InfoSeq: Quantification of Information Content in genomic datasets

Motivation

- We usually discuss the predictability of tracks using each other
- Uniform analysis of predictability of the data can reveal basic understanding of how genomic tracks relate to each other
- **Quantity:** How do we quantify the amount of information in genomic datasets?
 - Which datasets give the most amount of information?
- **Predictability:** How do we quantify (correct) predictability of one dataset from other datasets?
 - Which datasets can be predicted best from the others?

Entropic Measures from PrivaSeq

• Given *N* tracks (random variables) *X*₁, *X*₂, ..., *X_N*; we compute the predictability of *Y* as:

$$1 - \frac{H(Y|X_1, X_2, \dots, X_N)}{H(X_1, X_2, \dots, X_N, Y)}$$

where $H(X_1, X_2, ..., X_N, Y)$ is the joint entropy and $H(Y|X_1, X_2, ..., X_N)$ is the conditional entropy of Y given other tracks

• We estimate the entropies using the observed datasets as the sample entropies. Thus the quantities that we compute are data dependent

Entropic Measures from PrivaSeq

- Assign: $X_1 = H3K4me1, X_2 = H3K4me3, ..., X_N = H3K9me3$
 - H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3, H3K9me3, H3K9ac, H3K79me2, H4K20me1, Control,
 - Sequence indicators (A,C,G,T)
 - 1000 Genomes variants (<0.1 allele frequency)
 - Gene/Promoter/Exon indicators
 - Randomized Control (random Bernoulli track): Must have zero predictability.
- Bin the signal tracks with 1,000 base pair long bins and generate the sample for each track
- Use $\lfloor \log_{10}(total \ signal) \rfloor$ to generate the tracks
- Estimate the sample entropies from histograms

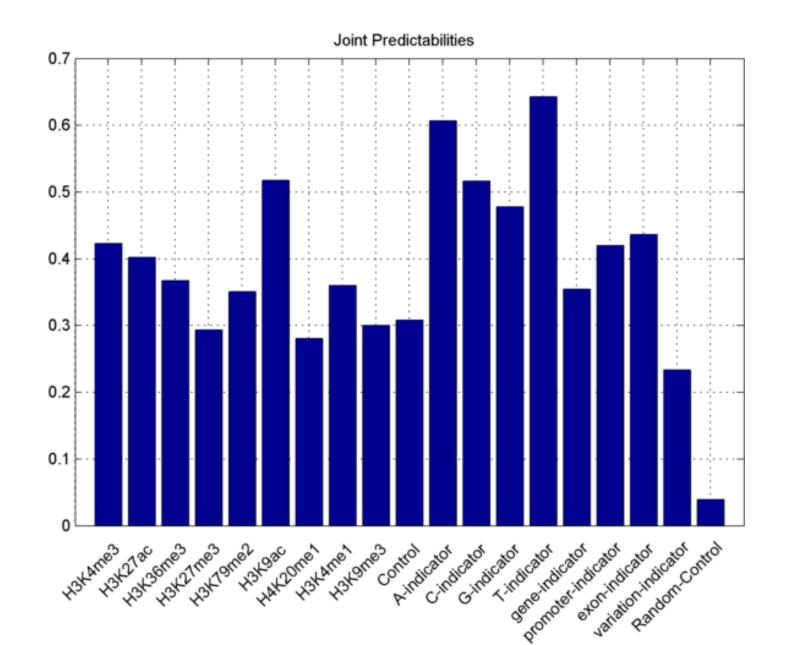
Predictability of Tracks Using Each of the others

Pairwise Predictabilities H3K4me3 H3K27ac 0.9 H3K36me3 H3K27me3 0.8 H3K79me2 H3K9ac 0.7 H4K20me1 H3K4me1 0.6 H3K9me3 Control 0.5 A-indicator 0.4 C-indicator G-indicator 0.3 T-indicator gene-indicator 0.2 promoter-indicator exon-indicator 0.1 variation-indicator Random-Control 0 Control cator Valiation Indicator PandomControl promoter indicator exonindicator geneindicator H3K9me3 Gindicator T-indicator Cindicator HAK20me1 Hatamet

Predicted Track

Predicting Track

Joint Predictability (one from others)



Next..

- *Effect of Data Representations:* How does different representations of the data effect quantification of information?
 - Peaks versus signal tracks
- How to spin this to not overlap with other efforts in ENCODE?