Annotation

- Gene Finding
  - Non-coding genes
  - Pseudo genes
- Transcript
  - Assembly
  - Abundance

Types of gene

protein-coding genes → protein

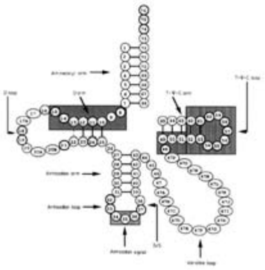
non-coding genes → structural or regulating RNA

tRNA
rRNA

miRNA
piRNA
snRNA

(pseudo genes)

tRNA



- Partial structure prediction
  - Quick, high FDR
- Full structure prediction
  - Slow, v.accurate
- tRNAscan-SE
  Prefilter then $O(MxN^2)$

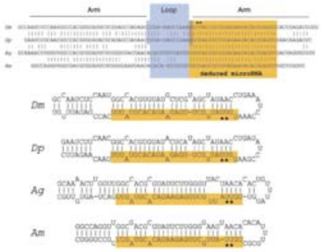
miRNA

- pre-miRNA forms stem-loop structure

- conserved across species
- miRNA homologous to target genes' 3' UTR
- Annotation via:
  - Secondary structure prediction

  e.g. MiPred, TripletSVM
  - Machine learning from training set

  e.g. miRAlign, MiRscan, miRDeep

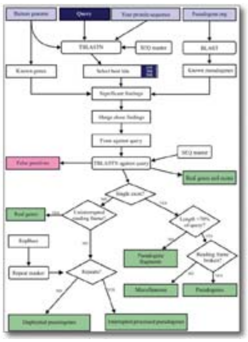


Hu *et al.* - Benchmark comparison of ab initio microRNA identification methods and software – Genet Mol Res 2012

Pseudogenes

- Often resembling (imperfect) copies of actual genes in the genome
- May have lost and/or never obtained:
  - complete ORF (they have many stop codons)
  - functional splice sites
  - regulatory regions
- May be transcribed or not

Pseudogenes

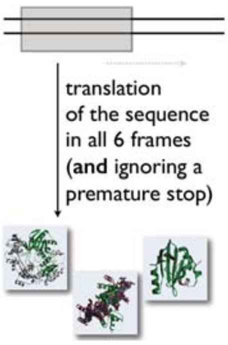


Strategy 1:

Search for sequence similarity to known genes, check for stop codons

taken from:
Ortutay & Vihinen
BMC Bioinformatics (2008)

Pseudogenes



translation of the sequence in all 6 frames (and ignoring a premature stop)

Strategy 2:

Search for similarity to known protein structures (and contain a stop).

does it resemble any of those structural domains?

Open Issues

- Coding exon prediction - accuracy
- Non-coding exon prediction (full length cDNA projects)
- polyA site prediction
- Transcription factor binding site identification (L11/12)
  - Low complexity, distance, chromatin, clustering
- Alternate splicing (High early estimates for gene number?)
- Pseudogenes
- Nested/ overlapping genes
- Small structural RNA genes e.g. miRNA
- Replication origins

See the ENCODE project:

http://genome.ucsc.edu/ENCODE/

Annotation

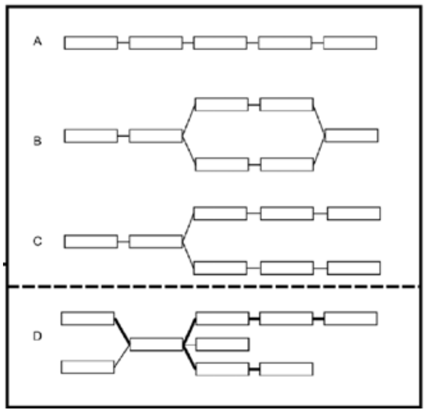
- Transcript
  - Assembly
  - Abundance

Oases – *de novo* transcriptome assembly

- Much wider variation of coverage than in genome

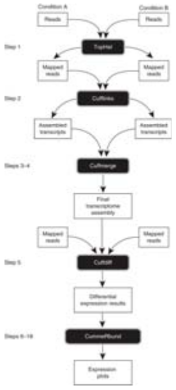
- Basically as velvet, but run multiple k-mer sizes
- Different filtering on graph – remove any low coverage nodes (<3x), nodes with low proportion coverage of predecessor (<10%)
- Bubbles may represent alternative splicing
- Produce possible lists of transfrags per k and merge via *de Bruijn* graph

- V. high memory consumption, alternatives: Trinity, TransABySS



Tuxedo pipeline

- TopHat
  - Aligns RNAseq reads to genome using bowtie
- Cufflinks
  - Assembles transcripts and calculates expression
- Cuffdiff
  - Calculates differential expression between conditions
- CummeRbund
  - Visualises expression data using R

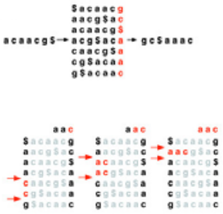


Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Trapnell *et al.* Nature Protocols

Bowtie 2

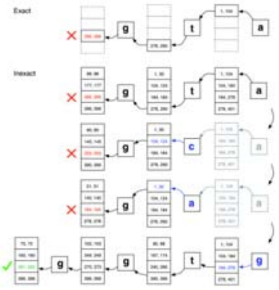
- V. efficient short read aligner used by Tuxedo
- Allows more scoring options, longer reads than Bowtie, mostly faster
- Indexes genome with Burrows–Wheeler transform (faster than hashing)
- Greedy, randomized, depth-first search

BWT



Bowtie

Match: ggta

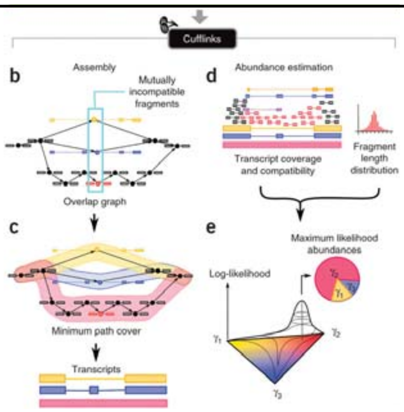


TopHat

- Aligns RNAseq reads to genome using bowtie
- Identifies ‘islands’ where entire reads pile up (exons)
- Looks for reads joining islands with splice sequences "GT-AG", "GC-AG" and "AT-AC“
- Can also attempt to join nearby islands with canonical splice junction

Cufflinks

- Assembles transcripts using reference genome
- Constructs a parsimonious set of transcripts that "explain" the reads
- Finds minimum path cover on the directed acyclic graph describing alignments


Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ1
γ2
γ3

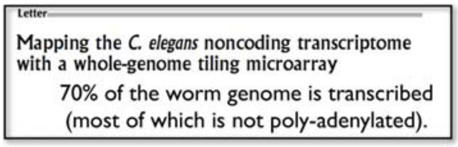
Abundance

- Estimate relative abundance of splice variants:
  - probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated
  - numerically maximizes a function that assigns a likelihood to all possible sets of relative abundances of the yellow, red and blue isoforms

- Per transcript abundance (expression) given in FPKM – Fragments per kilobase per million reads

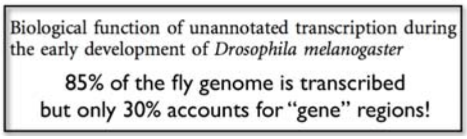
Human Genome Organization

"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."*

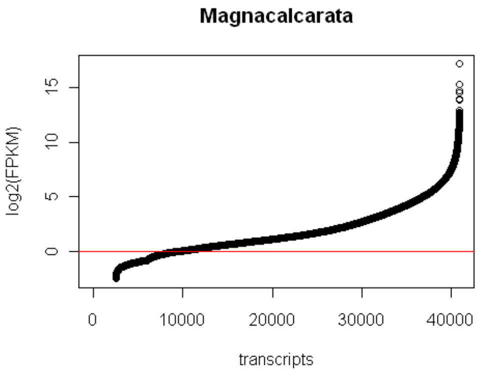

Letter

Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray

70% of the worm genome is transcribed (most of which is not poly-adenylated).

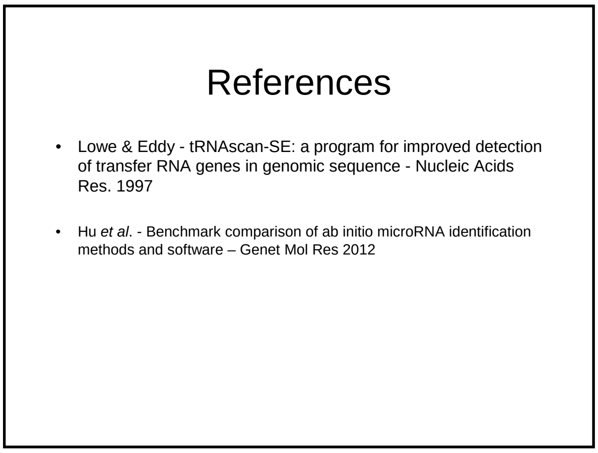

Biological function of unannotated transcription during the early development of *Drosophila melanogaster*

85% of the fly genome is transcribed but only 30% accounts for "gene" regions!



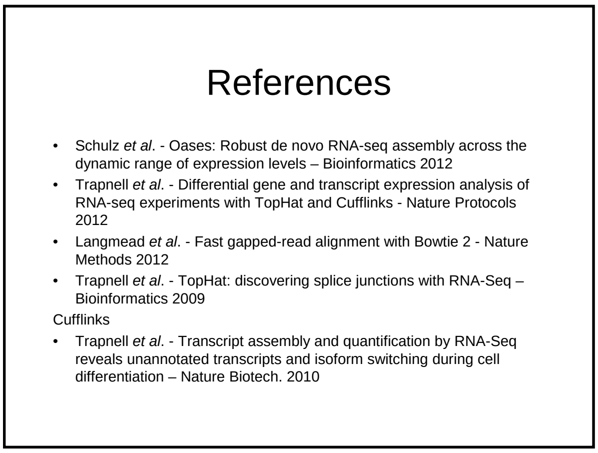
References

- Genes, gene structure, types of genes:
  - Genomes 3 by T. Brown or any other reasonable text book.
- Model genomic regions:
  - Ashburner *et al.* - An exploration of the sequence of a 2.9-Mb region of the genome of Drosophila melanogaster: the Adh region – Genetics 1999
  - ENCODE Project Consortium - The ENCODE (ENCyclopedia Of DNA Elements) Project – Science 2004
  - https://www.encodeproject.org/publications
  - http://www.modencode.org/publications/pubs



References

- Lowe & Eddy - tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence - Nucleic Acids Res. 1997
- Hu *et al*. - Benchmark comparison of ab initio microRNA identification methods and software – Genet Mol Res 2012

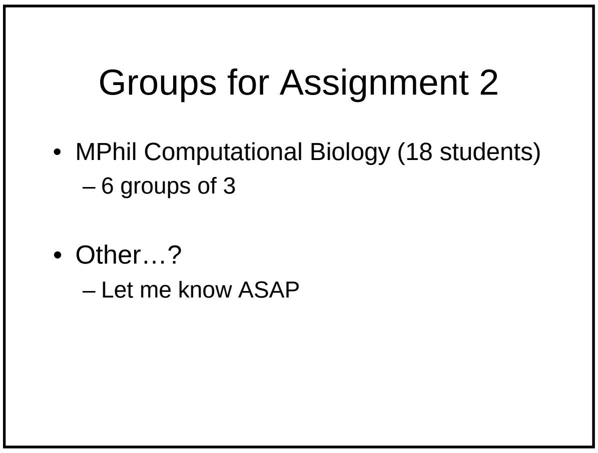


References

- Schulz *et al*. - Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels – Bioinformatics 2012
- Trapnell *et al*. - Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks - Nature Protocols 2012
- Langmead *et al*. - Fast gapped-read alignment with Bowtie 2 - Nature Methods 2012
- Trapnell *et al*. - TopHat: discovering splice junctions with RNA-Seq – Bioinformatics 2009

Cufflinks

- Trapnell *et al*. - Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation – Nature Biotech. 2010



Groups for Assignment 2

- MPhil Computational Biology (18 students)
  - 6 groups of 3
- Other…?
  - Let me know ASAP