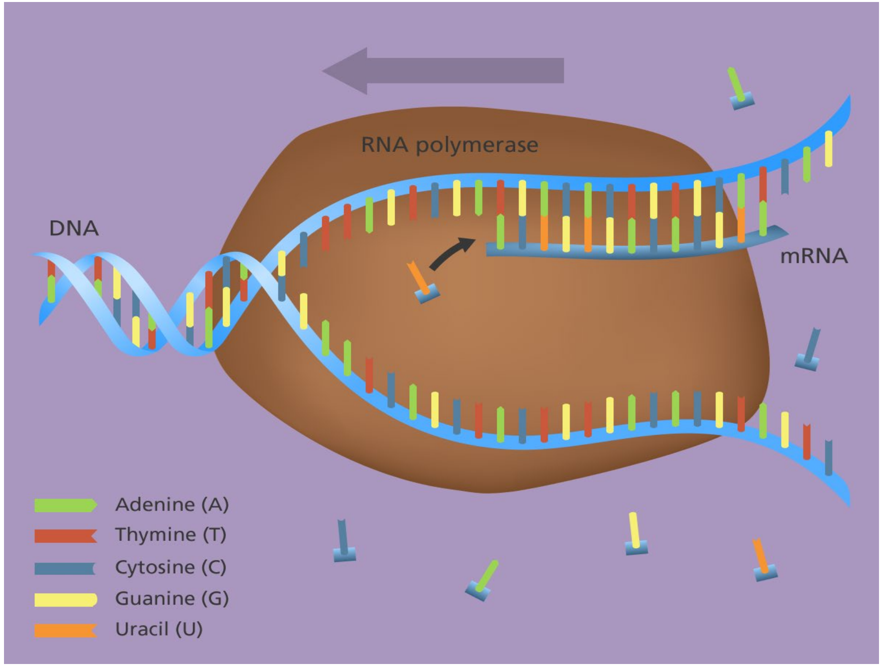
The Identification of Transcriptionally Active Regions

Dante Miller

University of Minnesota - Morris April 20, 2022

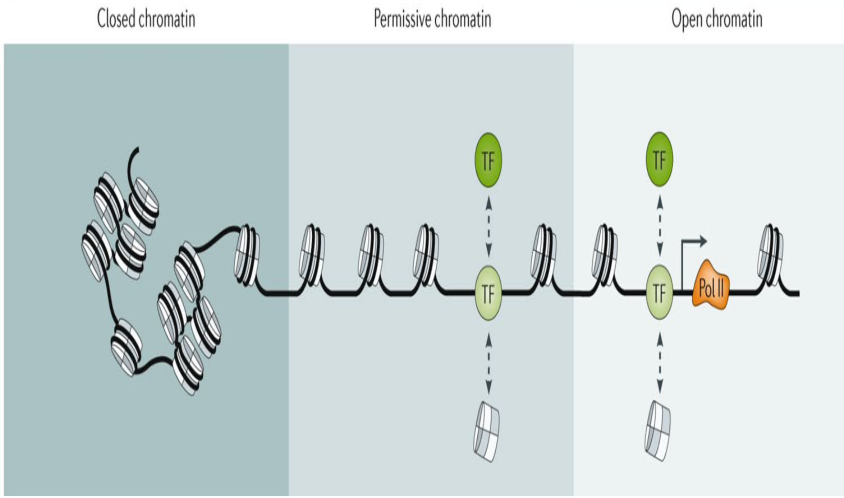
What is Transcription?

- One of the main processes for the flow of genetic information in an organism
- Four Key Components
  - Chromosomes
  - Proteins
  - Genes
  - RNA polymerase
- Transcription factors
- DNA



Chromatin Accessibility

- Chromosomes are made up of chromatin
- There are different types of chromatin
  - Closed accessibility
    - Transcription can not take place
  - Permissive accessibility
    - Regions are permissive of transcription factors f
  - Open accessibility
    - Regions can be accessed by transcription factors and the RNA polymerase enzyme accessibility



A figure on the different types of chromatin accessibility. The three blues areas of the image left from right represent closed, permissive, and open accessibility. [Klemm et al., 2019]

Diseases and Mutations

- Mutations are changes in a DNA sequence
    - Mistakes during the transcription or translation process
    - Caused by environmental factors
- Diseases are caused by mutations
    - Gene(s)
    - Caused by environmental factors
    - Damaged chromosomes
- Series of open chromatin regions are associated with diseases such as cancer [Wang et al., 2021]
- p53 gene can be related to cancer

Treatments and Medicine

- Identification of disease mechanisms is crucial in understanding
  - What mechanisms are associated with specific diseases
  - Determine potential treatments and medications for specific diseases
- Disease mechanisms can be better understood through studies about accessibility and transcriptionally active regions [Perrin et al., 2021]

Problems with current methods

- Nascent transcription methods (GRO-seq)
    - Identifies nascent transcripts (there are methods that identify mature mRNA)
- Current Issues
    - Expensive
    - Difficult to use
    - Requires large portions of cells
    - Struggle in identifying transcriptionally active regions from large sites
- Issues solved through the use of ATAC-seq and recurrent neural networks
    - Classification problem
    - Sequential data

Outline

- Background Information
    - Recurrent Neural Networks
    - ATAC-seq
    - Nascent Transcription (GRO-seq)
- Implementation
    - Data Attainment
    - Results
- Conclusion

Outline

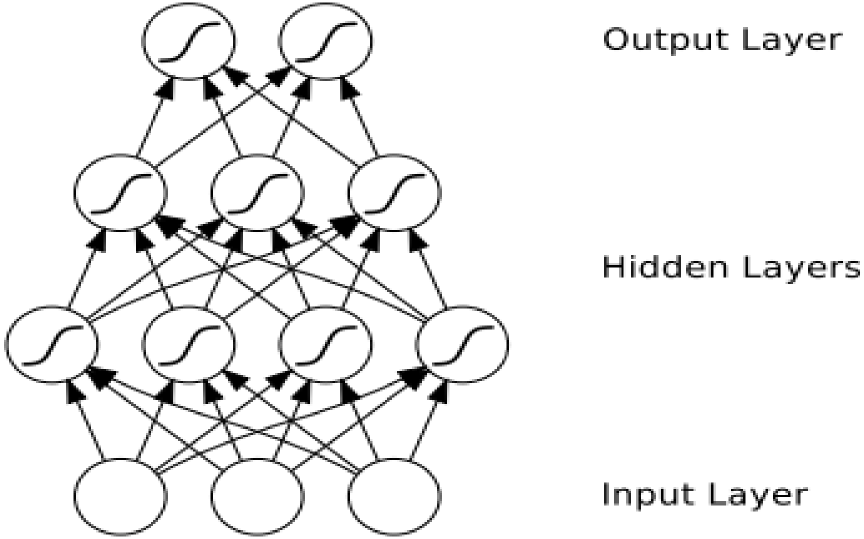
- **Background Information**
  - **Recurrent Neural Networks**
  - ATAC-seq
  - Nascent Transcription (GRO-seq)
- Implementation
  - Data Attainment
  - Results
- Conclusion

Machine Learning

- Machine learning
  - Study of statistical models that automatically improve themselves through experience and data
- Artificial neural networks
  - Subset of machine learning consisting of computing systems inspired by the physical processes observed in brains
- Feedforward neural networks (FNN)
  - A form of artificial neural networks

Feedforward Neural Networks - Organization

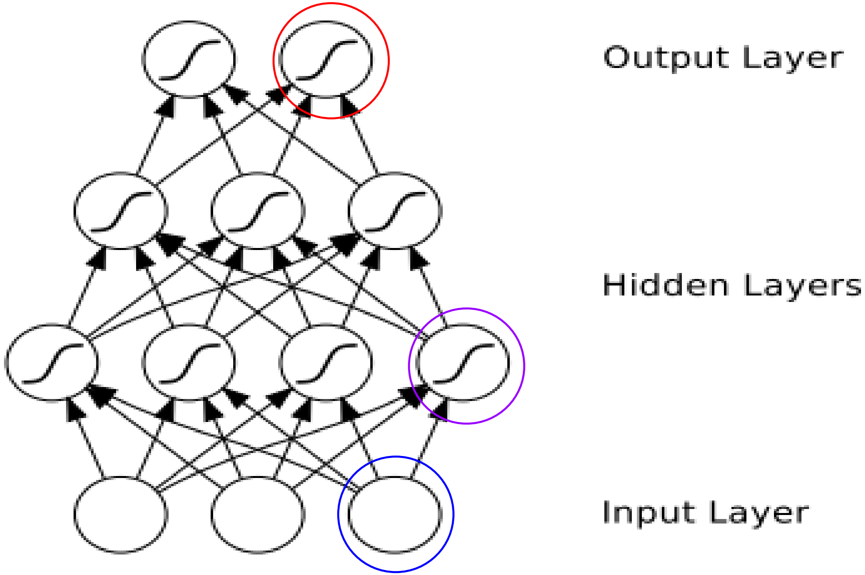
- Organized into
    - Input layer
    - One or more hidden layers
    - Output layer



Feedforward neural network. [Graves, 2012]

Feedforward Neural Networks - Nodes

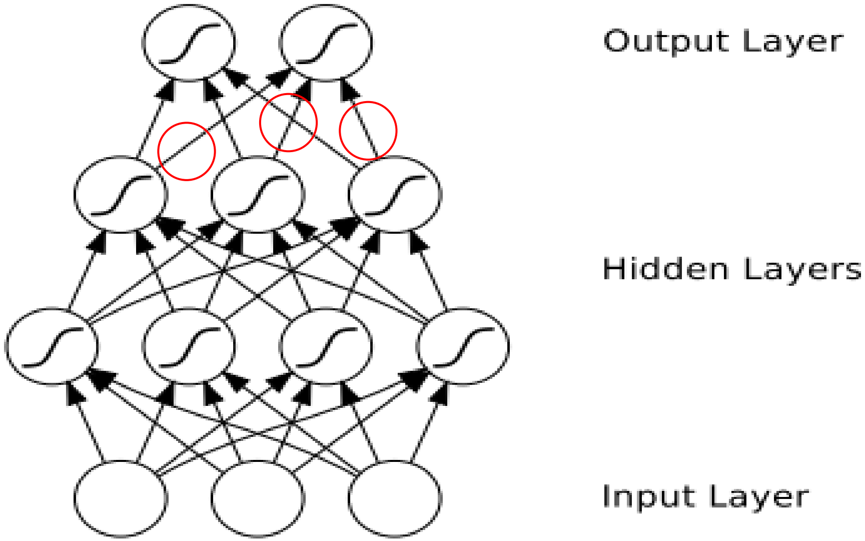
- Nodes
  - Computational units
    - Receives values from previous nodes
    - Performs computation
    - Outputs values to nodes in the next layer or as a final value
- Each node consists of an activation function
  - Map inputs that vary non-linearly to an output



Feedforward neural network. [Graves, 2012]

Feedforward Neural Networks - Edges

- Edges
  - Transmits values to nodes between layers
  - Contains a weight that determines how much the output value
- $$activation function(\sum_{i=1}^{3} x_i * w_i)$$



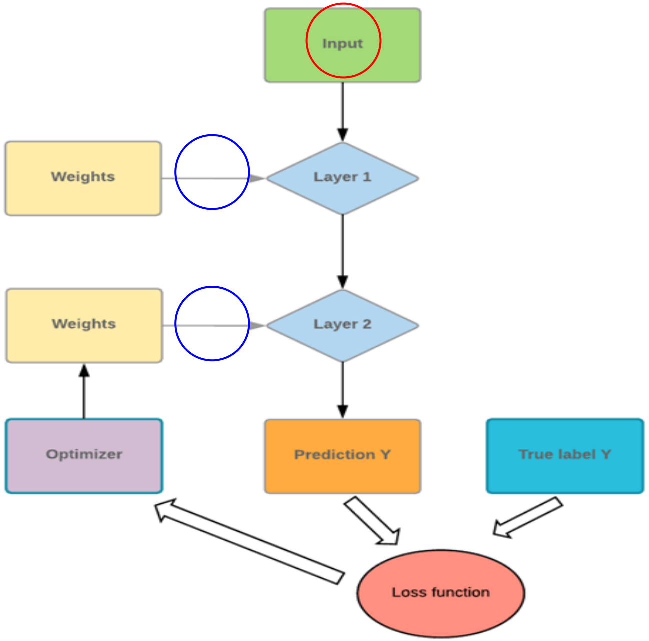
Feedforward neural network. [Graves, 2012]

Feedforward Neural Networks - Training Runthrough

- Training dataset
  - Train model to map inputs to an expected outputs for the variable being predicted
- Testing dataset
  - Make estimates on what the expected outputs are for the variable being predicted in the testing dataset
  - Compare predicted and actuals

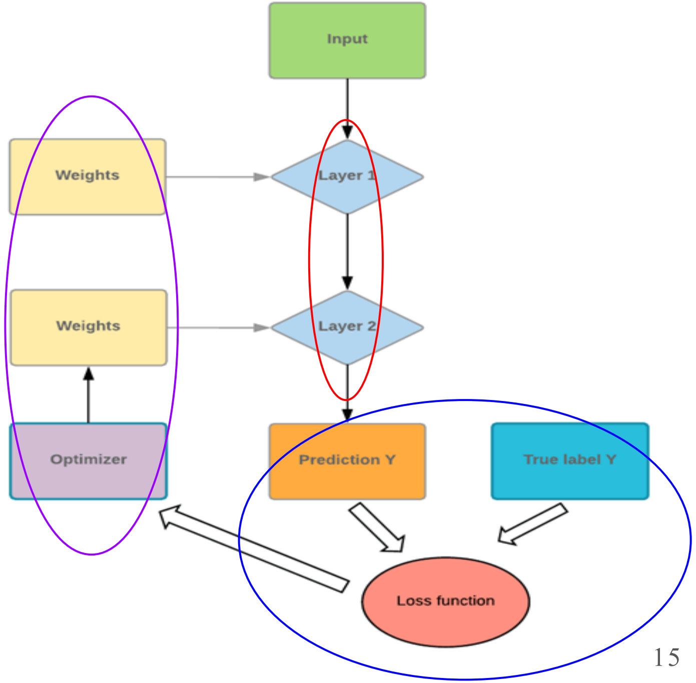
Feedforward Neural Networks - Training Runthrough cont.

- Provide the training dataset to the model
  - Choose the output variables (what is being predicted)
  - Choose the input variables (what is used to determine the outputs)
- Assign random weights



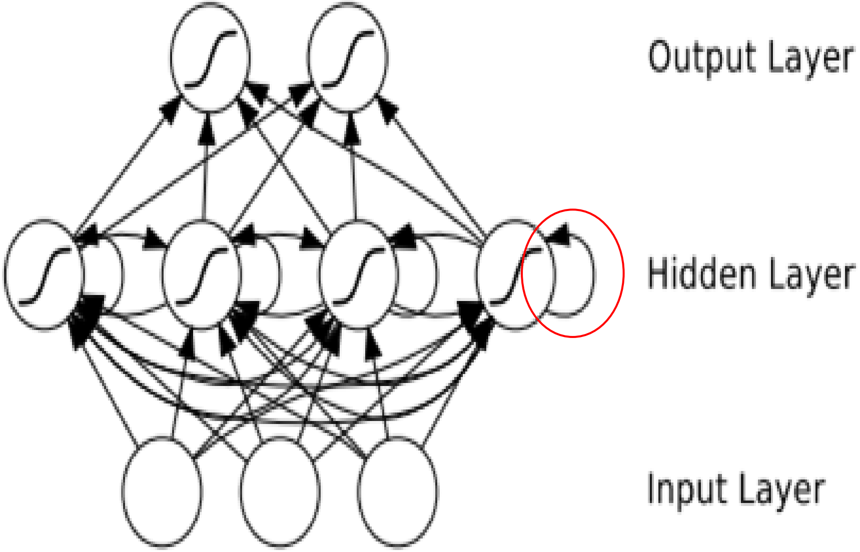
Feedforward Neural Networks - Training Runthrough cont.

- Forward Propagation
  - Mapping input values to an output value
- Loss Function
  - Determines the distance between the expected outputs and predicted outputs
- Back propagation
  - Done through an optimizer
  - Tuning the weights to lower the value produced by the loss function
- Forward and back propagation are ran until
  - Loss function value reaches a sufficient value



What are Recurrent Neural Networks?

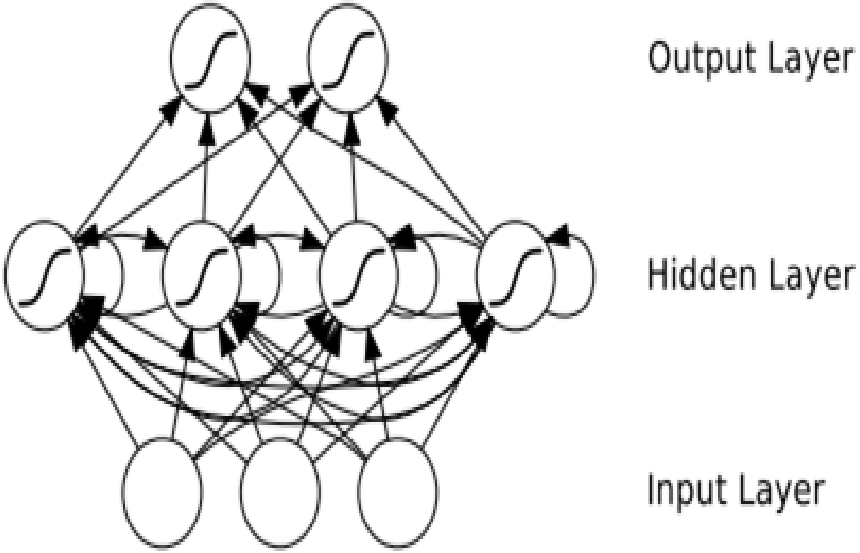
- Derived from Feedforward Neural Networks
  - Handles sequential data
  - Loops allow each previous observation to have an affect on the input in the next observation



Recurrent neural network. [Graves, 2012]

What is the Vanishing Gradient Problem?

- Vanishing Gradient Problem
  - Back propagation has little effect on the weights in earlier layers
- Gated recurrent unit
  - Solves the problem through the use of an update and reset gates
  - Decide what information is relevant to the output and can retain past information



Recurrent neural network. [Graves, 2012]

Outline

- **Background Information**
  - Recurrent Neural Networks
  - **ATAC-seq**
  - **Nascent Transcription (GRO-seq)**
- Implementation
  - Data Attainment
  - Results
- Conclusion

What is ATAC-seq?

- Determines the accessibility of chromatin
- Two parts of the methods
  - ATAC-seq method
  - Pipeline
    - Series of steps involving tools to analyze the ATAC-seq reads

Short read runs

- DNA sequencing
  - Technique that determines the exact sequence of organic molecules in a DNA sequence
- Short read runs (reads)
  - A set of raw sequences after amplifying and sequencing copied short fragments of the DNA

Attainment of ATAC-seq Reads

- Sequencing adapters
  - Short single strands of synthetic DNA or RNA
- Tn5 transposase
  - Inserts sequencing adapters at accessible regions of the genome
- Genome
  - The set of all genetic information about an organism
- Reads are stored in a database with annotations
  - Where the reads are from
  - The study where the genome was sequenced
  - The species that was sequenced

ATAC-seq Pipeline - Preprocessing

- Download the reads and annotations
- Check quality of the reads
- Determine presence of sequencing adapters
- Trim reads
    - Remove low quality reads
    - Sequencing adapters

ATAC-seq Pipeline - Mapping

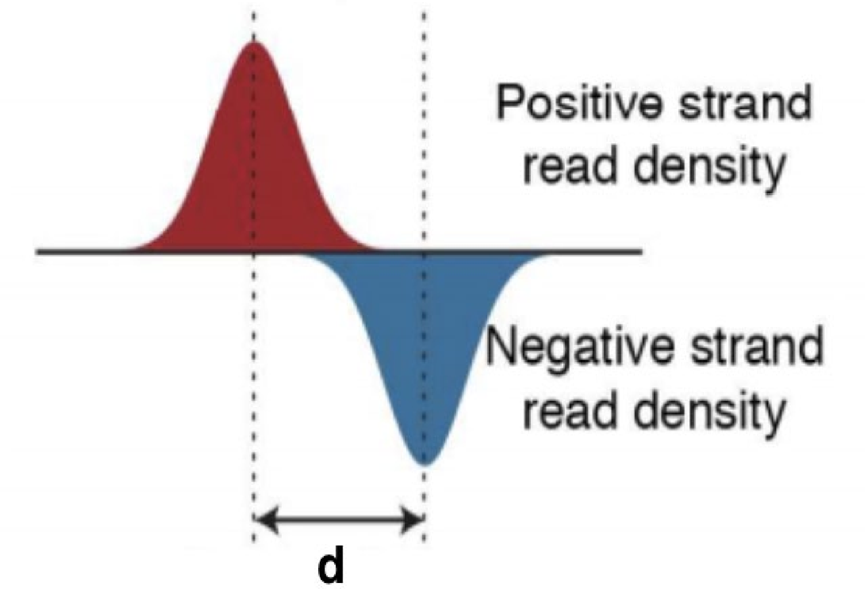
- Human reference genome
  - Database on one idealized individual of a species
- Map the reads to the reference genome of interest
  - Matches read to most similar regions

ATAC-seq Pipeline - Filters

- Filters are imposed on the mapped reads to remove
  - Uninformative reads
  - Duplicate reads
  - Low mapping quality reads
  - Reads that are not properly paired
- Insert size is checked
  - Length of space between the sequencing adapters
  - Determine the quality of reads

ATAC-seq Pipeline - Peak Calling

- Determine density distribution
  - Both strands of the DNA
  - On the two density distributions for both strands of the DNA
- Regions with significant differences
  - Peaks (open chromatin regions)



What is Nascent Transcription?

- Identifies nascent transcripts
- Two parts of the methods
  - Nascent transcription (GRO-seq)
  - Pipeline similar to ATAC-seq

Attainment of Nascent Transcription Reads

- Halts transcription
- Nucleus
    - Controls and regulates the activities of cells
    - Isolated from a cell population of interest
- Organic molecules labeled with tags (markers) are added
- Transcription is then restarted
    - Nascent transcript molecules are labeled
- Reads are produced with annotations

Outline

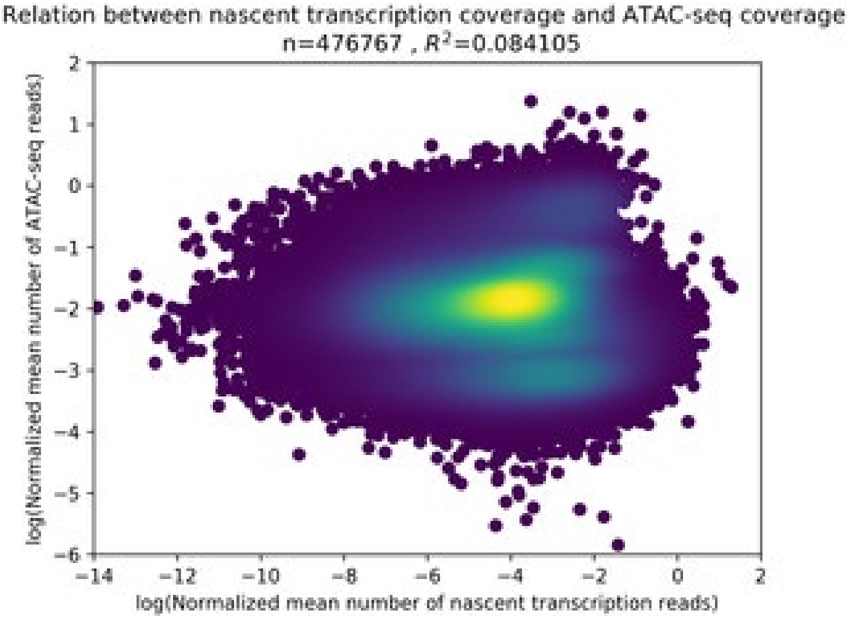
- Background Information
  - Recurrent Neural Networks
  - ATAC-seq
  - Nascent Transcription (GRO-seq)
- **Implementation**
  - **Data Attainment**
  - **Results**
- Conclusion

Implementation - Data Attainment and Pipeline

- Gene Expression Omnibus (GEO)
    - Database that consists of data on gene expression and DNA sequencing
- ATAC-seq and nascent transcription (GRO-seq) reads from GEO
    - Filtered under the criteria of the reads having matching cell types and diseases such as cancer
    - Nine sets of reads were attained
- Pipelines
    - Approximately half a million open chromatin regions were identified from the ATAC-seq pipeline
    - 29 percent of open chromatin regions were labeled as transcriptionally active from the nascent transcription (GRO-seq) pipeline
        - Used as the expected outputs

Implementation - Correlation Between Accessibility and Nascent Transcription

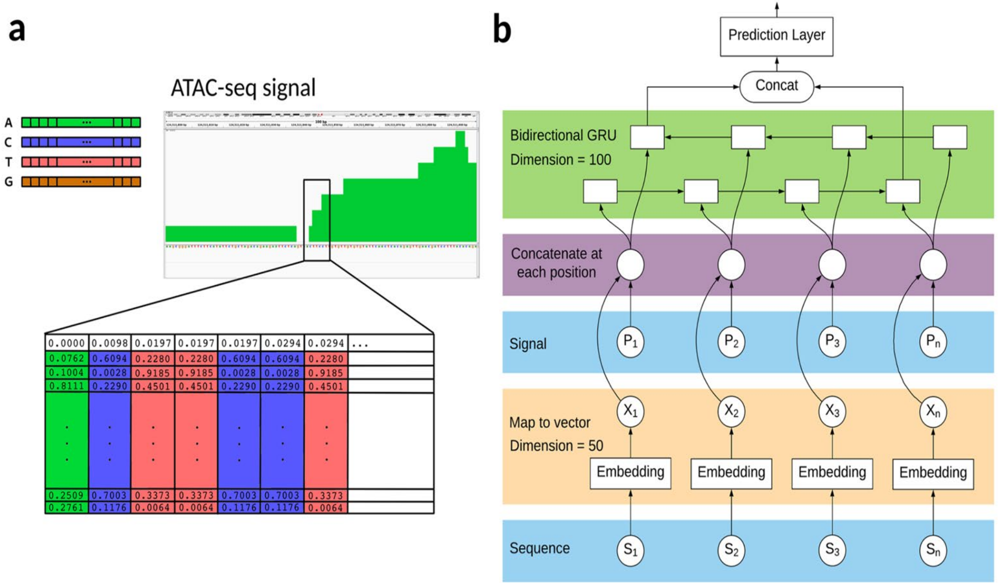
- Low correlation between accessibility and nascent transcription
  - Each point is a open chromatin region
  - Comparing the mapped open chromatin reads from ATAC-se and nascent transcription (GRO-seq)



Relationship between accessibility and nascent transcription.
[Tripodi et al., 2020]

Implementation - Vector Encoding

- Developed a vector encoding
  - Sequence
    - For each open chromatin region
  - Signal
    - Number of mapped ATAC-seq reads for each open chromatin region divide by millions mapped



0.0000	0.0098	0.0197	0.0197	0.0197	0.0294	0.0294	...
0.0762	0.6094	0.2280	0.2280	0.6094	0.6094	0.2280	
0.1004	0.0028	0.9185	0.9185	0.0028	0.0028	0.9185	
0.8111	0.2290	0.4501	0.4501	0.2290	0.2290	0.4501	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
0.2509	0.7003	0.3373	0.3373	0.7003	0.7003	0.3373	
0.2761	0.1176	0.0064	0.0064	0.1176	0.1176	0.0064	

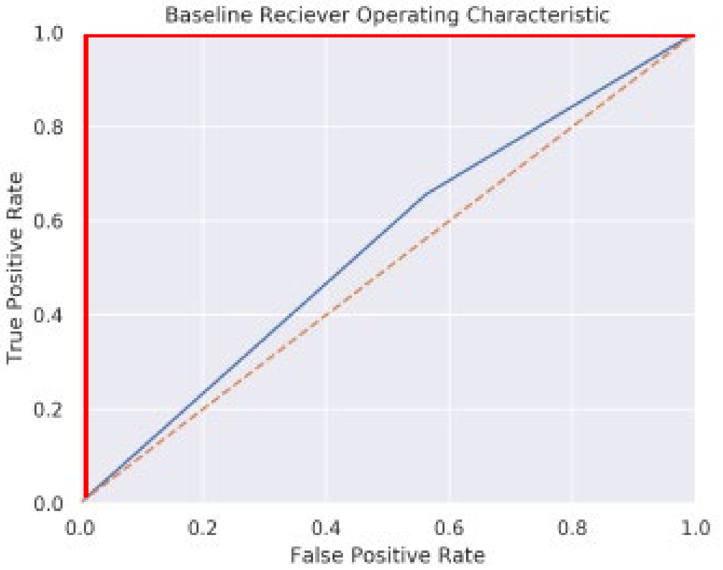
Vector encoding signal and sequence to summarize each open chromatin regions. [Tripodi et al., 2020]

Implementation - Measures

- True positive rate
  - TP / (TP + FN)
    - True Positive (TP)
      - Predicted as transcriptionally active and actually is
    - False Negative (FN)
      - Predicted as not transcriptionally active but is actually is
- False positive rate
  - FP / (FP + TN)
    - False Positive (FP)
      - Predicted as transcriptionally active but actually is not
    - True Negative (TN)
      - Predicted as not transcriptionally active and actually is not
- F1 Score
  - TP / (TP + (½)(FP+FN))
  - Measure of a model's accuracy on a dataset

Implementation - Baseline

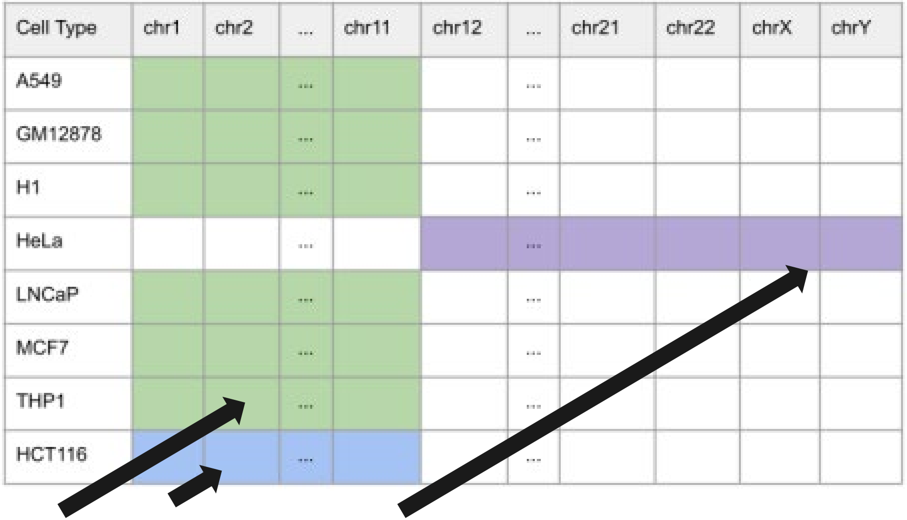
- Developed a baseline method
  - Predicts the probability of whether an open chromatin region is transcriptionally active or not based on the distribution of mapped ATAC-seq reads per open chromatin region labeled as transcriptionally active or not



Baseline method for classifying open chromatin regions as transcriptionally active or not. [Tripodi et al., 2020]

Implementation - RNN Model Template

- Leave-one-out training(LOOT)
  - Model trained eight times
  - Leaves a cell type out each time
  - Trained for each of the cell types
- Data split
  - 90 percent training data
  - 10 percent testing data

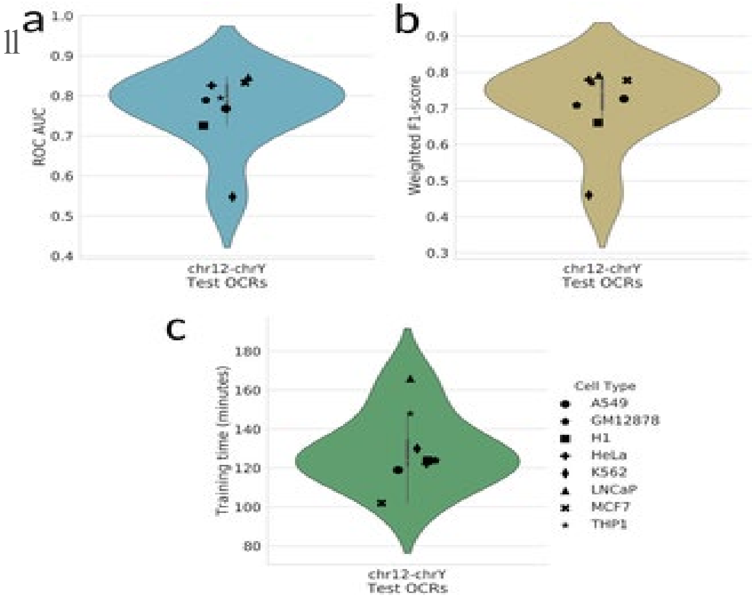


Cell Type	chr1	chr2	...	chr11	chr12	...	chr21	chr22	chrX	chrY
A549	🟩	🟩	🟩 ...	🟩		...				
GM12878	🟩	🟩	🟩 ...	🟩		...				
H1	🟩	🟩	🟩 ...	🟩		...				
HeLa			...		🟪	🟪 ...	🟪	🟪	🟪	🟪
LNCaP	🟩	🟩	🟩 ...	🟩		...				
MCF7	🟩	🟩	🟩 ...	🟩		...				
THP1	🟩	🟩	🟩 ...	🟩		...				
HCT116	🟦	🟦	🟦 ...	🟦		...				

Training, validating, and testing sets. [Tripodi et al., 2020]

Implementation - RNN Performance

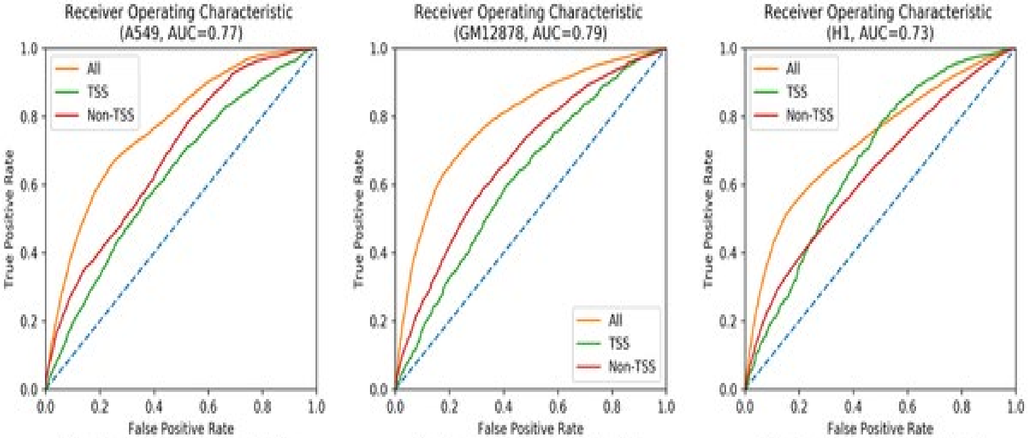
- ROC AUC score is above 70 percent for majority of the cell
  - K562 had the worse quality
- F1 score
  - F1 score is above 65 percent for majority of the cells
- Training time
  - The RNN models for each of the cell types took over 100 minutes



Classifier performance of RNNs. [Tripodi et al., 2020]

Implementation - RNN Performance cont.

- Transcription start site (TSS)
  - The location where the first DNA nucleotide is transcribed into RNA
  - TSS only represent a small fraction of all transcriptionally active regions



Error analysis using ROC curve for comparing predicted vs actuals. [Tripodi et al., 2020]

Outline

- Background Information
  - Recurrent Neural Networks
  - ATAC-seq
  - Nascent Transcription (GRO-seq/ PRO-seq)
- Implementation
  - Data Attainment
  - Results
- **Conclusion**

Conclusion

- Transcriptionally active regions can be identified accurately through the use of ATAC-seq and recurrent neural networks
- Maximizes the benefits of ATAC-seq so that disease mechanisms can be more efficiently studied

Acknowledgement

- I would like to show my appreciation to Dr. Elena Machkasova for her guidance and continuous support over the literature review process. I would like to also thank Dr. Peter Dolan and Chineng Vang for their feedback over the literature review process.

References

- Alex Graves, 2012, Supervised Sequence Labelling with Recurrent Neural Networks
- Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf, 2019, Chromatin accessibility and the regulatory epigenome
- Hannah J. Perrin, Kevin W. Currin, Swarooparani Vadlamudi, Gautam K. Pandey, Kenneth K. Ng, Martin Wabitsch, Markku Laakso, Michael I. Love, and Karen L. Mohlke, 2021, Chromatin accessibility and gene expression during adipocyte differentiation identify context-dependent effects at cardiometabolic GWAS loci
- Ignacio J. Tripodi, Murad Chowdhury, Margaret Gruca, and Robin D. Dowell, 2020, Combining signal and sequence to detect RNA polymerase initiation in ATAC-seq data
- Jiayin Wang, Liubin Chen, Xuanping Zhang, Yao Tong, and Tian Zheng, 2021, OCRDetector: Accurately Detecting Open Chromatin Regions via Plasma Cell-Free DNA Sequencing Data

Questions?