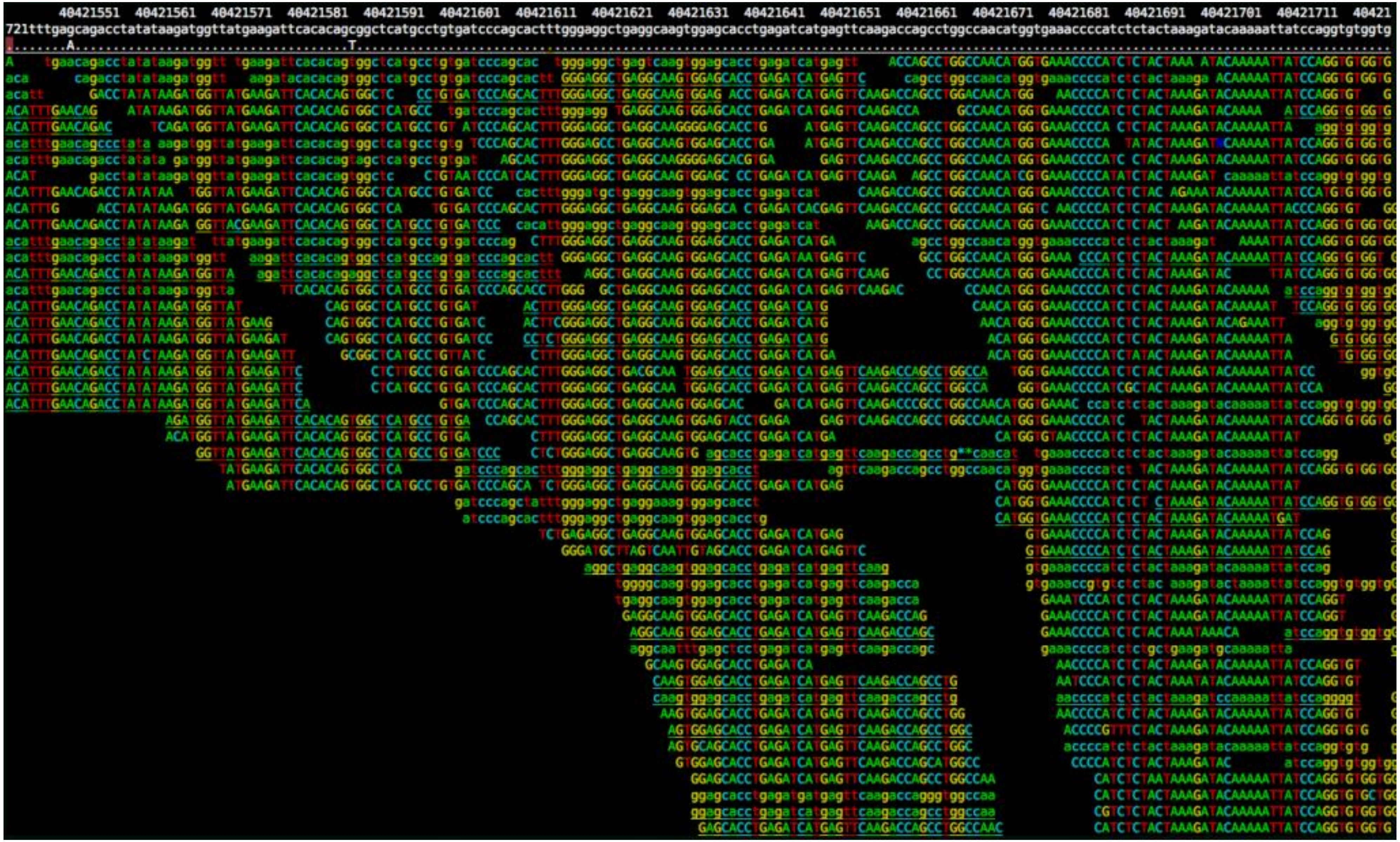
Using Gene Sets To Analyze Genomic Compression

Eric You
Mentor: Dr. Gil Alterovitz
7th Annual Primes Conference
May 21 2017

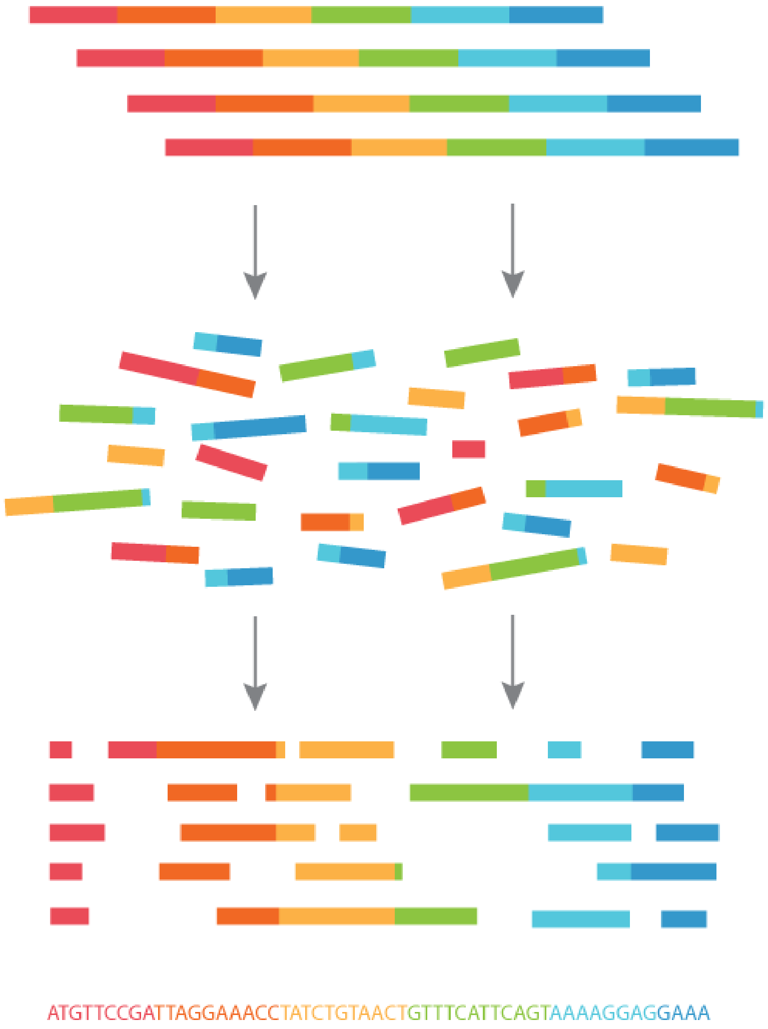


```
     40421551  40421561  40421571  40421581  40421591  40421601  40421611  40421621  40421631  40421641  40421651  40421661  40421671  40421681  40421691  40421701  40421711  40421
721tttgagcagacctatataagatggttatgaagattcacacagcggctcatgcctgtgatcccagcactttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatggtgaaaccccatctctactaaagatacaaaaattatccaggtgtggtg
.......A.................................T..............................................................................................................................
```

Next Generation Sequencing (NGS) readout from autism’s genetics

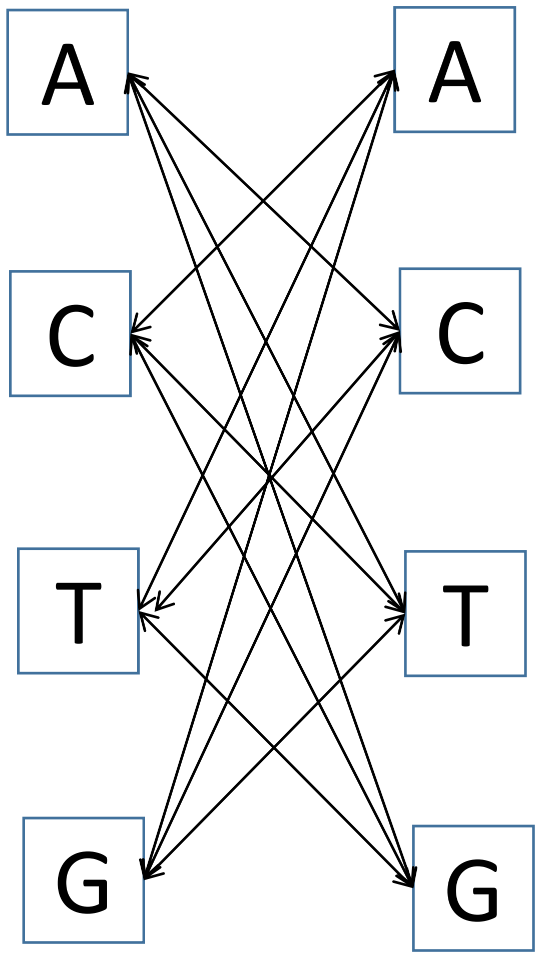
An Introduction to Genomic Data

- Next Generation Sequencing (NGS)
  - **Easy** human genomic data
- NGS for genetics research
  - **Precise** detection of variants
  - **Personalized** drug and medicines
- NGS data
  - **Large**, **difficult** to store and process

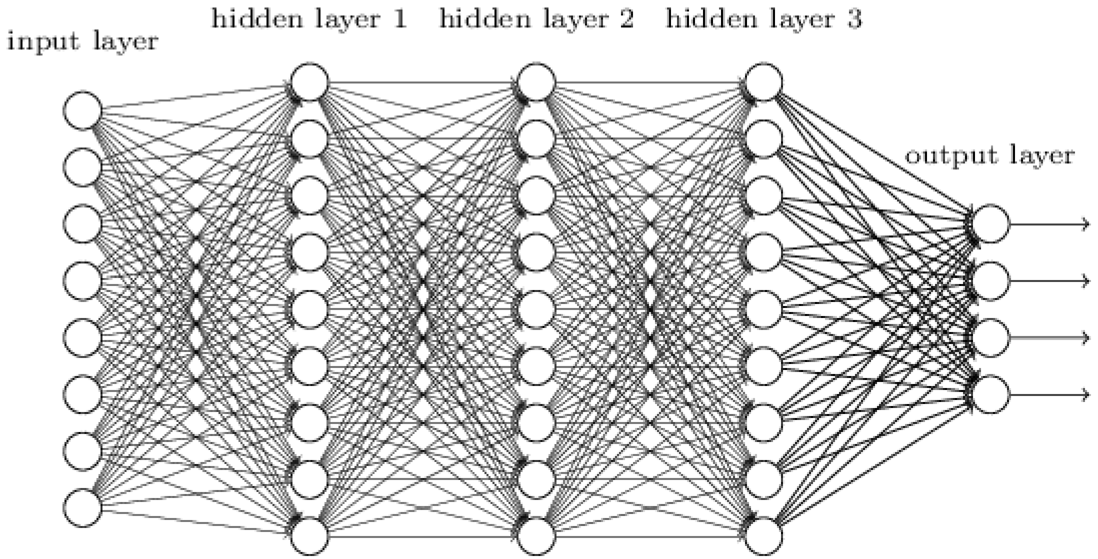


An Introduction to Genomic Compression

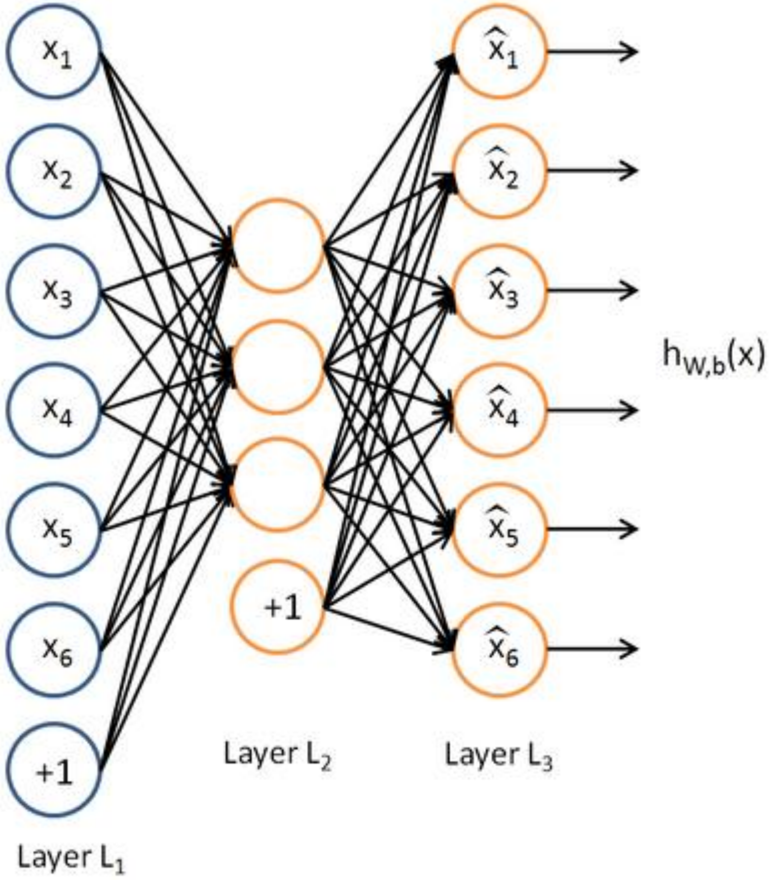
- Genomic compression would greatly improve handling
- Conventional compression can be improved through intrinsic patterns in **variant data (SNPs)**
- Group’s ongoing focus on **autoencoders** and **convolutional networks**



Current Work in Compression



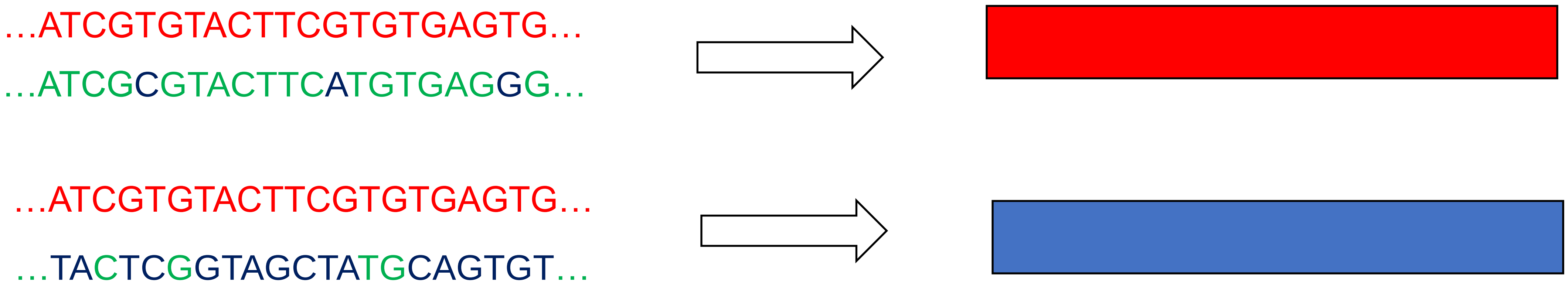
Fully connected convolutional neural network



autoencoder with smaller hidden layer

Current Research

- Correlate compression results to properties of genomic data
  - Gene sets organize data
- Quantify differences in genomic data
- Similar genomic data should be similar after compression!



…ATCGTGTACTTCGTGTGAGTG…

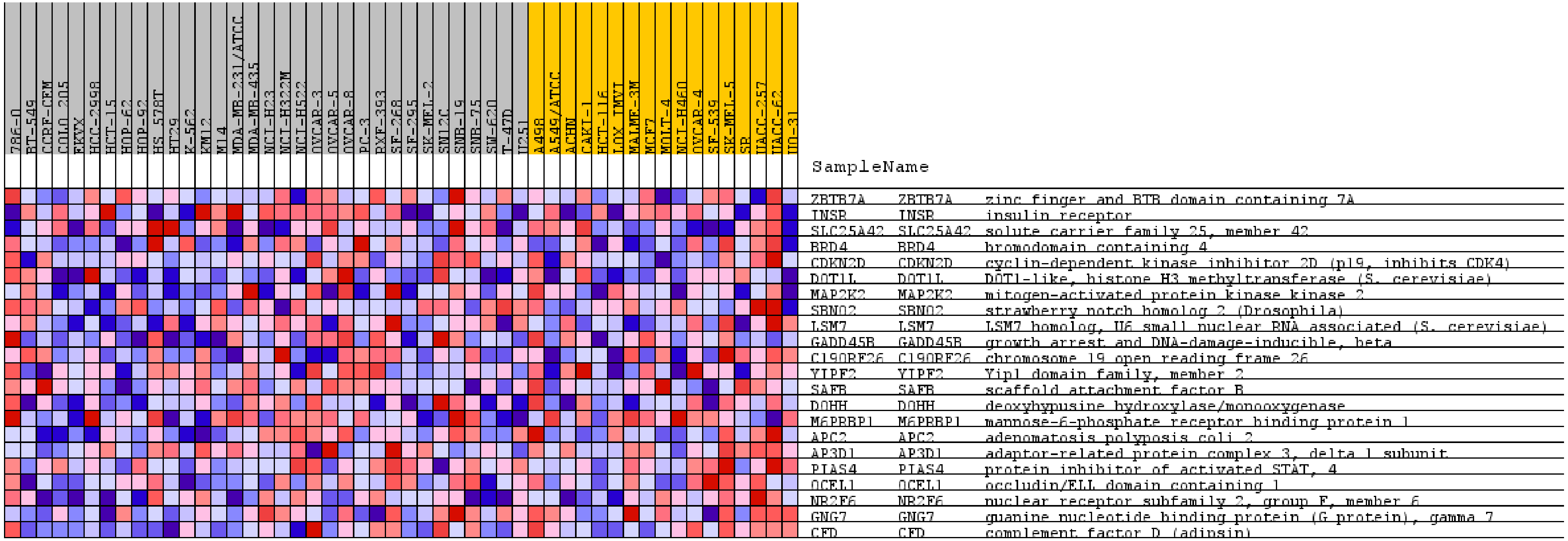
…ATCGCGTACTTCATGTGAGGG…

…ATCGTGTACTTCGTGTGAGTG…

…TACTCGGTAGCTATGCAGTGT…

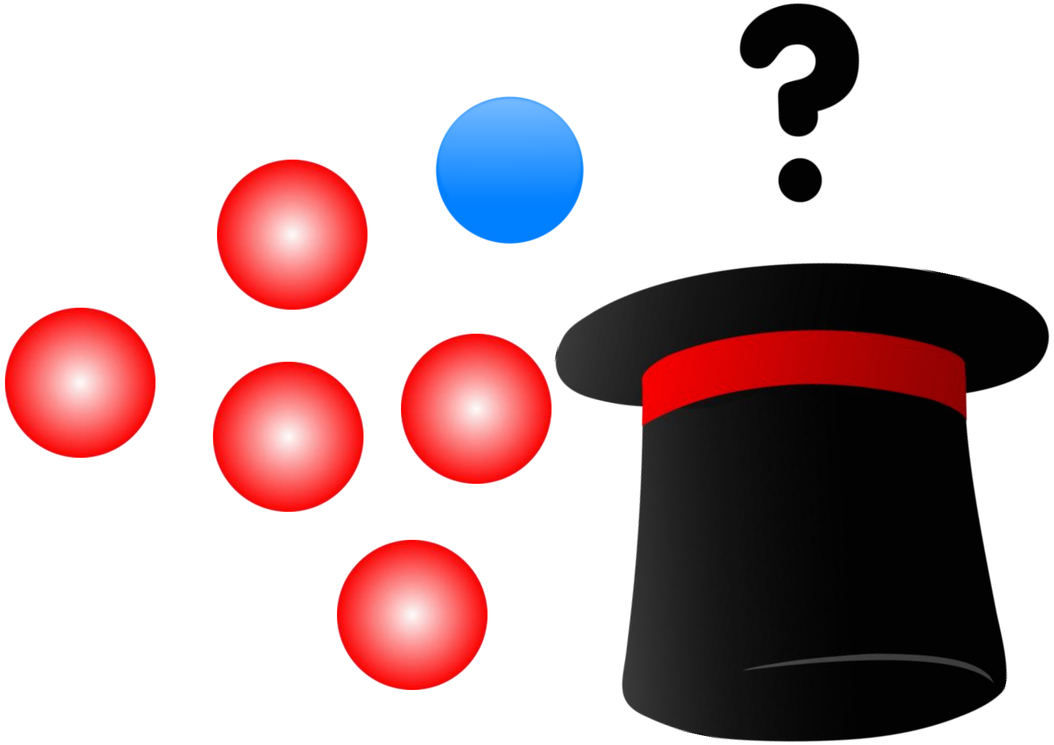
Analysis of Gene Sets

- Gene Ontology (GO) consortium classification
  - Assigns biological significance
- Gene Set Enrichment Analysis (GSEA) finds hidden patterns



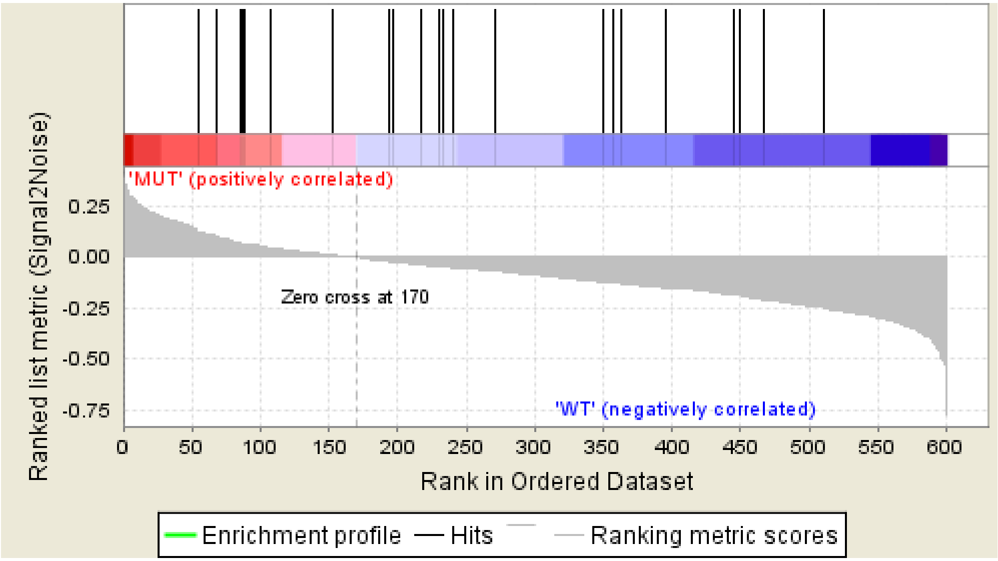
Gene Set Enrichment Analysis

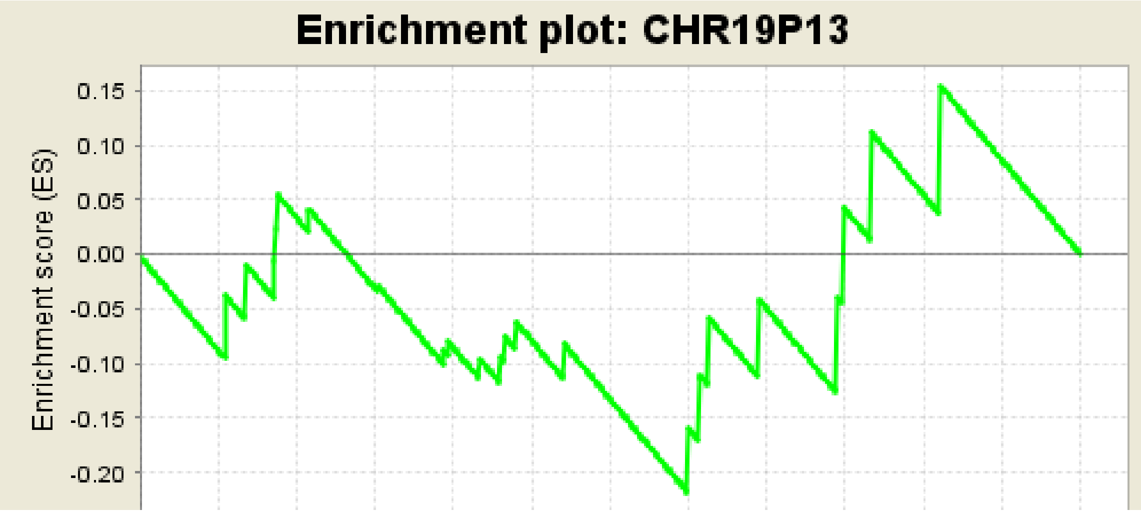
- In gene expression profiles
  - no individual gene may be statistically significant
  - significant genes with no biological theme
  - effects on pathways improperly described



Gene Set Enrichment Analysis

- Rank expression data set
- Running sum down gene list
- Enrichment Score (ES) calculated for each gene set

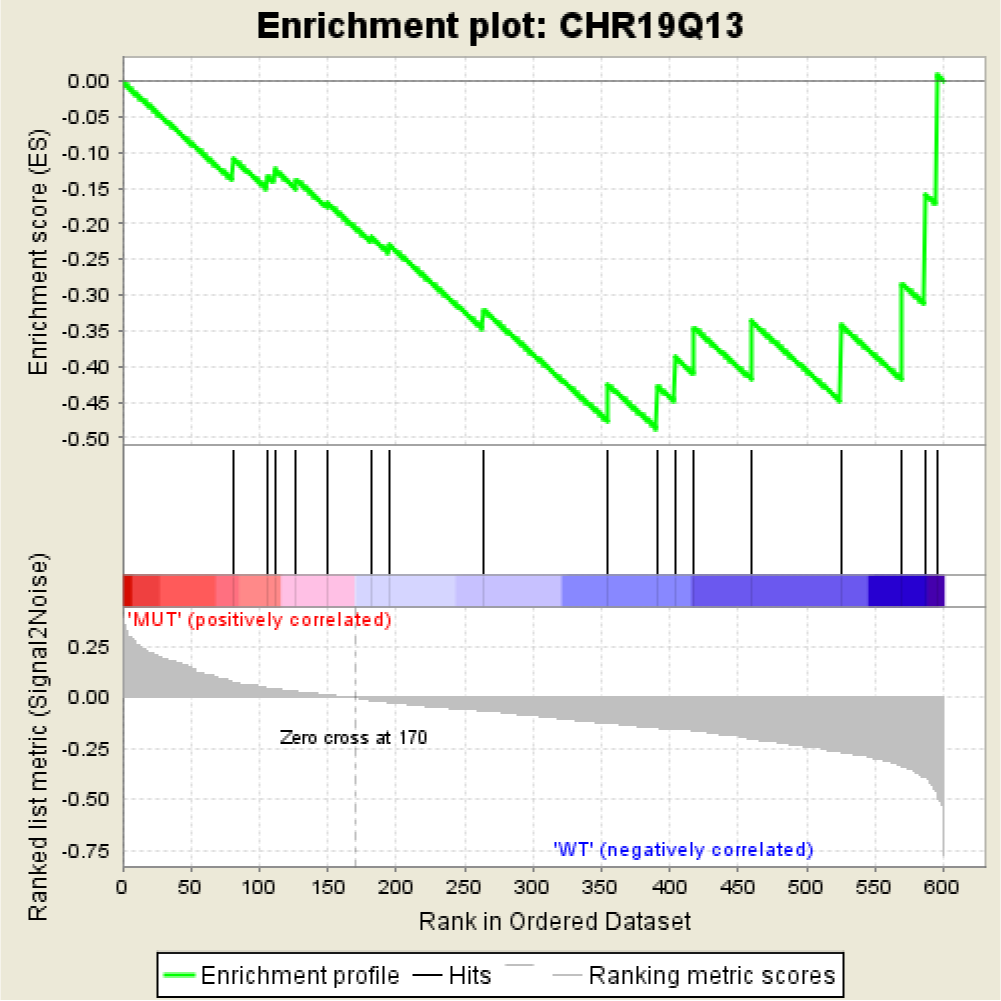




- Phenotype labels permuted and ES calculated vs null
- Nominal p-value estimation

VCF Enrichment Analysis

- VCF files require preparation
  - Annotate with GO labels
- Running sum down VCF labels to calculate ES
- Similarity score calculated using Kolmogorov-Smirnov test



Future Work

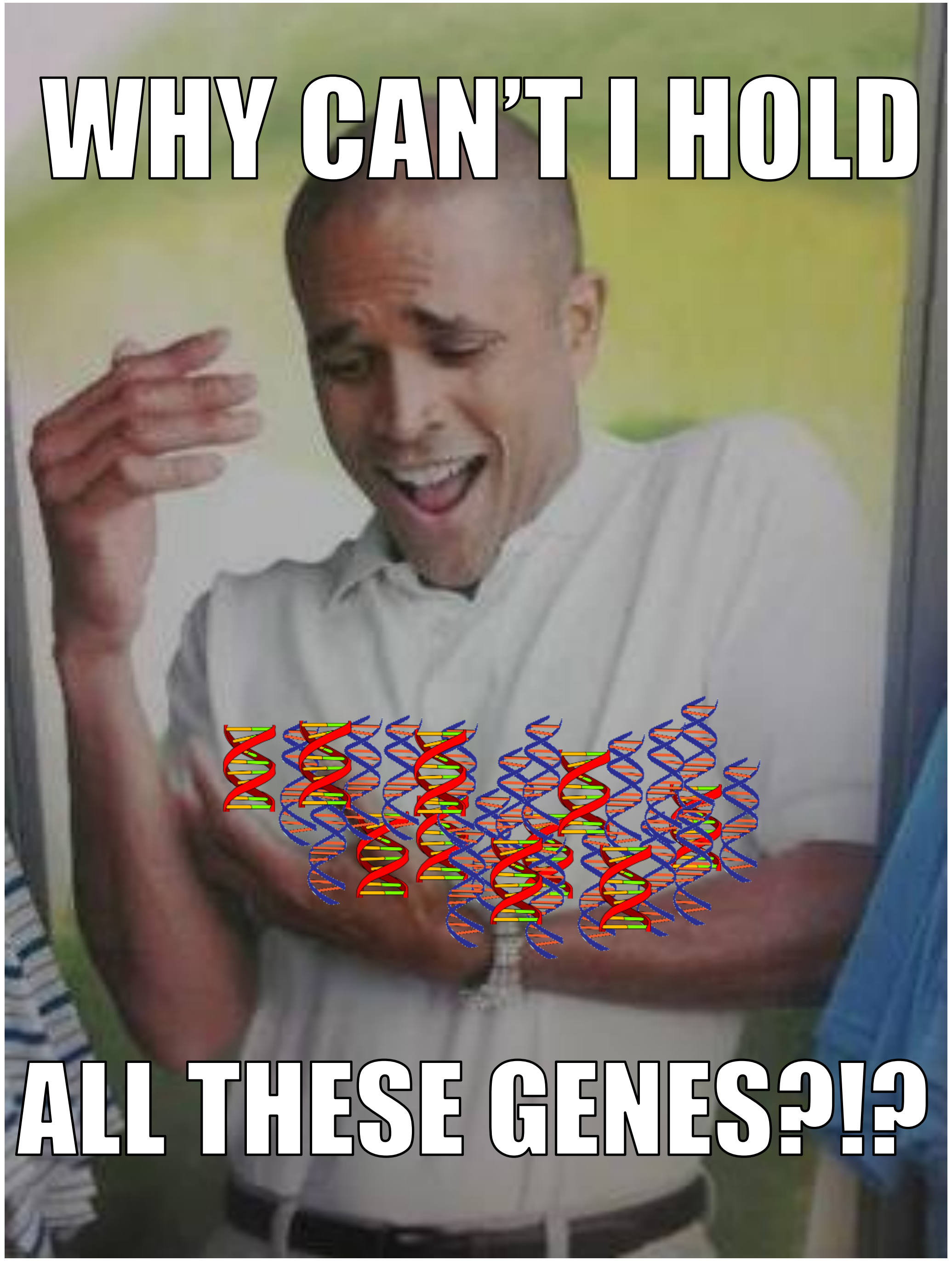
- Analyze more genomic data, including FASTA, to be able to find more potential biological patterns
- Improve the metric for comparing VCFs to be more robust and include various genomic data types
- Experiment with convolutional network layers to be able to take advantage of biological patterns
- Investigate this process to test whether biological patterns exist in other genomic datasets

Conclusions

- Genomic data is difficult to use without compression
- Optimal compression methods have to use intrinsic patterns
- Gene sets allow for quantification of biological significance and patterns
- Labeling VCF files allow for us to analyze intrinsic patterns
- Current algorithms are not able to show much compression towards VCF, but FASTA

Acknowledgements

- **MIT PRIMES** for allowing us to take on this exciting and challenging research opportunity
- **Dr. Gil Alterovitz and Maksym** for their guidance and support
- **Adithya, Kalyan, and other members of our group** for their previous work and assistance
- **My parents** for their support


WHY CAN'T I HOLD
ALL THESE GENES?!?