# Enrichment and Analysis of Sequence Motifs in Genomic Variant Calls

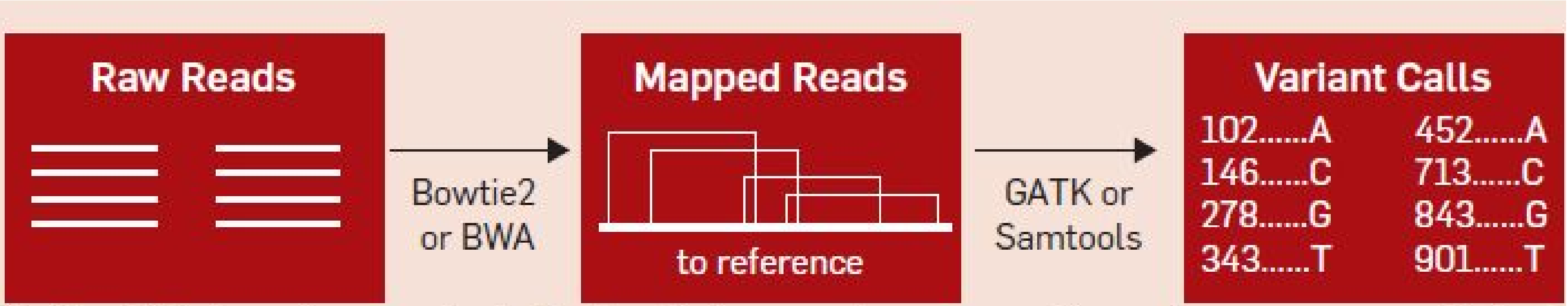
Adithya Vellal
Mentor: Dr. Gil Alterovitz

7th Annual PRIMES Conference
May 21 2017

# Intro to Next Generation Sequencing

- Individual genomes can be sequenced inexpensively
- 3 major parts
  - Raw Genomic Sequence Data(FASTA/FASTQ)
  - Sequence Alignments(SAM/BAM)
  - Genomic Variant Calls(VCF/BCF)
- Finding and analyzing patterns in data crucial to better understanding diseases and drugs

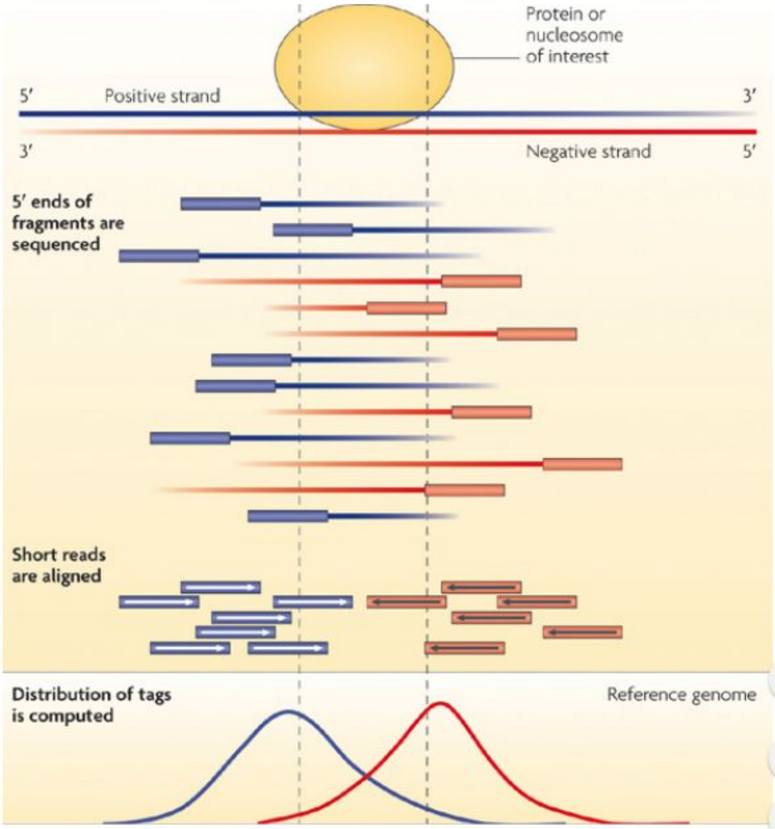
**NGS Pipeline**



# Gene Regulation and ChIP Seq

- Various factors control RNA transcription
  - Regulation of gene expression
- Transcription Factor Binding Sites(TFBS)
  - Represented by sequence motifs
- Chromatin Immunoprecipitation + NGS → ChIP-Seq
  - Peak analysis to determine binding location

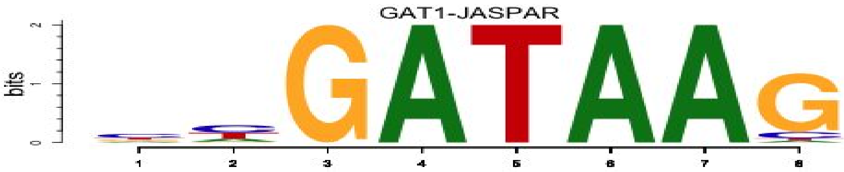
**ChIP-seq Process**

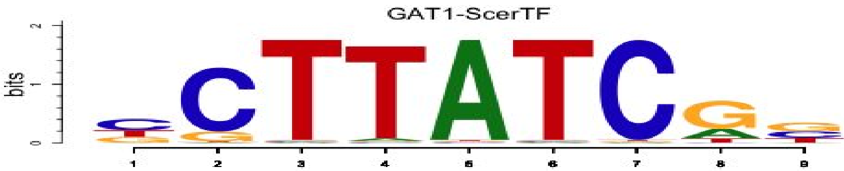

Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome

# Binding Motifs

- Short sequences which represent binding sites
  - ~10 base pairs in length
- Determined using ChIP Seq
  - ENCODE and JASPAR databases
  - Slow and expensive process
  - No way to find common patterns between TFBS
- Not 100% specific
  - Difficult to model effects of variants on TF binding

## Example TF Binding Motifs

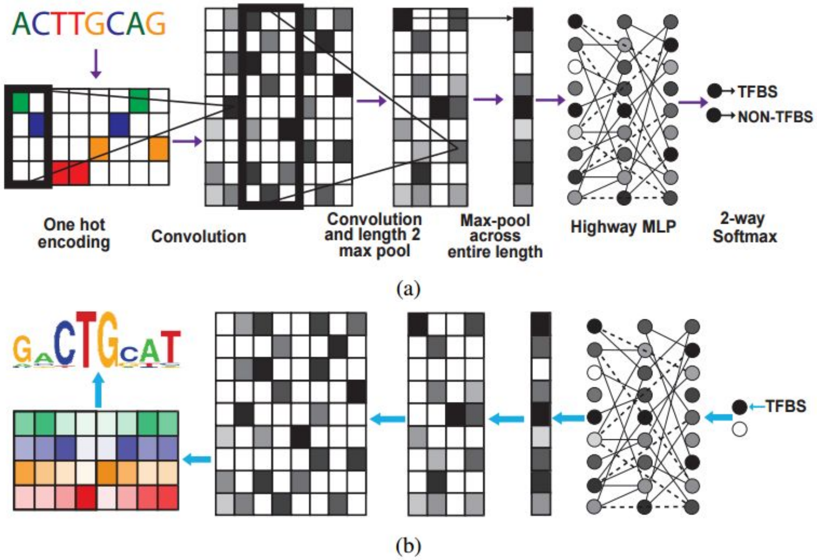

GAT1-JASPAR


GAT1-ScerTF

# Existing Work

- DeepMotif(Lanchantin et al.)
  - Convolutional Neural Network to classify TFBS
  - Individual network for each TF
  - Visualization techniques to predict new motifs
- Shi et al.
  - Random forest classifier predicts effects of SNPs on TF binding

**DeepMotif Network Architecture**

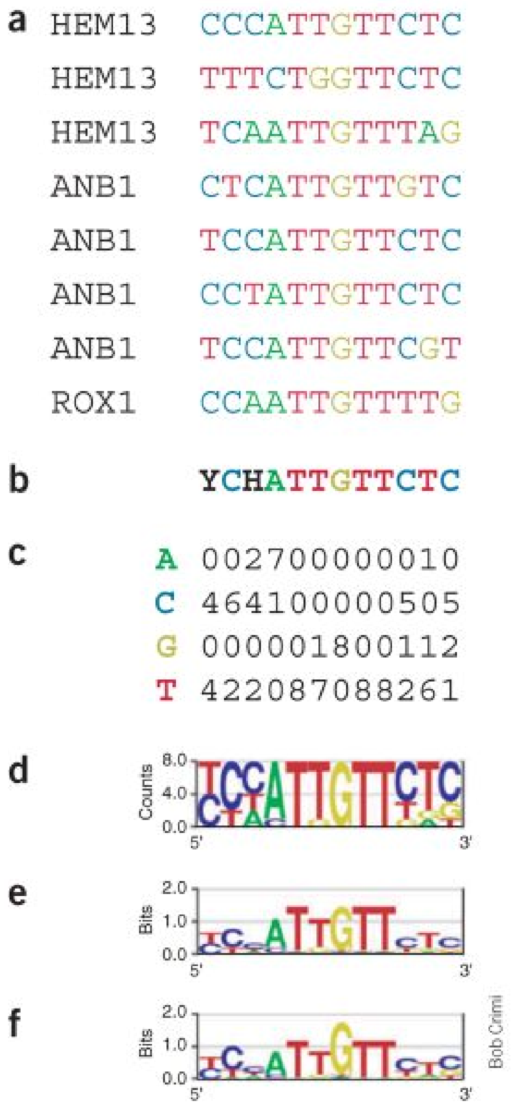


(a)

(b)

# Motif Representation

- Consensus sequence
  - “Ideal” representation
- Position Weight Matrix(PWM)
  - Measures effect of each base on binding energy
  - Easy search of novel sites with high predicted affinity
- Sequence Logo
  - Bases scaled by information content



a

| | |
|---|---|
| HEM13 | CCCATTGTTCTC |
| HEM13 | TTTCTGGTTCTC |
| HEM13 | TCAATTGTTTAG |
| ANB1 | CTCATTGTTGTC |
| ANB1 | TCCATTGTTCTC |
| ANB1 | CCTATTGTTCTC |
| ANB1 | TCCATTGTTCGT |
| ROX1 | CCAATTGTTTTG |

b

YCHATTGTTCTC

c

| | |
|---|---|
| A | 002700000010 |
| C | 464100000505 |
| G | 000001800112 |
| T | 422087088261 |

Bob Crimi

# Intro to Motif Identification

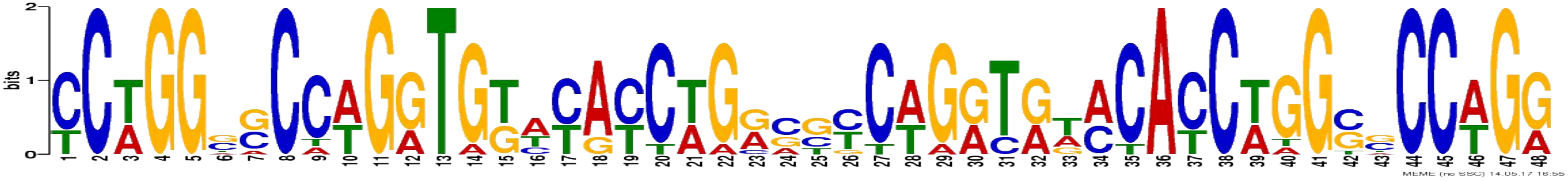
- Data Preparation and Preprocessing
  - Integrate variants into reference genomic sequence
  - Remove all ambiguous bases
  - Segment sequence data into sections of length 100,000
- MM Motif Identification Algorithm
  - **E-value:** expected # of similar motifs found in a sequence of similar length
  - **P-value:** probability that a random sequence would have a stronger motif score than the sequence of interest

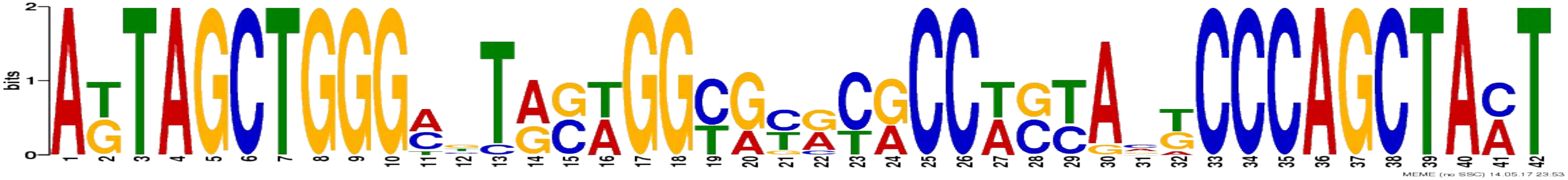
# Motif Identification Cont.

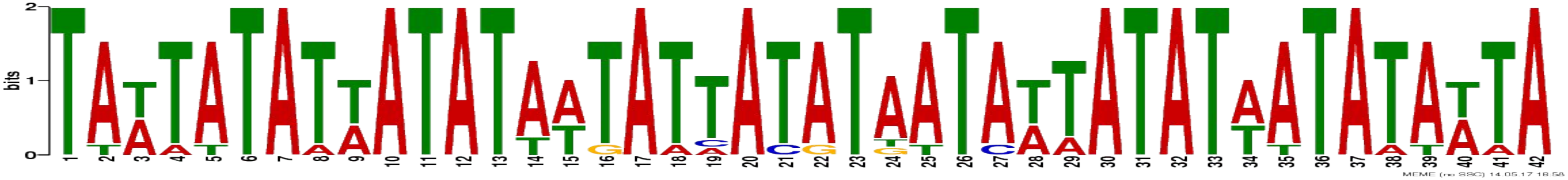
- 100 sequence segments analyzed
- Highest scoring motif in each segment recorded

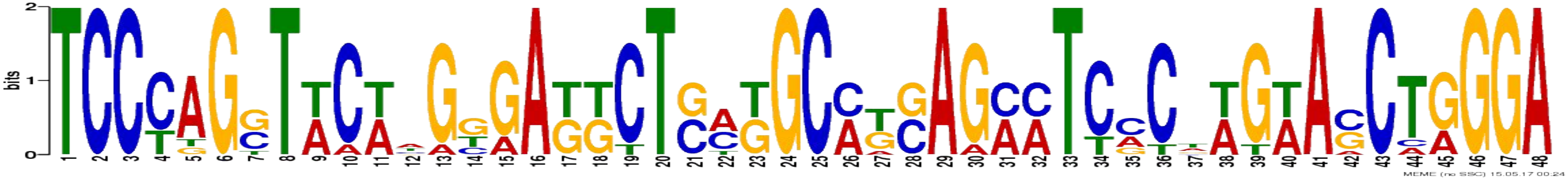
| **Motif Width(bp)** | **Relative Frequency** | **Avg. E-value** | **Avg. P-value** |
|---|---|---|---|
| 42 | 0.51 | $1.8 * 10^{-11}$ | $2.5 * 10^{-16}$ |
| 41 | 0.15 | $5.5 *10^{-13}$ | $1.0 * 10^{-15}$ |
| 48 | 0.12 | $5.6 * 10^{-10}$ | $2.1 * 10^{-17}$ |

# Sample Identified Motif Logos









# Motif Enrichment in ChIP Seq Data

- Analyze ChIP seq peak data for the TF of interest
- Looks for “best” site for motif in each sequence
- Statistic of measurement is **E-value**
- Using pre-determined set of motifs from identification step leads to better results

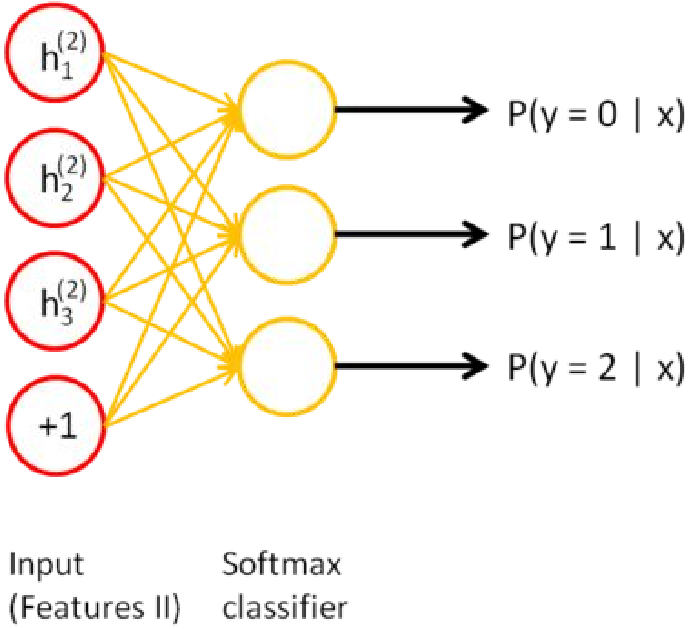
# TFBS Classification Algorithm Outline

- Deep learning model
    - **Convolutional neural network(CNN)**
- Predicts effects of all variants on binding affinity at TFBS
- Training Data: ChIP seq peak calls(ENCODE)
    - Based on enrichment results
- Binary classification of TFBS
- Evaluation Metric: $\mathbf{\Delta\ P(TFBS) = P_{var}(TFBS) - P_{ref}(TFBS)}$

# Network Architecture and Evaluation

- One-hot encoding to form images from sequence data
- Layer structure(Lanchantin et al.)
  - Convolutional layer(4 x 2 feature map)
  - ReLU Layer
  - Max pooling layer(2 x 1)
  - Fully connected layer
- Final max pooling layer + softmax layer
  - Outputs TFBS probabilities

**FC + Softmax Layers**



# Future Work

- Testing and evaluation of convolutional network
- Development of generalized network for all TFBS
  - Currently individual networks required for each one
  - Visualization could help in understanding network
- Testing network with especially compressible data
  - Potential association between effective compression and sequence motifs/TFBS

# Conclusions

- Understanding patterns in sequence motifs is essential to furthering our knowledge of gene regulation
- Motif identification and enrichment can provide valuable insight into patterns found in sequence motifs
- Deep learning provides a simple and effective paradigm for predicting the effects of variants on TF binding

# Acknowledgements

- **MIT PRIMES** for providing this excellent and challenging research opportunity
- **Dr. Gil Alterovitz** for all his guidance and support
- **My parents** for their support