Regulatory Annotations [@ Mark, 9']
• Most cancer mutations fall in non-coding regions
• Coding/noncoding ratios vary across cancer types
• Exact number of non-coding mutations in “functional regions” depend greatly annotation def'n (in contrast, to situation with coding)

Many non-coding mutations, but interpretation very dependent on definitions
Non-coding Annotations: Overview

Several collections of information "tracks" related to non-coding features

**Sequence features, incl. Conservation**

- Large-scale sequence similarity comparison
- Identify large blocks of repeated and deleted sequence:
  - Within the human reference genome
  - Within the human population
  - Between closely related mammalian genomes
- Identify smaller-scale repeated blocks using statistical models

**Functional Genomics**

ChIP-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription

Signal processing of raw experimental data:
- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks

A) PEAKS
1. DNase peaks at the UCSC genome browser (on many cell lines)
2. The regulation track at the UCSC genome browser, with compilation of TF ChIP-seq peaks from uniform processing (individual peaks are annotated with TF and cell line)
3. Blacklist Regions

B) PROMOTERS
Annotated GENCODE TSSes (also, TSSes with FANTOM CAGE support)

C) ENHANCERS (Supervised) **

D) UNSUPERVISED SEGMENTATIONS, (Including enhancers) **
ChromHMM, SegWay, HiHMM....

E) HOT/LOT REGIONS
1. Enhancer-target gene connection
2. TF-target network connectivity
3. TADs: Topologically Associated Domain

F) CONNECTIVITY **

G) MOTIFS & MODELS
for TF binding

H) RNA
1. A matrix of expression data of known genes (or exons) for protein-coding genes & known ncRNAs (on many cell lines)
2. Novel RNA contigs track, i.e., possible novel transcripts (ie Transcriptionally Active Regions or TARs)
3. Novel junctions

[ preliminary gdoc https://docs.google.com/spreadsheets/d/1zRQJ6O7Cqzqyf6K28pEY3BvcOp7bz__tF0mnWclbylQ/edit#gid=104785709 ]
Supervised v Unsupervised Enhancer Sets

- Unsupervised: Group the His mark patterns at regions in the genome in ~20 characteristic states, some of which may correspond to enhancers. Have an annotation for each base in genome.
- Supervised: Learn a pattern of His marks associated with enhancers, with a gold standard. Then apply to the genome.

[Hoffman et al ('12), Nat. Methods; Ernst & Kellis ('12), Nat. Methods; Yip et al. ('12), Gen. Biology]
Highly simplified regulatory sites

- **Candidate enhancers:** The master list of TSS-distal DHS peaks annotated with
  - H3K27ac enrichment (percentile over background) in a cell-type-specific manner.
  - TF ChIP-seq peaks across cell-types

- **Candidate promoters:** The master list of TSS-proximal DHS peaks annotated with TF ChIP-seq peaks across cell types.
Relating Non-coding Annotation to Networks & Protein-coding Genes

Regulatory elements

Assigning proximal sites to target genes

Assigning distal sites to targets

~500K Edges

~26K Edges; ~5K per cell line

[ Cheng et al., Bioinfo. ('11); Gerstein et al. Nature ('12) ]
Associating enhancers (& generalized "DRMs") with target genes

- **Motivation:** Histone modifications predict gene expression levels

  1. Find correlated enhancer-target pairs
  2. Find TFs binding enhancers in cell lines with strong HMs
  3. Draw distal edges from TFs to targets

- Among **~2.2M putative enhancers** (found using supervised methods), significant associations between **~332K of them & ~17K genes**

[Yip et al., *Genome Biology* (2012); Fu et al., *Genome Biology* (2014)]
Some Key Issues for Cancer Genomics Annotation

• For our "enhancer set", should we use the **encode simplified one** (H3K27ac), a specific one based on machine learning, or an extraction from segmentation?

• Should we use a **universal or cell-line specific annotation**? How do we match cell-line specific annotations to particular cancers?

• Given the uncertainties in variant calling, particularly related to rearrangements, should we use very specific & **small regulatory annotations (eg motifs)**?

• Is it useful to **associate as much as possible of the non-coding genome with genes**, particularly known cancer genes (eg via weak enhancer target linkages)?