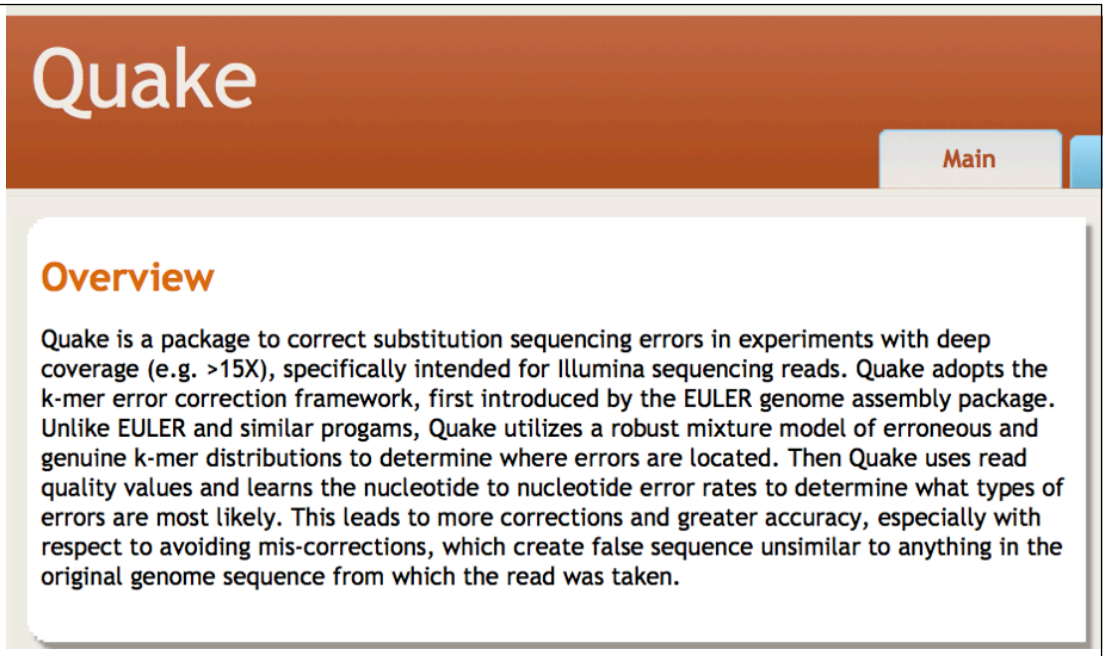


http://www.cbcb.umd.edu/software/jellyfish/
Guillaume Marcais and Carl Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (2011) 27(6): 764-770

http://www.cbcb.umd.edu/software/quake/
Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biology 11:R116 2010.

https://cutadapt.readthedocs.io/en/stable/

Cutadapt Removes Adapter Sequences From High-throughput Sequencing Reads

Marcel Martin

Abstract

When small RNA is sequenced on current sequencing machines, the resulting reads are usually longer than the RNA and therefore contain parts of the 3' adapter. That adapter must be found and removed error-tolerantly from each read before read mapping. Previous solutions are either hard to use or do not offer required features, in particular support for color space data. As an easy to use alternative, we developed the command-line tool cutadapt, which supports 454, Illumina and SOLiD (color space) data, offers two adapter trimming algorithms, and has other useful features.
Cutadapt, including its MIT-licensed source code, is available for download at http://code.google.com/p/cutadapt/

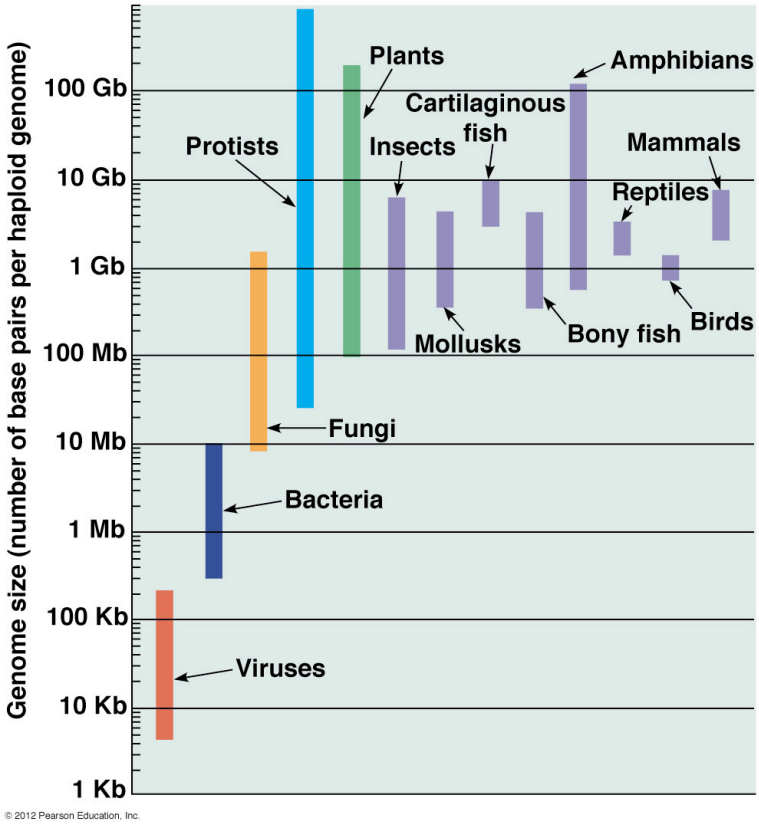
What's in a genome?

- Genome size
- Genome structure

Annotation

- Repeat Finding

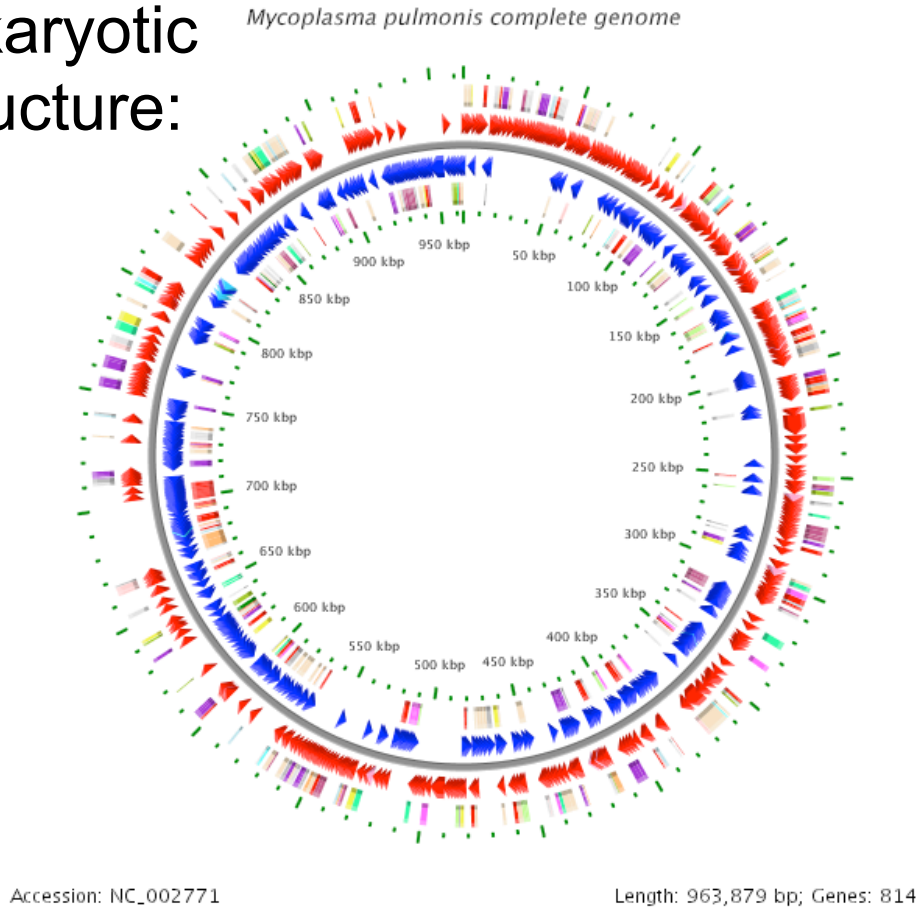
Comparative Genome Size



Organism	Genome size (Kb)	Approx no. genes	Notes
Human mitochondrion	16.5	37	
Epstein-Barr virus	172	80	Causes mononucleosis
Nanoarchaeum equitans	401	552	Parasitic Archaea, smallest known genome of a 'true' organism
Encephalitozoon cuniculi	2,508	1,997	Parastic eukaryote
Deinococcus radiodurans	3,254	3,157	2 chromosome, 2 plasmids; highly radiation resistant
Vibrio cholerae	4,033	3,800	2 chromosomes; causes cholera
Escherichia coli	4,639	4,377	4,290 of these are protein-coding
Saccharomyces cerevisiae	12,496	5,770	Budding yeast; eukaryote
Caenorhabditis elegans	100,258	20,532	First multicellular eukaryote to be sequenced
Arabidopsis thaliana	115,410	~25,000	Flowering plant
Drosophila melanogaster	122,654	13,927	Fruit fly
Tetraodon nigroviridis	342,420	27,918	Pufferfish; very compact genome
Rice	390,000	37,544	
Dog	2,400,000	19,300	
Human	3,300,000	20,769	

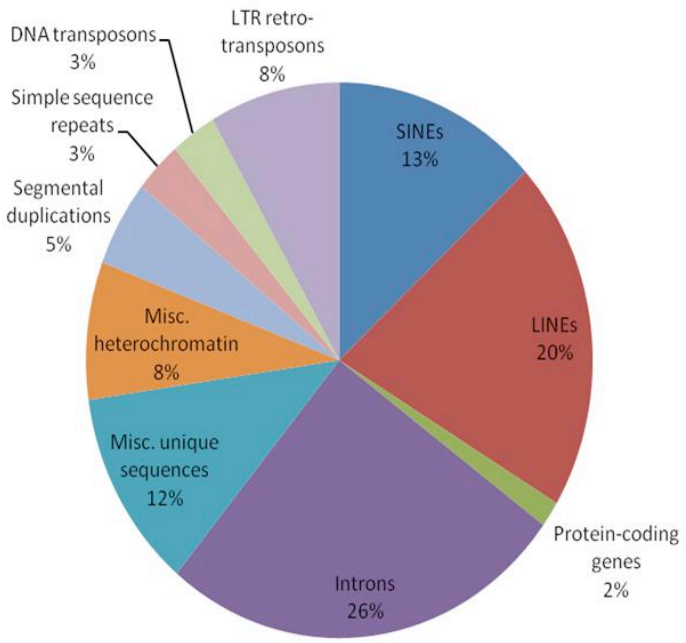
Typical prokaryotic genome structure:

- circular
- no introns
- high gene density



http://microbewiki.kenyon.edu/index.php/File:Genome.png

Human genome: not homogeneous



Two genomes:

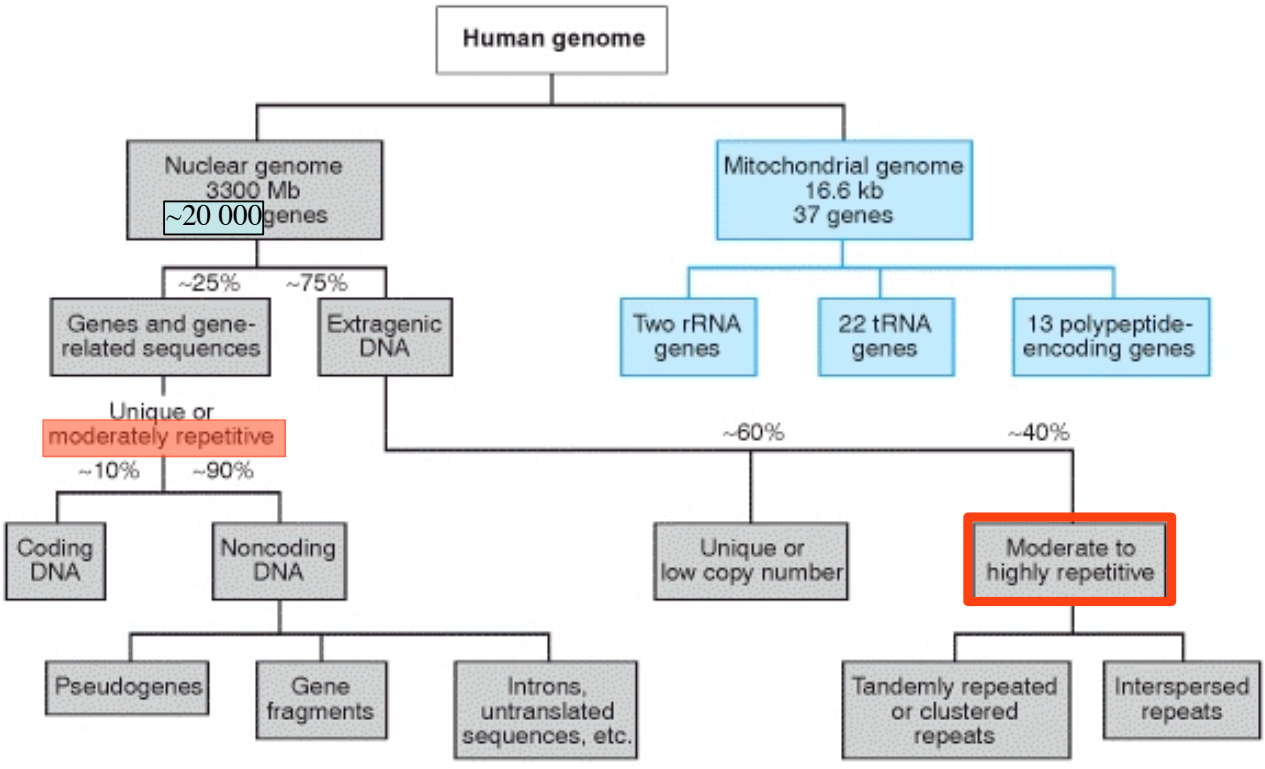
- Mitochondrial:
  16.6 kb - 37 genes
- Nuclear:
  3,300,000 kb - ~20,000 genes

Linear

Complex gene structure

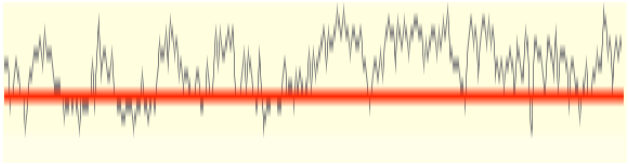
Low gene density

Human genome: not homogeneous

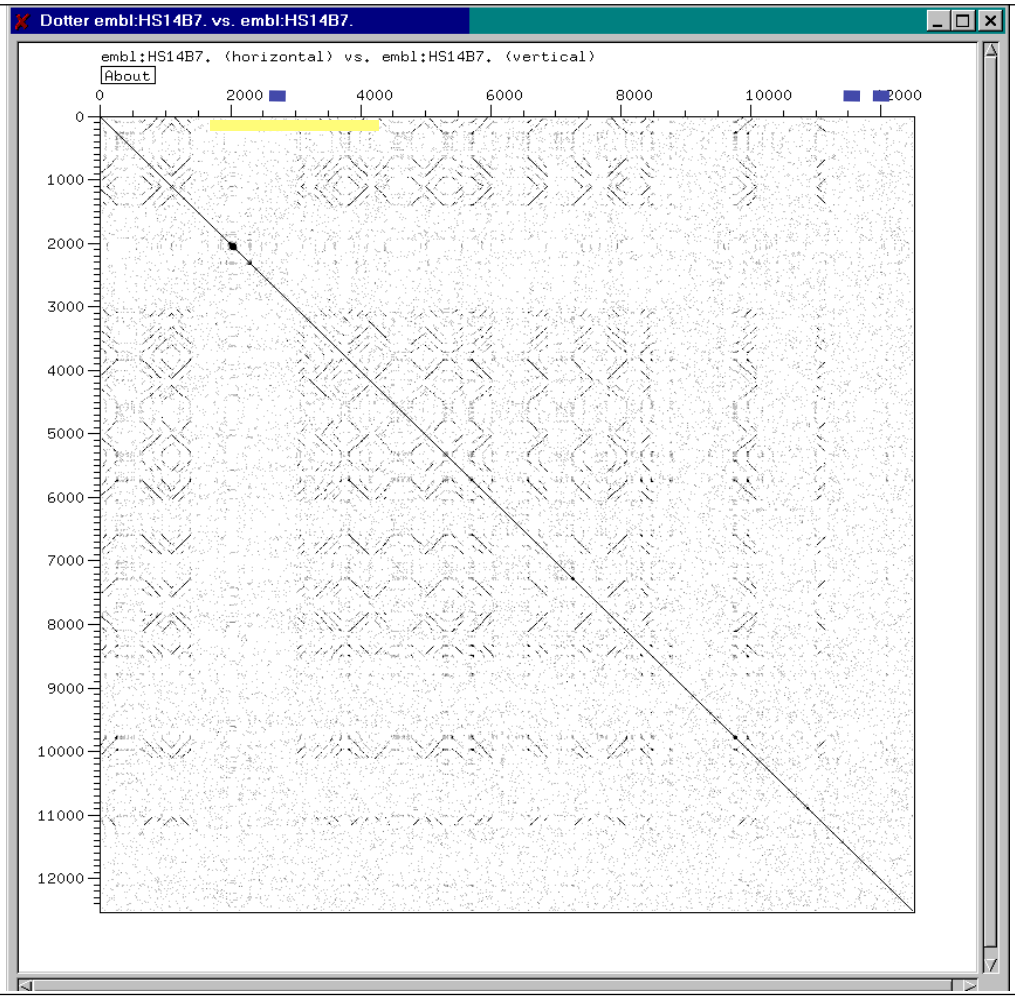


What's in a (Human) Genome? I

Varying %GC: 1Mb of 16p13.3

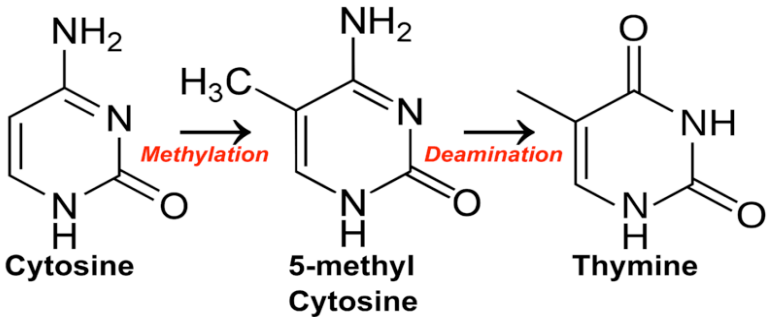


- 95% Intergenic sequence
  - ~50% genome is repetitive
    - Low complexity/ tandem repeats/ centromeric/ telomeric
  - Mobile elements (selfish DNA)
    - Transposons/ retroviruses
      - Long interspersed repeats (LINE)
        - » E.g. 6kb L1 LINE, old, often partial.
      - Short interspersed sequences (SINE)
        - » E.g. ~10e6 copies of 300bp ALU repeat, young, mostly complete.
- %GC; Alu density; gene density all related (LINE inversely)

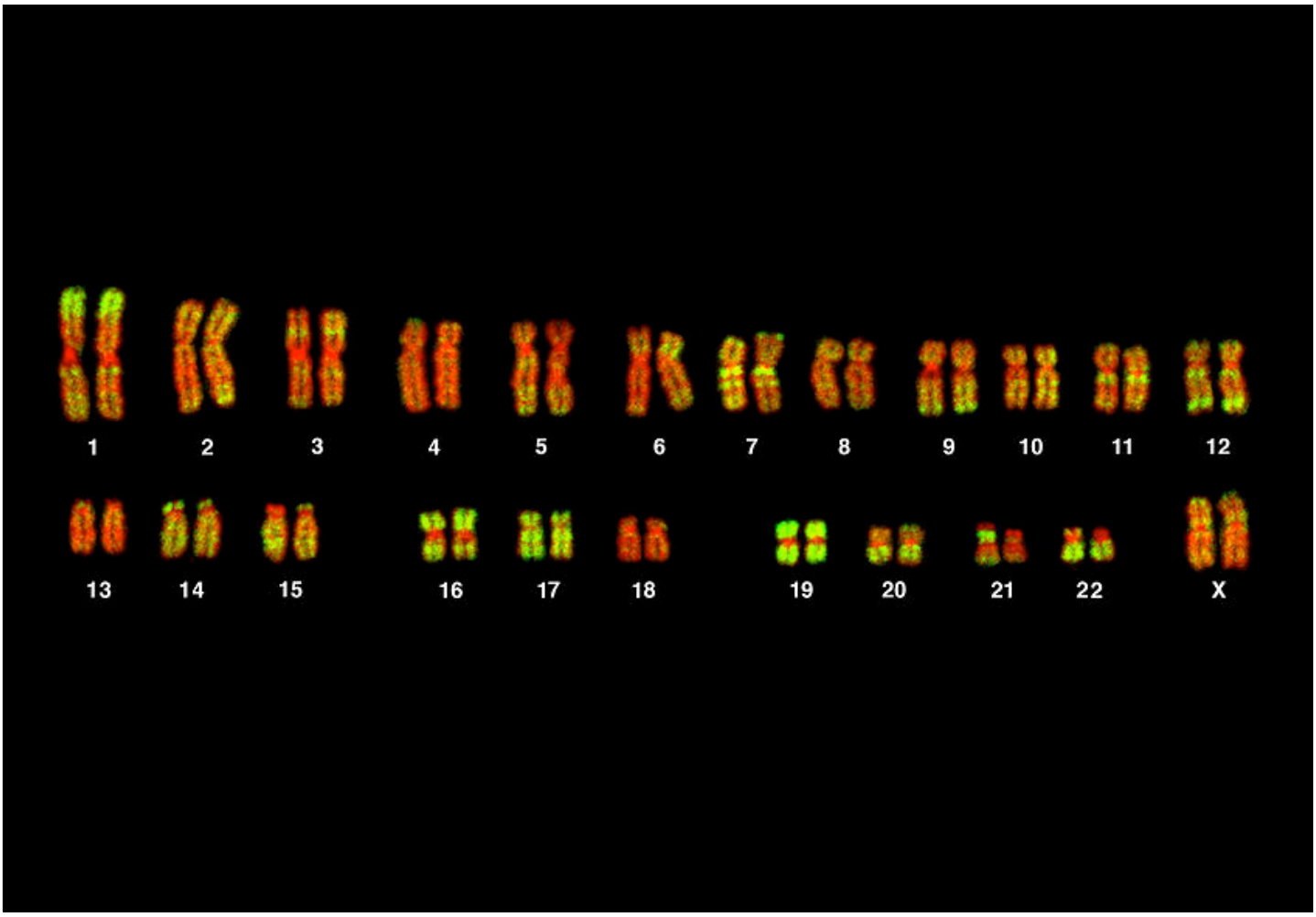


CpG Islands

- C modified by methylation to methyl-C
- Methyl-C frequently mutates to T

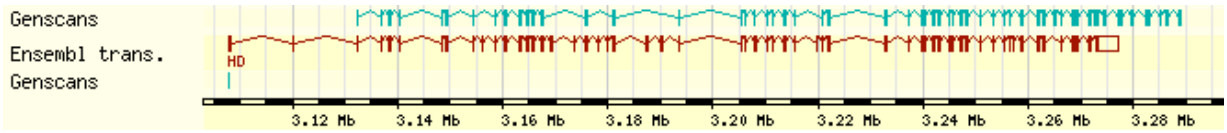


- CpG is normally ~5 fold under-represented from genomic GC% (in intergenic regions)
- CpG more frequent where methylation is suppressed (e.g. gene promoter regions)

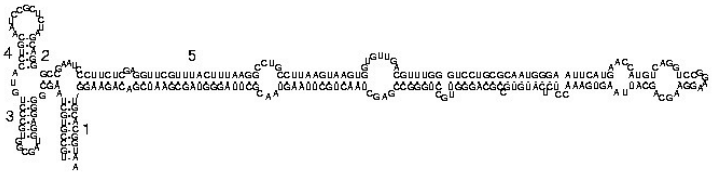


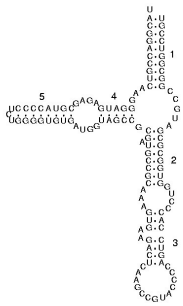
What's in a Genome? II

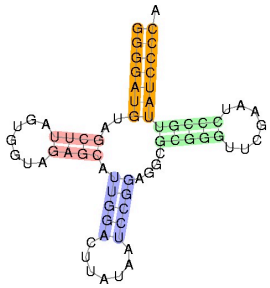
- Genes
  - Protein Coding Genes
    - Alternate splicing, nesting



  - Structural RNA genes
    - Ribosomal RNA
    - tRNA
    - Many other small structural RNAs
      - splicing, gene regulation
      - microRNA genes (miRNA)

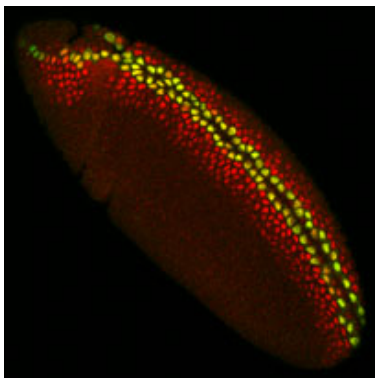






What's in a Genome? III

- Regulatory Sequences
  - Gene transcription
  - DNA replication
  - Gene splicing
  - (Gene re-arrangment: Ig, TCR genes)
  - (Chromatin packing)



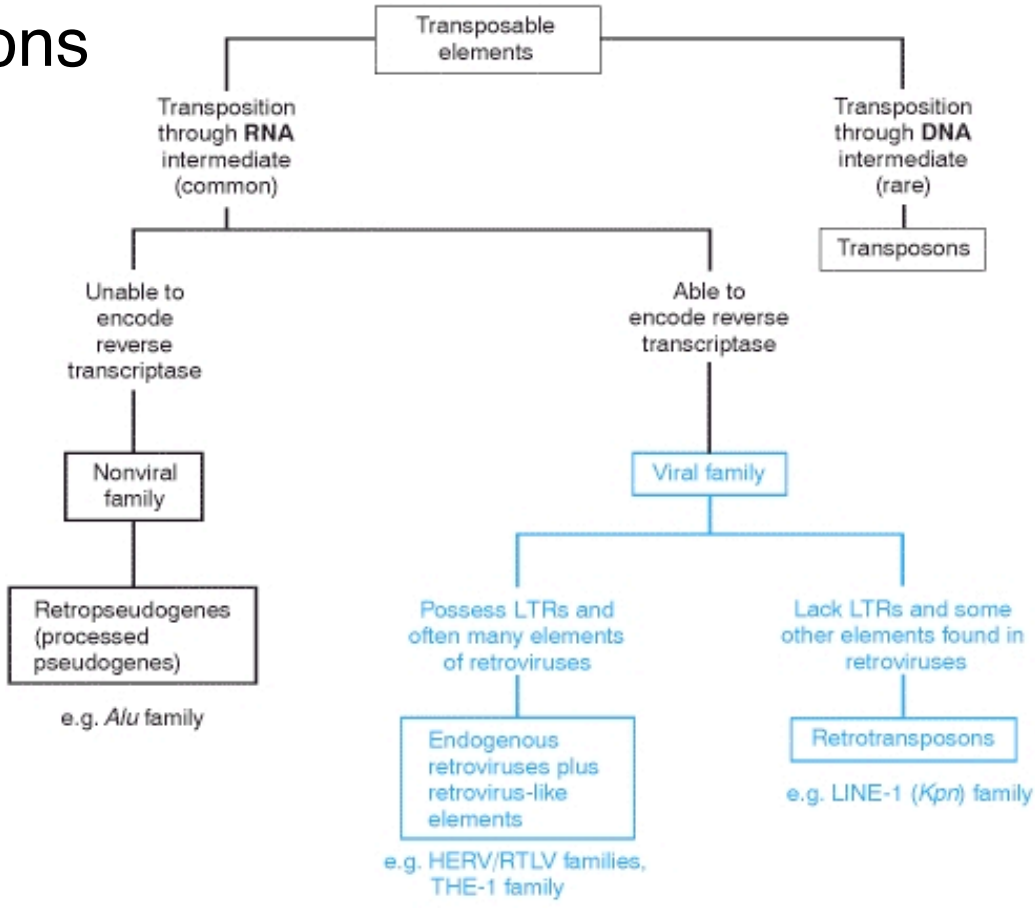
Signals are small

Signals are often conserved between organisms

Indian Muntjac

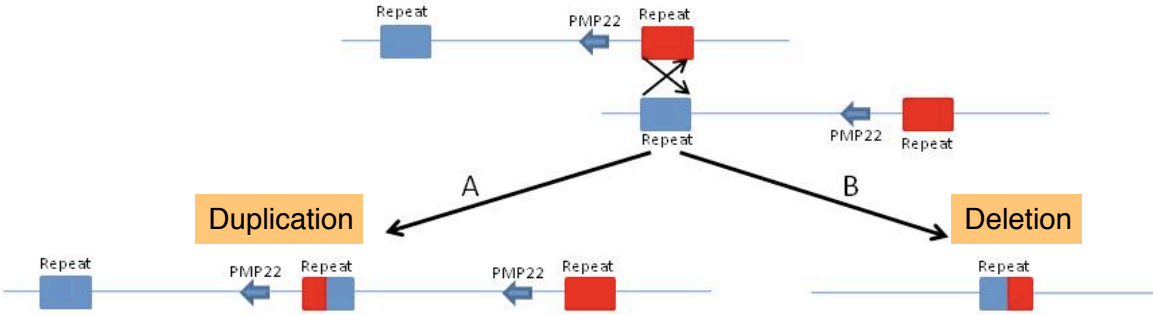
Chinese Muntjac

Transposons



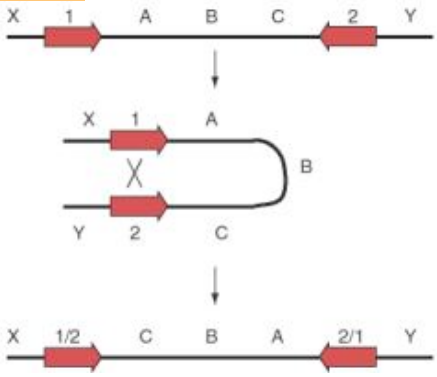
http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=hmg&part=A642

Recombination



http://www.ncbi.nlm.nih.gov/projects/dbvar/images/Information_fig1.png

Inversion



http://www.bx.psu.edu/~ross/workmg/TranspositionCh9_files/image046.jpg

Annotation

- Repeat Finding

Existing Databases

- RepBase
  - All types of repeats; actual sequence

  http://www.girinst.org/repbase/index.html

- Dfam
  - Alignments, HMMs and match lists of repeats

  http://dfam.org/

RepeatMasker

- Screens using modified RepBase library
- Alignment with engine similar to blast but optimised for repeats

- Masks ~50% of human sequence

de novo repeat finding

- Start from assembled genome: High-coverage k-mers, align, greedy extension e.g. RepeatScout
- Start from reads: calculate overlaps, generate graph – vertices repeats elements, edges overlaps and cluster e.g. RECON
- Combined in RepeatModeller

Problems

- Speed
- False positive rate – esp. repetitive elements in genes
- Classification
- Saha *et al*. - Empirical comparison of ab initio repeat finding programs – Nucleic Acid Research 2008

References

Benson - Tandem repeats finder: a program to analyze DNA sequences – Nucleic Acid Research 1999

Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2010
http://www.repeatmasker.org

RepeatScout: Price AL, *et al*. - De novo identification of repeat families in large genomes. Bioinformatics. 2005

RECON: Bao Z, Eddy SR – Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002

RepeatModeller: http://www.repeatmasker.org/RepeatModeler.html

Saha *et al*. - Empirical comparison of ab initio repeat finding programs – Nucleic Acid Research 2008

Online resources to explore in your own time - I

Resource centres:
The EBI: www.ebi.ac.uk
The NCBI: www.ncbi.nlm.nih.gov
UCSC: genome.ucsc.edu

Model organism databases:
Budding yeast: www.yeastgenome.org
Worm: www.wormbase.org
Fly: www.flybase.org
Mouse: www.informatics.jax.org
Rat: rgd.mcw.edu
Zebrafish: zfin.org

Database building tool: www.intermine.org
InterMine databases:

http://yeastmine.yeastgenome.org
http://www.mousemine.org/mousemine
https://phytozome.jgi.doe.gov/phytomine/begin.do
https://apps.araport.org/thalemine
http://www.humanmine.org
http://targetmine.mizuguchilab.org
http://mitominer.mrc-mbu.cam.ac.uk
http://ratmine.mcw.edu
http://intermine.wormbase.org
http://zmine.zfin.org
http://intermine.modencode.org
http://www.flymine.org

Ontologies:
Gene ontology:
www.geneontology.org
Sequence ontology:
www.sequenceontology.org

DNA sequence/ genomes:
Ensembl: www.ensembl.org
Short read archive: www.ncbi.nlm.nih.gov/sra

Proteins:
www.uniprot.org
www.ebi.ac.uk/interpro

RNA:
rfam.sanger.ac.uk

Pathways:
www.reactome.org
www.genome.jp/kegg/

Online resources to explore in your own time - II

A selection of tools:

GBrowse genome browser: gmod.org/wiki/Gbrowse
CGL genome annotation manipulation:
www.yandell-lab.org/software/cgl.html
bio{perl, python, java, ruby}.org: much useful functionality
mummer overview genome comparisons: mummer.sourceforge.net

Galaxy: powerful online analysis system: galaxy.psu.edu

For next generation sequencing related questions/discussions:
http://seqanswers.com

Pubmed - http://www.ncbi.nlm.nih.gov/pubmed

WoK - http://wok.mimas.ac.uk/

Google - http://scholar.google.co.uk