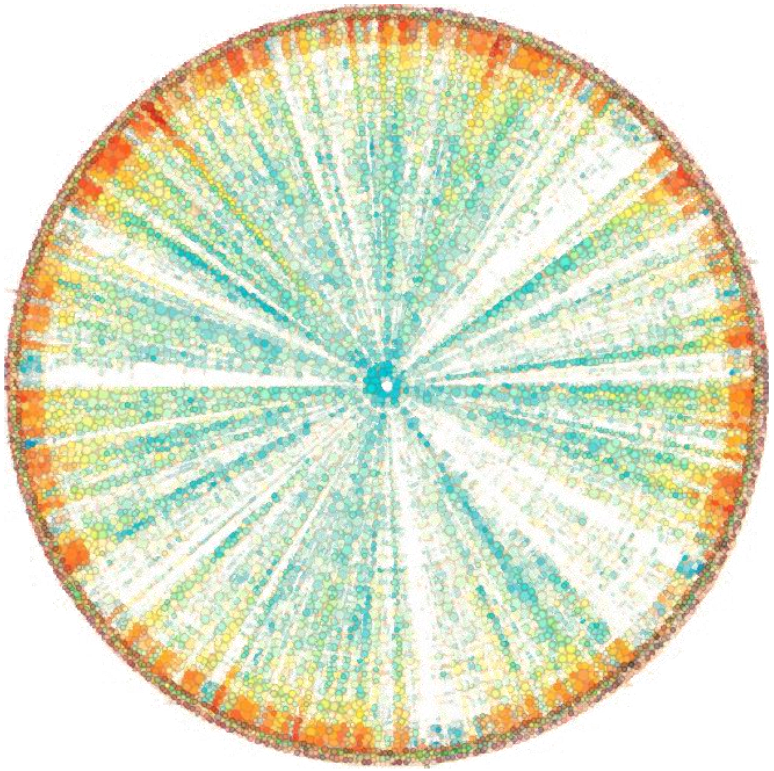
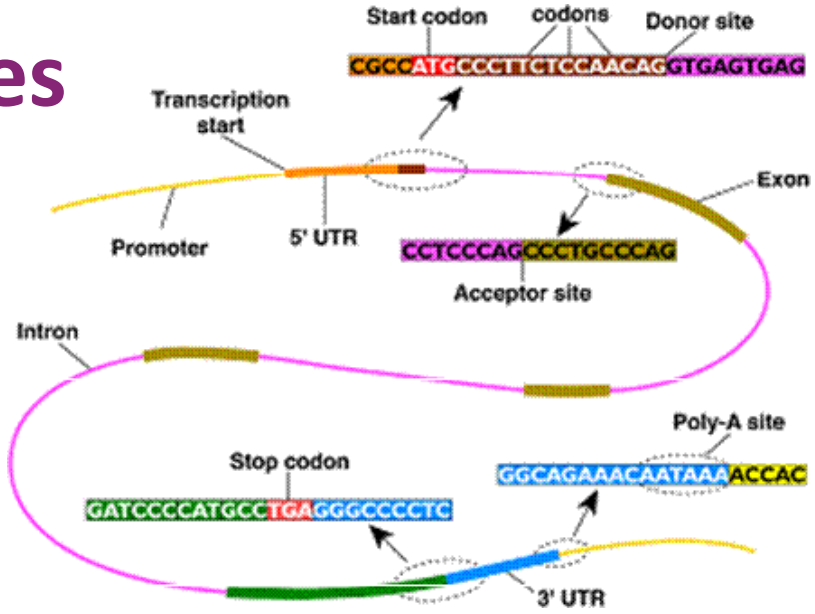
Comparative Genomics and Visualisation

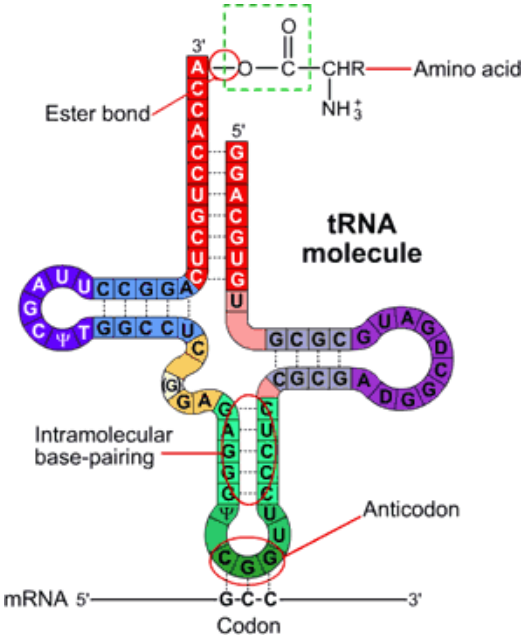
- **Genome Features/Functional Components**
  - numbers and types of features (genes, ncRNA, regulatory elements, etc.)
  - organisation of features (synteny, operons, regulons, etc.)
  - complements of features
  - selection pressure, etc.



Genome Features

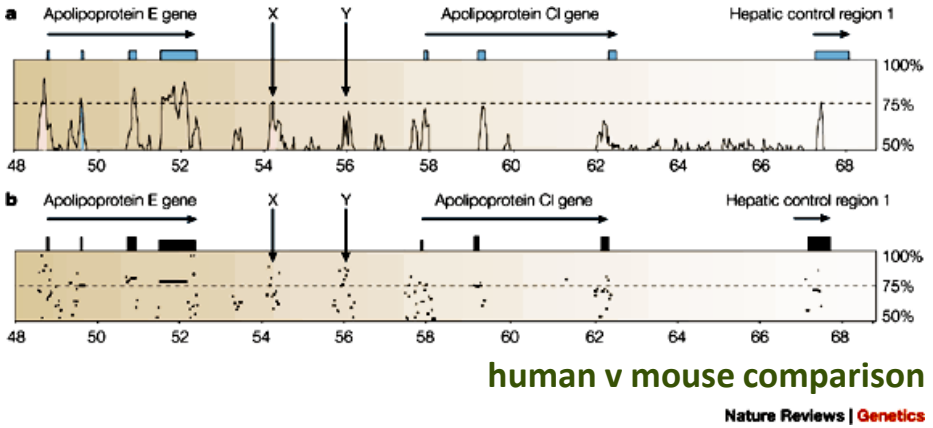
- **Genes:**
  - translation start
  - introns
  - exons
  - translation stop
  - translation terminator
- **ncRNA:**
  - tRNA – transfer RNA
  - rRNA – ribosomal RNA
  - CRISPRs – bacterial and archaeal defence (genome editing)
  - many other classes (including enhancers)





Genome Features

- **Regulatory sites**
  - Transcription start site (TSS)
  - RNA polymerase binding sites
  - Transcription Factor Binding Sites (TFBS)
  - Core, proximal and distal promoter regions



human v mouse comparison

Nature Reviews | Genetics

- **Repetitive Regions and Mobile Elements**
  - Tandem repeats
  - (retro-)transposable elements
    - *Alu* has ≈50,000 active copies in human genome
  - Phage inclusion (bacteria/archaea)

Pennacchio & Rubin (2001) *Nat. Rev. Genet.* doi:10.1038/35052548

Genome Feature Identification

- **Gene Finding**:
  1. Empirical (evidence-based) methods:
     - Inference from known protein/cDNA/mRNA/EST sequence
     - Inference from mapped RNA reads
  2. *Ab initio* methods:
     - Identification of sequences associated with gene features:
       - TSS, CpG islands, Shine-Dalgarno sequence, stop codons, etc.
  3. Inference from genome comparisons/conservation

Liang *et al.* (2009) *Genome Res.* doi:10.1101/gr.088997.108
Brent (2007) *Nat. Biotech.* doi:10.1038/nbt0807-883
Korf (2004) *BMC Bioinf.* doi:10.1186/1471-2105-5-59

Genome Feature Identification

- **Finding Regulatory Elements (short, degenerate)**:
  1. Empirical (evidence-based) methods:
     - Inference from protein-DNA binding experiments
     - Inference from coexpression
  2. *Ab initio* methods:
     - Identification of regulatory motifs (profile/other methods):
       - TATA, sigma-factor binding sites, etc.
     - statistical overrepresentation
     - Identification from sequence properties
  3. Inference from sequence conservation/genome comparisons

Zhang *et al.* (2011) *BMC Bioinf.* doi:10.1186/1471-2105-12-238
Kilic *et al.* (2013) *Nucl. Acids Res.* doi:10.1093/nar/gkt1123
Vavouri & Elgar (2005) *Curr. Op. Genet. Devel.* doi:10.1016/j.gde.2005.05.002

Genome Feature Identification

- All prediction methods result in errors
- All experiments have error
- Genome comparisons can help correct errors
- **[OPTIONAL ACTIVITY] – useful for exercise**
  - `predict_CDS.md` Markdown
- Other options for prokaryotic genecalling:
  - Glimmer (http://ccb.jhu.edu/software/glimmer/index.shtml)
  - GeneMarkS (http://opal.biology.gatech.edu/)
  - RAST (http://rast.nmpdr.org/)
  - BASys (https://www.basys.ca/), etc.
- Options for eukaryotic genecalling:
  - GlimmerHMM (http://ccb.jhu.edu/software/glimmerhmm/)
  - GeneMarkES (http://opal.biology.gatech.edu/gmseuk.html)
  - Augustus (http://augustus.gobics.de/), etc.