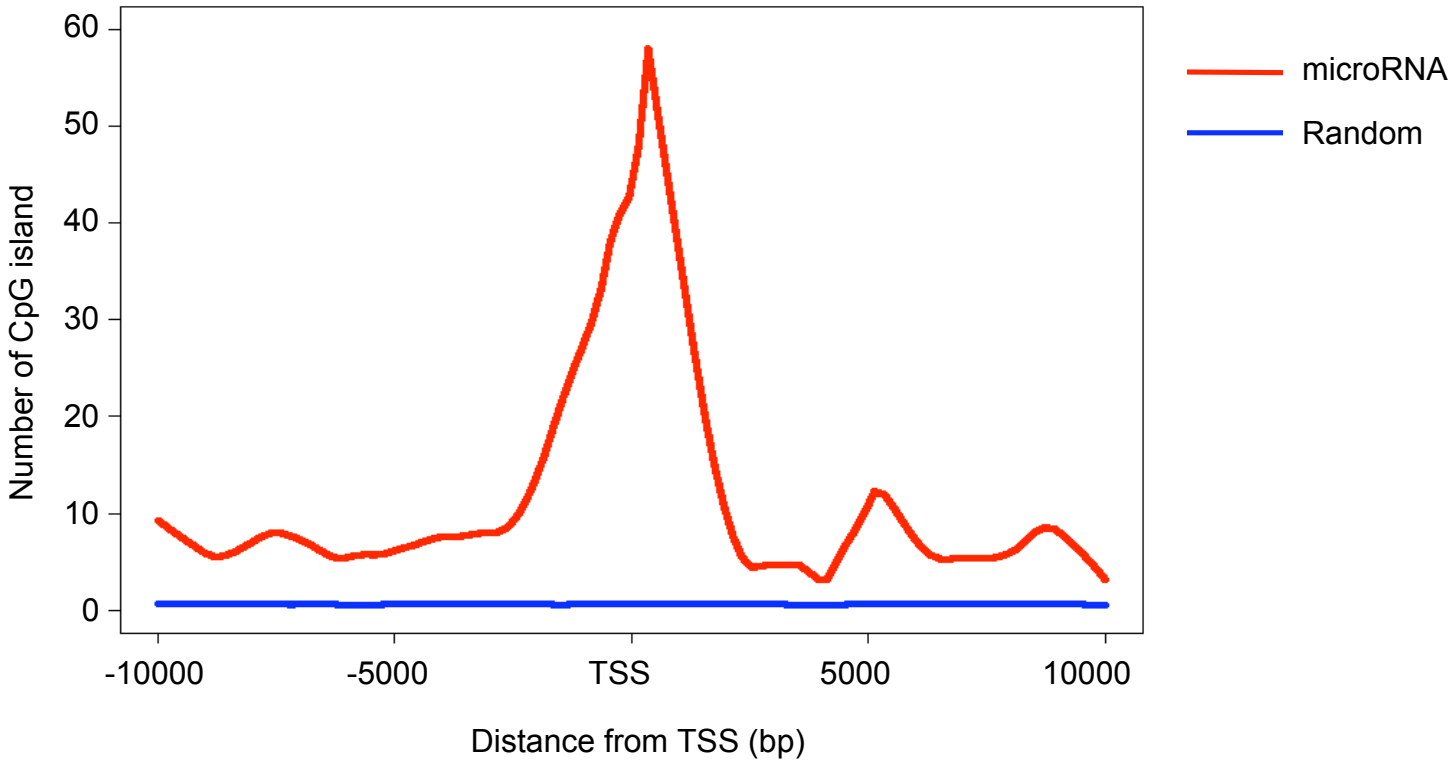
- Higher conservation score is observed around transcription start site



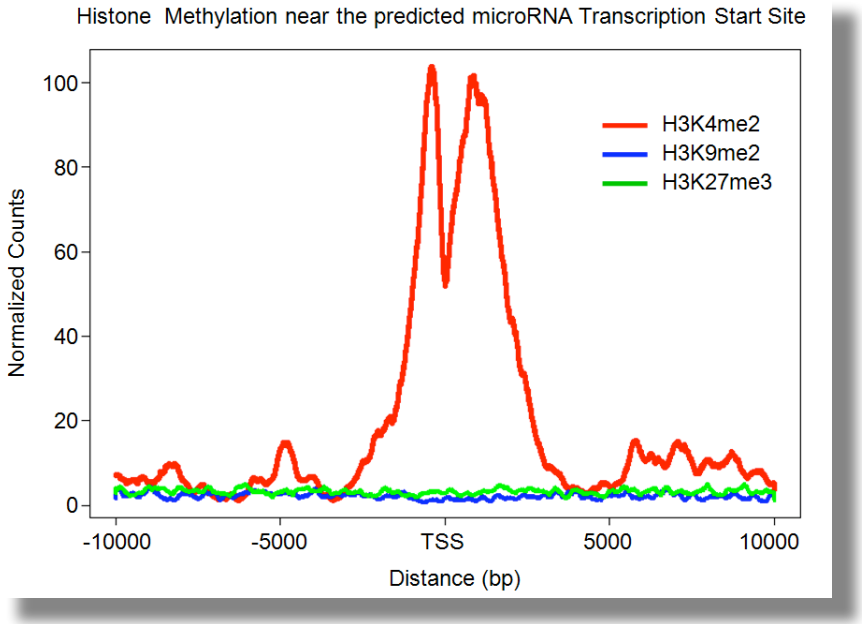
Conservation scores are retrieved from UCSC Genome Browser

Validation II – CpG islands

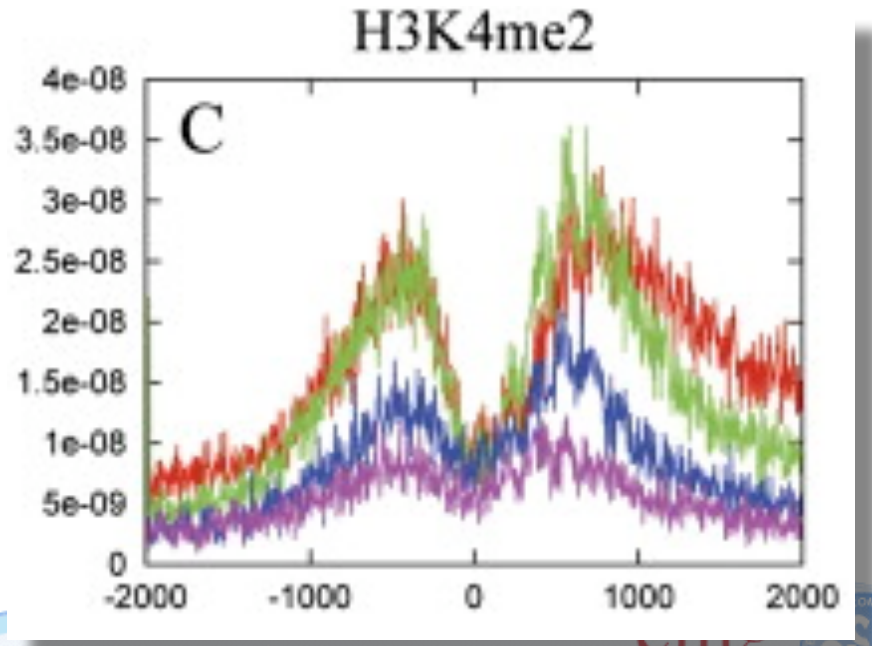
- Number of total annotated CpG islands away from the transcription start site (TSS) in 83 identified microRNAs.



Histone markers (other study)



- H3K4Me2 patterns for identified microRNA promoters

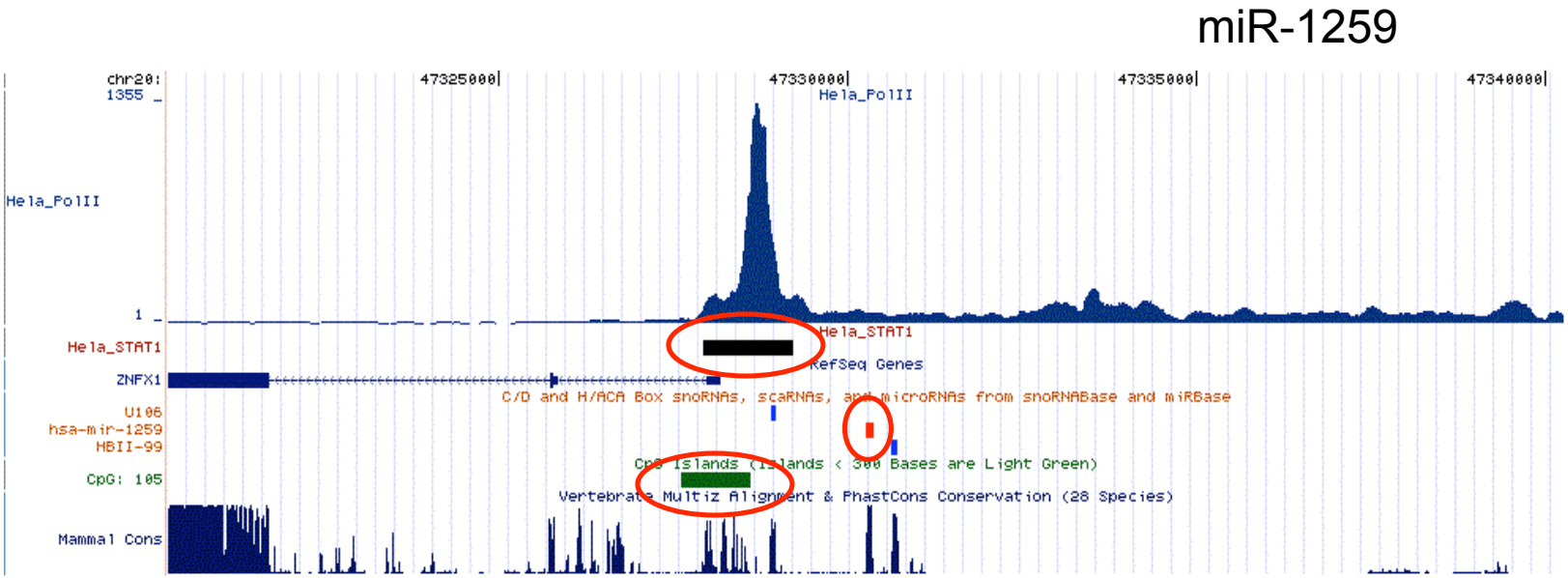


- H3K4Me2 patterns gene published in Barski et al. Cell 2007

Relationship of predicted microRNA promoters with STAT1 binding

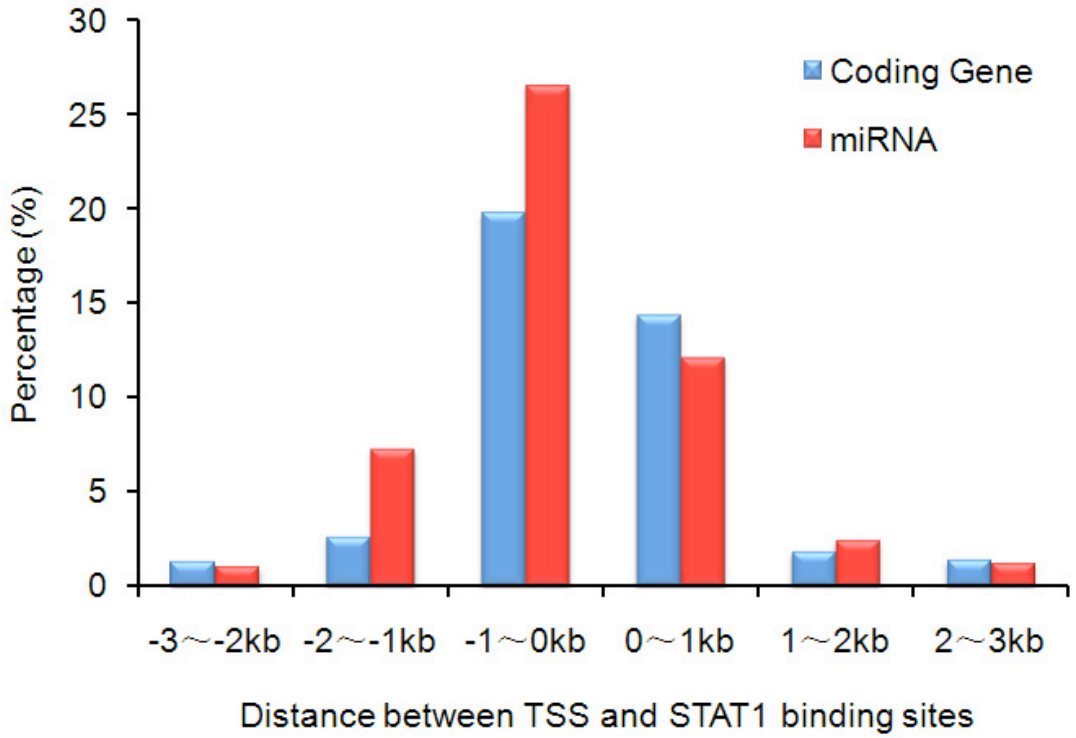
Relationship with STAT1 binding

- Promoter regions of 41 microRNAs (49.4%, out of 83) contain or overlap with STAT1 enriched regions.



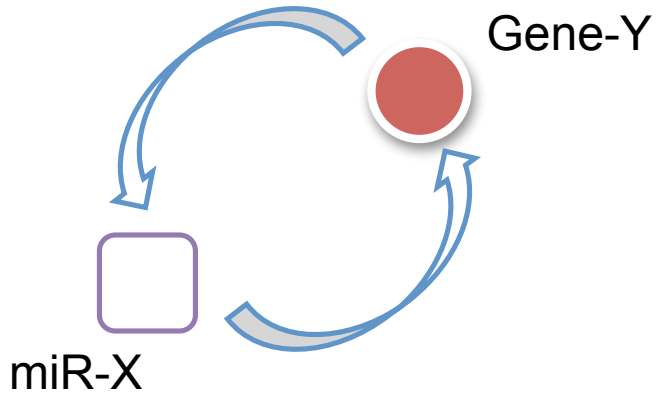
Relationship with STAT1 binding

- Promoter regions of 41 microRNAs (49.4%, out of 83) contain or overlap with STAT1 enriched regions.



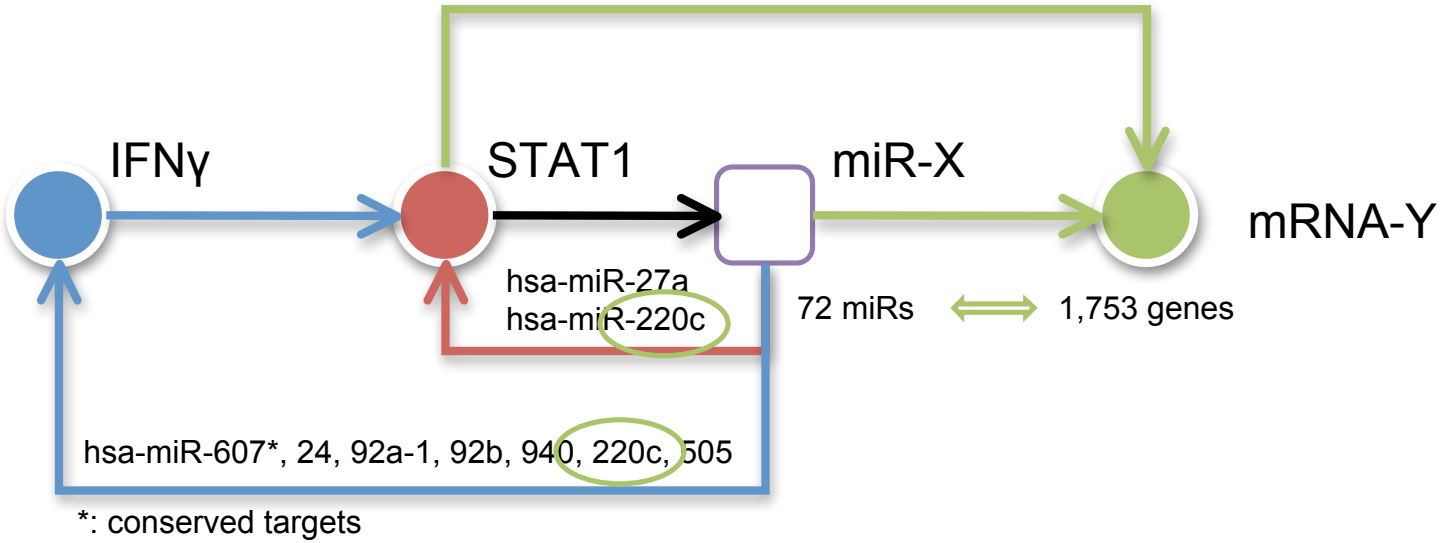
microRNA – mediated regulatory network?

- 83 microRNAs are actively transcribed
- 41 of which are STAT1 targets
- 6,254 genes are STAT1 targets (within -1,000 to +500-bp from TSS)



microRNA – mediated regulatory network?

- Feed back loops
- Feed forward loops



hsa-miR-220 is deregulated in 5 types of cancers, including:

•B-cell chronic lymphocytic leukemia
•Glioblastoma
•Lung cancer
•Pancreatic cancer
•Papillary thyroid carcinoma

Summary

- Genome-wide RNA polymerase II binding data (from ChIP-seq) is a good resource for identifying regulatory regions of microRNA;
- Promoters of 83 microRNAs were identified in HeLa cells.
- Important step to study microRNA regulatory network;
- Similar methodology can also be applied to study regulatory mechanisms of other non-coding RNA, such as pi-RNA;
- For coding genes, tissue-specific alternative promoters can be also identified.

Acknowledgement

- Collaborators
  - Tim Huang (OSU)
  - Ken Nephew (Indiana U.)
  - Changyu Shen (IUSM)

- CAMDA Committee
  - Simon Lin
  - Pan Du

- Data Provider
  - Gerstein group

- Funding Source
  - US NIH-AA017941
  - US NIH-CA113001
  - China 863 High-Tech Program 2007AA02Z302
  - China Natural Science Foundation 60901075

INGEN

cmg Center for Medical Genomics @ Indiana University School of Medicine

Result summary

- Focusing on 419 annotated intergenic microRNAs;
- Computational model identified 83 microRNAs predicted to be transcribed in HeLa cells (FDR ≤ 20%);

RNA-seq experiment