Practical 2 – Read QC – Removing adaptors

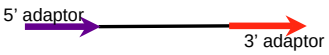
Get data file '/local/data/genome informatics/practical_2/part2.tar.gz'

Unpack with *tar zvxf part2.tar.gz*

Files:

- 1m Illumina sequencing reads — 1m_raw_short_reads
- Adaptor sequences — adaptors.seq

N.B. The 5' adaptor is also the sequencing primer



1) Write a simple program to find and remove any 3' adaptor sequences from the reads.
   What issues should you consider?
2) Investigate the complexity of the dataset (graph)
   How can the dataset be represented more compactly?

Annotation

- What is a gene?
- Gene Structure
- Gene Finding
  - By alignment / homology

What is a gene?

Human Genome Organization

"A DNA segment that contributes to a phenotype or function. In the absence of demonstrated function, it may be characterized by sequence, transcription or homology."

Sequence Ontology Consortium

"A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions."

Types of gene

protein-coding genes ⟶ protein

non-coding genes → structural or regulating RNA

- tRNA
- rRNA
- miRNA
- piRNA
- snRNA

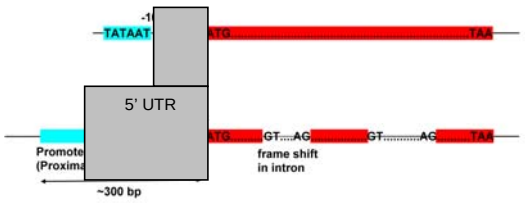
(pseudo genes)

Gene Structure - Prokaryote

- ORF
- Standard promoter sequence
  - Pribnow box: TATAAT

- Start Codon often embedded in a Shine-Dalgarno consensus sequence – Ribosome binding site

Gene Structure - Eukaryote



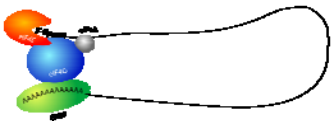
- Varying promoter sequences
  - TATA box – 24% of human genes
- Start Codon often embedded in a Shine-Dalgarno (bacteria) or Kozak (eukaryotes) consensus sequence – Ribosome binding site

UTRs – 5’ - Transcription Start Site - Leader Sequence

- Experimentally often poorly defined.
  - usually: longest cDNA/EST
- Forms complex secondary structure to regulate translation

UTRs – 3’

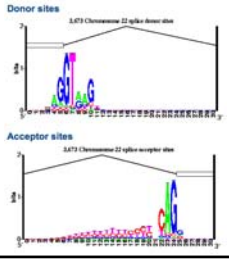
- Contains poly-adenylation signal.
  - 70% of all pre-mRNAs contain AAUAAA (variants exist).
- 3’ UTRs are the target sequences for microRNAs (in animals)



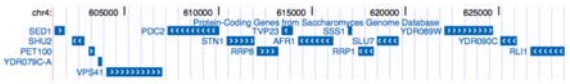
- 3’ UTRs contain additional regulatory sequences for transcript stability and/or translation efficiency.

Splice sites

- different molecular mechanisms of splicing exist.
- splicing is a tightly controlled process
- donor/acceptor site consensus sequences are highly conserved between species.

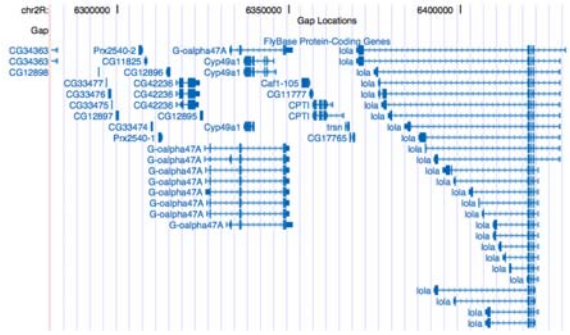


Genome complexity



Saccharomyces cerevisae

Genome complexity



Drosophila melanogaster

How to find human genes?

Via human cDNA or EST sequences
Via vertebrate cDNA or EST sequences

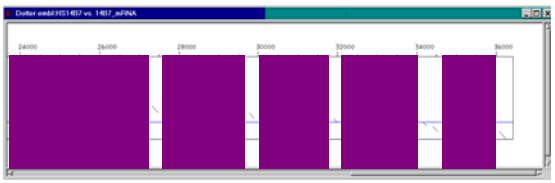
Finding similarity in genome to known proteins

Ab initio - using statistical gene finders.

Main genome annotation sites:

http://www.ensembl.org
http://genome.ucsc.edu

In practice, dynamic programming is used to piece together the exons once they have been roughly located



Same approach for *Protein* to genomic sequence alignment. Method must be tolerant of frameshifts, sequence errors, introns: e.g. GeneWise

Genomic Sequence

Miniseq 

ENSEMBL

Targeted Genewise

Human protein sequences
Swissprot/trembl/Refseq

pmatch give rough gene positions

blast and miniseq — rough exons

Genewise (intron, error, frameshift tolerant) — accurate exons

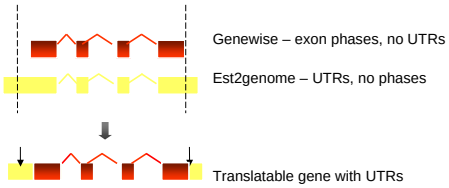
Untranslated Regions (UTRs)

Human and
vertebrate
cDNA sequences

blast and miniseq — rough exons

est2genome
accurate

Combining predictions



Ensembl gene annotation summary

Human protein sequences
Swissprot/trembl/Refseq

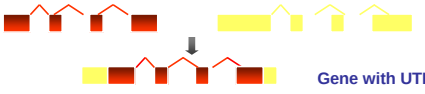
find protein matches

human and
vertebrate cDNA
sequences

blast and miniseq

genewise

est2genome



Gene with UTRs

References

http://www.ebi.ac.uk/Tools/psa/genewise/help/

- Birney *et al.* - GeneWise and Genomewise – Genome Research 2004

http://bioweb2.pasteur.fr/docs/EMBOSS/est2genome.html

- Mott, R. - EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. - Comp. Appl. Biosci. 1997