Machine Learning for Genomics

Introduction to Genomics and Types of Genomic Data

Sarwan Ali

Department of Computer Science
Georgia State University

June 23, 2025

Understanding Genomic Data Types

Today's Learning Journey

1. Introduction to Genomics
2. DNA Sequences
3. RNA-seq Data
4. ChIP-seq Data
5. ATAC-seq Data
6. Variant Data (VCF)
7. Data Integration Challenges

What is Genomics?

Definition

Genomics is the comprehensive study of an organism's entire DNA sequence, including all genes and non-coding sequences.

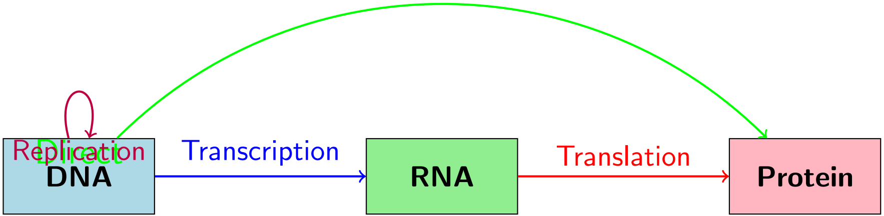
Key Concepts:

- Genome: Complete set of DNA
- Gene: Functional unit of heredity
- Chromosome: Structure containing DNA
- Nucleotides: Building blocks (A, T, G, C)

Applications:

- Disease diagnosis
- Drug discovery
- Personalized medicine
- Evolution studies

Central Dogma of Molecular Biology



Why This Matters for ML

Each step generates different types of data that require specific computational approaches and machine learning techniques.

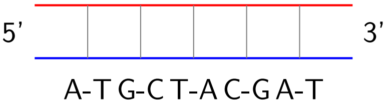
DNA Sequences: The Foundation

What are DNA Sequences?

Linear sequences of nucleotides (A, T, G, C) that encode genetic information.

Characteristics:

- 4-letter alphabet: {A, T, G, C}
- Double-stranded (complementary)
- Human genome: 3.2 billion base pairs
- Contains coding and non-coding regions



Example:

ATCGTACGGCTACGAT

DNA Sequence Data Formats

FASTA Format:

```
>seq1 description
ATCGATCGATCG
TACGTACGTACG
>seq2 description
GCTAGCTAGCTA
```

FASTQ Format:

```
@seq1
ATCGATCG
+
IIIIIIII
```

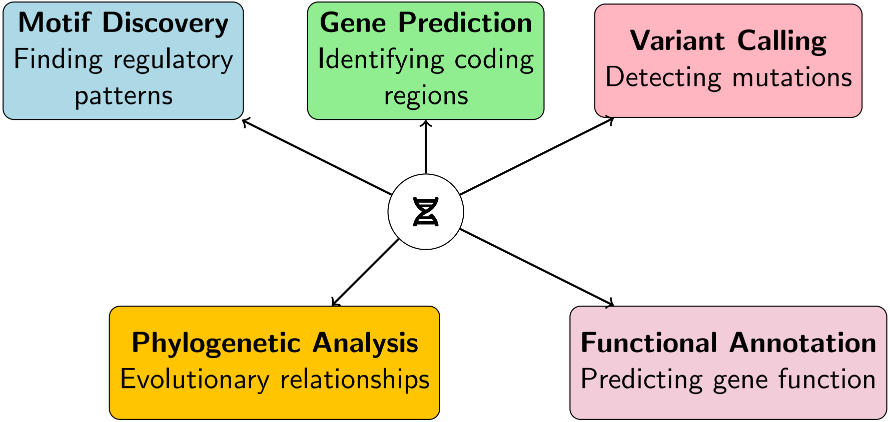
Key Properties:

- **Header**: Sequence identifier
- **Sequence**: Actual nucleotides
- **Quality**: Sequencing confidence (FASTQ)
- **Length**: Variable (genes to genomes)

ML Considerations

- Variable length sequences
- Sequence representation
- Feature extraction methods

ML Applications with DNA Sequences



RNA-seq: Measuring Gene Expression

What is RNA-seq?

RNA sequencing measures the quantity and sequences of RNA molecules in a biological sample, providing a snapshot of gene expression.

Process Overview:

1. RNA extraction from cells
2. Reverse transcription to cDNA
3. Library preparation
4. High-throughput sequencing
5. Computational analysis

Data Characteristics:

- Quantitative: Expression levels
- Qualitative: Transcript sequences
- Temporal: Expression over time
- Conditional: Different treatments
- High-dimensional: 20,000+ genes

RNA-seq Data Types and Formats

Count Matrix:

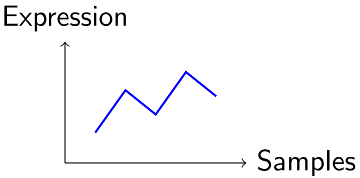
Gene	Sample1	Sample2	Sample3
Gene1	1500	1200	1800
Gene2	500	800	600
Gene3	2000	1900	2100
...	...	...	...

Expression Units:

- Raw counts
- RPKM/FPKM
- TPM (Transcripts Per Million)
- Log-transformed values

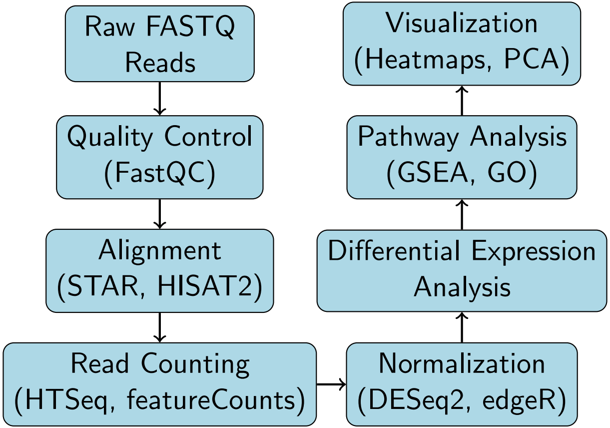
File Formats:

- **FASTQ**: Raw sequencing reads
- **SAM/BAM**: Aligned reads
- **GTF/GFF**: Genome annotations
- **CSV/TSV**: Count matrices



Gene Expression Profile

RNA-seq Analysis Pipeline



ML Applications in RNA-seq

Classification Tasks:

- Disease vs. healthy samples
- Tumor subtype classification
- Drug response prediction
- Cell type identification

Clustering Tasks:

- Co-expression analysis
- Sample clustering
- Gene module discovery
- Trajectory analysis

Common ML Methods:

- **Dimensionality Reduction**: PCA, t-SNE, UMAP
- **Clustering**: k-means, hierarchical
- **Classification**: SVM, Random Forest, Neural Networks
- **Feature Selection**: LASSO, mutual information

Challenges

High dimensionality, batch effects, normalization, missing values

ChIP-seq: Protein-DNA Interactions

What is ChIP-seq?

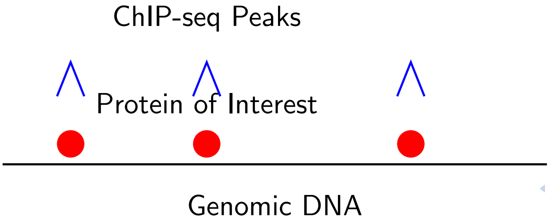
Chromatin Immunoprecipitation followed by sequencing identifies genome-wide protein-DNA binding sites and histone modifications.

ChIP-seq Protocol:

1. Cross-link proteins to DNA
2. Fragment chromatin
3. Immunoprecipitate target protein
4. Reverse cross-links, Sequence purified DNA

Applications:

- Transcription factor binding
- Histone modifications
- Chromatin accessibility
- Regulatory element discovery, Epigenetic studies



ChIP-seq Data Characteristics

Data Structure:

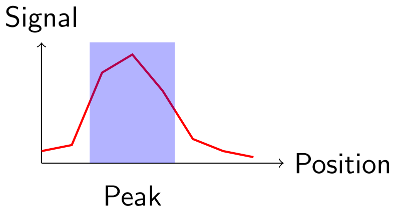
- **Reads**: Short DNA sequences
- **Peaks**: Enriched regions
- **Signal**: Read coverage
- **Controls**: Input/IgG samples

File Formats:

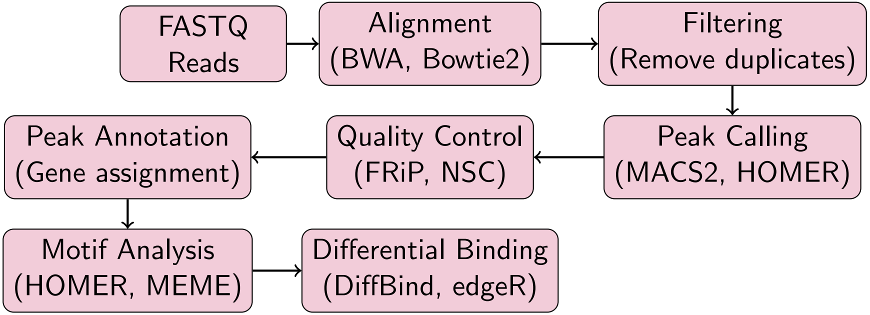
- FASTQ (raw reads)
- BAM (aligned reads)
- BED (peak coordinates)
- BigWig (signal tracks)
- narrowPeak/broadPeak

Peak Example (BED format):

```
chr1 1000 1500 peak1 100
chr1 2000 2300 peak2 150
chr2 5000 5200 peak3 200
```



ChIP-seq Analysis Workflow



Key Metrics

FRiP: Fraction of Reads in Peaks, **NSC**: Normalized Strand Correlation, **Peak Width**: Narrow vs. Broad peaks

ML Applications in ChIP-seq

Peak Prediction:

- Supervised learning for peak calling
- Feature engineering from signal
- CNN for peak detection
- Transfer learning across cell types

Motif Discovery:

- Unsupervised pattern discovery
- Deep learning for motif recognition
- Sequence-to-binding prediction

Regulatory Prediction:

- Enhancer-promoter interactions
- Gene regulation modeling
- Chromatin state prediction
- Multi-omics integration

Common Approaches:

- **CNNs**: Sequence pattern recognition
- **RNNs**: Sequential dependencies
- **Random Forest**: Feature importance
- **HMMs**: Chromatin states

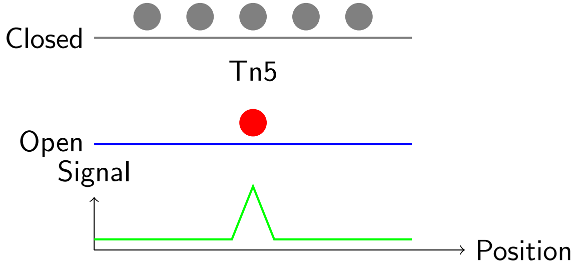
ATAC-seq: Chromatin Accessibility

What is ATAC-seq?

Assay for Transposase-Accessible Chromatin using sequencing identifies regions of open, accessible chromatin genome-wide.

ATAC-seq Protocol:

1. Isolate nuclei from cells
2. Tn5 transposase tagmentation
3. Simultaneous fragmentation and tagging
4. PCR amplification
5. High-throughput sequencing



Advantages:

- Fast and simple protocol, Single-cell compatible
- High resolution, No antibodies required

ATAC-seq vs ChIP-seq

Aspect	ATAC-seq	ChIP-seq
Target	Open chromatin regions	Specific protein binding
Specificity	General accessibility	Protein-specific
Protocol	Simple, fast (1 day)	Complex, long (3-4 days)
Cell number	Low (500-50,000)	High (¿1 million)
Antibody	Not required	Required
Resolution	Nucleotide level	100-200 bp
Applications	Regulatory regions, nucleosome positioning	TF binding, histone modifications

Complementary Nature

ATAC-seq identifies *where* chromatin is accessible, while ChIP-seq identifies *what proteins* are bound there.

ATAC-seq Data Analysis

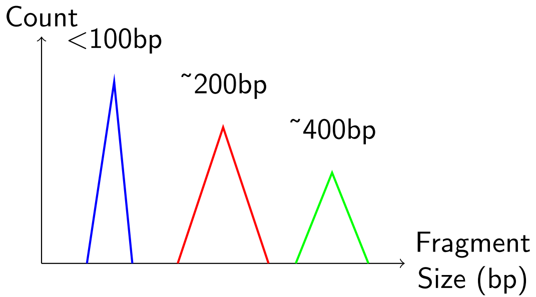
Data Processing Steps:

1. Quality control (FastQC)
2. Adapter trimming
3. Alignment to reference genome
4. Remove mitochondrial reads
5. Peak calling (MACS2)
6. Fragment size analysis

Quality Metrics:

- TSS enrichment score
- Fragment size distribution
- FRiP score
- Library complexity

Fragment Size Pattern:



Nucleosome Ladder

ML Applications in ATAC-seq

Peak Classification:

- Promoter vs enhancer prediction
- Cell type-specific accessibility
- Developmental stage classification
- Disease state identification

Single-cell ATAC-seq:

- Cell clustering and annotation
- Trajectory inference
- Dimensionality reduction
- Batch effect correction

Integration Tasks:

- ATAC + RNA-seq integration
- Multi-modal cell identification
- Regulatory network inference
- Chromatin state prediction

ML Methods:

- **Matrix factorization**: Topic modeling
- **Graph neural networks**: Cell relationships
- **Autoencoders**: Dimensionality reduction
- **Transformer models**: Sequence patterns

Variant Call Format (VCF)

What is VCF?

Variant Call Format is a standardized text file format for storing gene sequence variations against a reference genome.

VCF Structure:

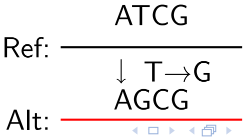
- **Header**: Metadata and format info
- **Columns**: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, samples
- **Variants**: SNPs, INDELs, CNVs, SVs

Example VCF Entry:

```
chr1 1000 .  A G 60 PASS DP=30;AF=0.5 GT:DP 0/1:15
```

Variant Types:

- **SNP**: A→G
- **Insertion**: A→AGT
- **Deletion**: AGT→A
- **MNP**: AT→GC
- **SV**: Large variants



Variant Annotation and Effects

Functional Consequences:

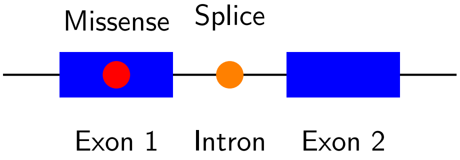
- **Synonymous**: No amino acid change
- **Missense**: Amino acid substitution
- **Nonsense**: Premature stop codon
- **Frameshift**: Reading frame alteration
- **Splice site**: Affects splicing
- **Regulatory**: Non-coding effects

Annotation Tools:

- VEP (Variant Effect Predictor)
- ANNOVAR
- SnpEff
- CADD scoring

Clinical Significance:

- Pathogenic
- Likely pathogenic
- Uncertain significance
- Likely benign
- Benign



ML Applications with Variant Data

Pathogenicity Prediction:

- Disease variant classification
- GWAS signal prioritization
- Rare variant interpretation
- Pharmacogenomics predictions

Population Genetics:

- Ancestry inference
- Population stratification
- Selection signatures
- Demographic modeling

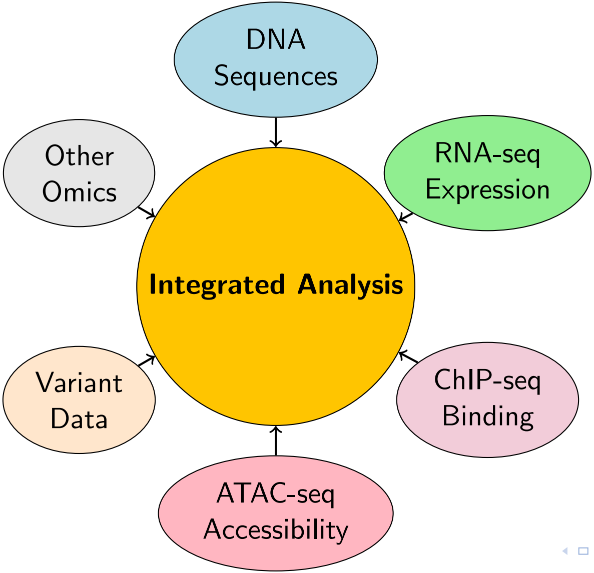
Feature Engineering:

- Sequence context features
- Conservation scores
- Functional annotations
- Population frequencies
- Protein structure impacts

ML Approaches:

- **Ensemble methods**: Random Forest, XGBoost
- **Deep learning**: CNNs for sequence context
- **Graph networks**: Protein interaction effects
- **Multi-task learning**: Multiple phenotypes

Multi-omics Data Integration



Integration Challenges and Solutions

Major Challenges:

- **Scale differences**: Different data sizes
- **Noise levels**: Varying signal quality
- **Missing data**: Incomplete measurements
- **Batch effects**: Technical variations
- **Temporal dynamics**: Different time scales
- **Sample alignment**: Matching across assays

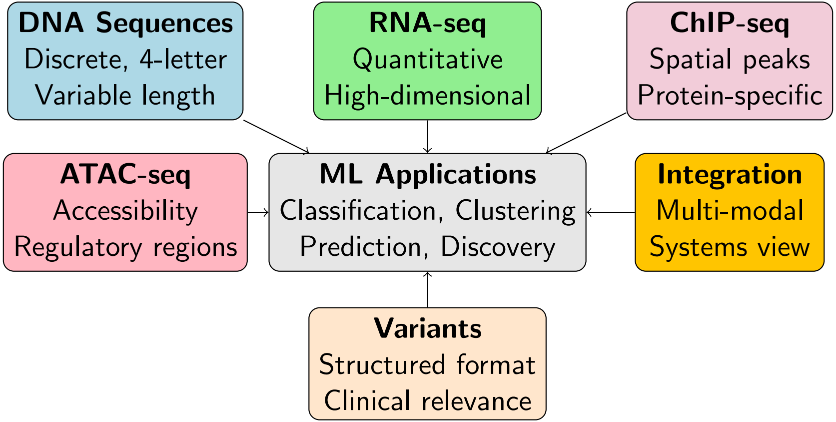
ML Solutions:

- **Multi-view learning**: Joint representation
- **Transfer learning**: Cross-domain knowledge
- **Graph neural networks**: Relationship modeling
- **Variational autoencoders**: Latent integration
- **Multi-task learning**: Shared features
- **Attention mechanisms**: Importance weighting

Best Practices

Start with pairwise integration, validate with independent data, consider biological priors, and maintain interpretability.

Summary: Genomic Data Landscape



Next Steps: Practical Considerations

Data Preprocessing:

- Quality control procedures
- Normalization strategies
- Feature engineering
- Dimensionality reduction
- Batch effect correction

Model Selection:

- Problem-specific architectures
- Validation strategies
- Interpretability requirements
- Computational constraints

Evaluation Metrics:

- Biological relevance
- Statistical significance
- Reproducibility
- Generalization ability
- Clinical utility

Ethical Considerations:

- Data privacy and security
- Bias and fairness
- Informed consent
- Result interpretation
- Clinical responsibility

Questions & Discussion

Next: Hands-on analysis with real genomic datasets

Contact: [sali85@student.gsu.edu]