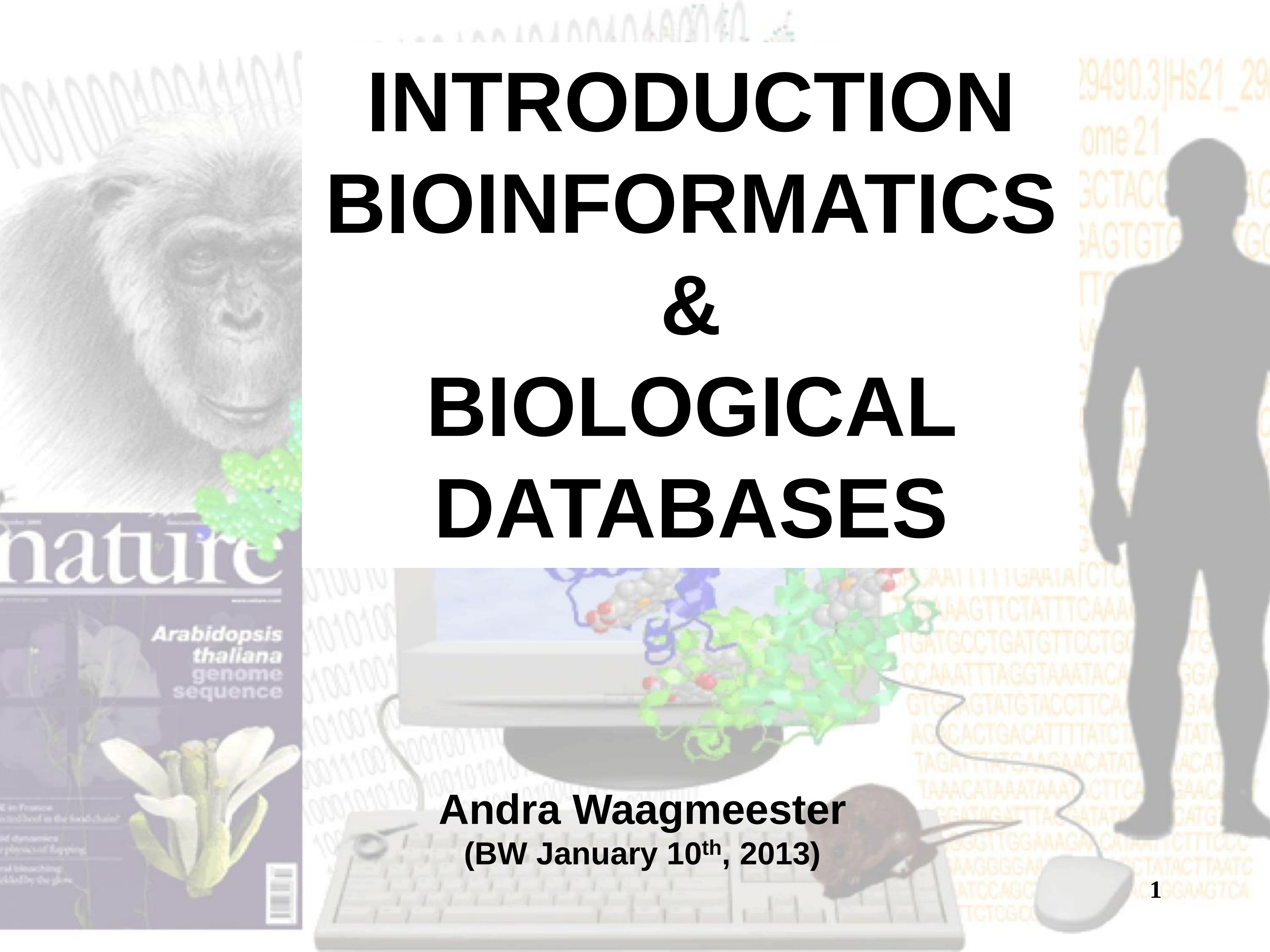


INTRODUCTION BIOINFORMATICS & BIOLOGICAL DATABASES

Andra Waagmeester

(BW January 10th, 2013)

BLOCKCOORDINATOR

- **Dr. Ir. Chris Evelo** **chris.evelo@maastrichtuniversity.nl**

ORGANISATION

- **Dr. Susan Coort (BGK)** **susan.coort@maastrichtuniversity.nl**
- **Dr. Lars Eijssen (MLW)** **l.eijssen@maastrichtuniversity.nl**
- **Andra Waagmeester (BW)andra.waagmeester@maastrichtuniversity**

OTHER INSTRUCTORS (left to right)

- **Dr. Egon Willighagen** **egon.willighagen@maastrichtuniversity.nl**
- **Stan Gaj, MSc** **s.gaj@maastrichtuniversity.nl**
- **Martina Kutmon, MSc** **martina.kutmon@maastrichtuniversity.nl**
- **Anwesha Dutta, MSc** **anwesha.dutta@maastrichtuniversity.nl**

Practical information

- Lectures will mostly be in **Dutch**
  Slides and practical exercises are in **English**

- All the practical sessions, should be signed off.

- You are required to study the literature provided for each practical session, <u>before</u> the session starts.
  - ➢ For the first session, the literature can be studied in the session

- The Bioinformatics trajectory (BMW2003) in year 2 is present in periods 1, 3, 4 and 5. Per block a bioinformatics exam will be given. This exam is separate from the block exam. In the end you will get **one grade** for the Bioinformatics trajectory

- **First exam Bioninformatics trajectory**
  - ➢ Tuesday April 4rd , 13.00-16:00
    <u>Open</u> book exam.

Course Material

ELEUM:

- The ***slides*** of the lecture will become available after the lecture.
- Per practical session ***literature references*** are provided.
- The ***exercises*** are available before the start of each practical session.
- The ***answers*** to the questions are available a week after each practical session.

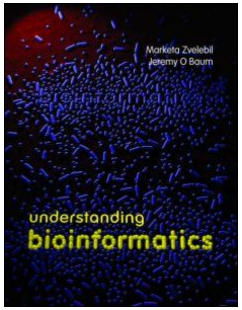
Course Material

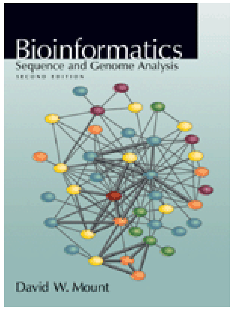
ELEUM:

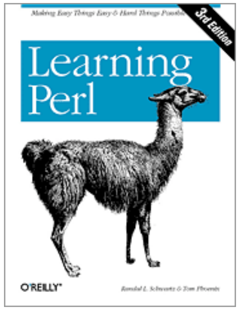
- The ***slides*** of the lecture will become available after the lecture.
- Per practical session ***literature references*** are provided. We strongly recommend you to read these before the session.
- The ***exercises*** are available before the start of each practical session.
- The ***answers*** to the questions are available a week after each practical session.

BOOKS (available at the “studielandschap”)

- **Understanding bioinformatics**
  M Zvelebil and J.O. Baum
- **Bioinformatics:**
  Sequence and genome analysis
  David W. Mount
- **Bioinformatics and Functional Genomics**
  J. Pevsner
- **Learning Perl**
  Randal L. Schwartz and Tom Phoenix







Subjects of bioinformatics track in BW2.3

1. Introduction Bioinformatics and Biological Databases

2. Protein Structures

Introduction Programming:

Matlab

2 sessions in

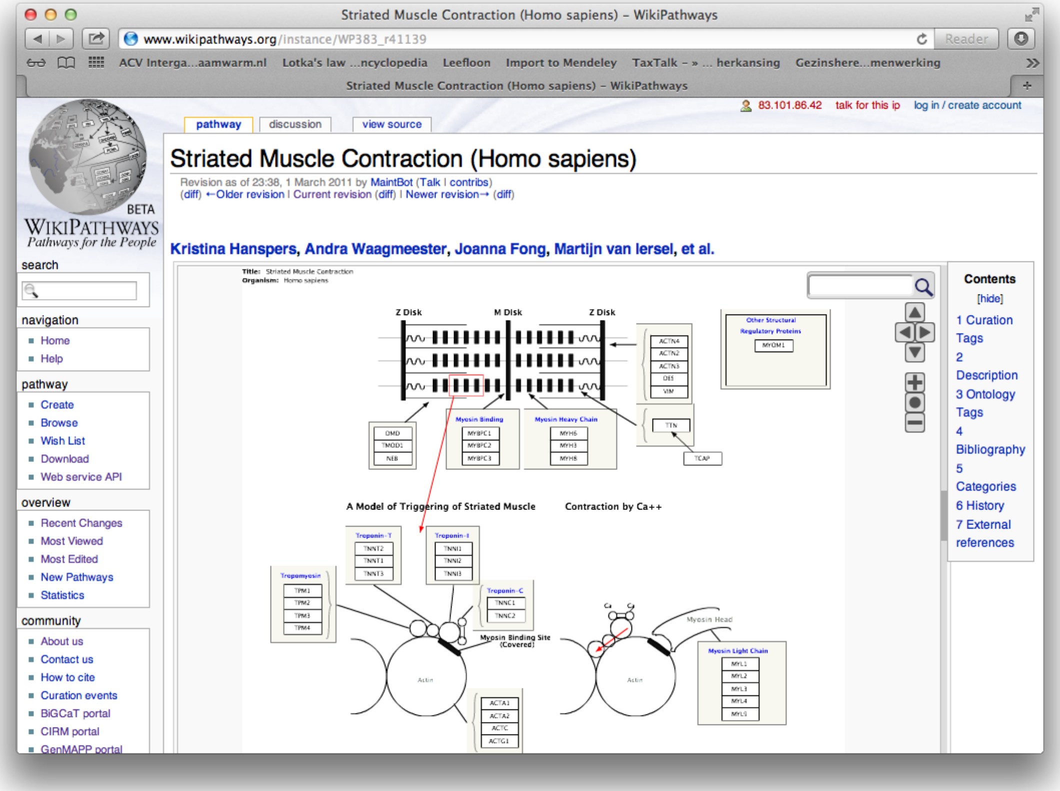
Next block 2.4 the programme will continue:

Gene Expression data

Pathway analysis – Network analysis

Introduction to Bioinformatics

Pathways



Striated Muscle Contraction (Homo sapiens) – WikiPathways

www.wikipathways.org/instance/WP383_r41139

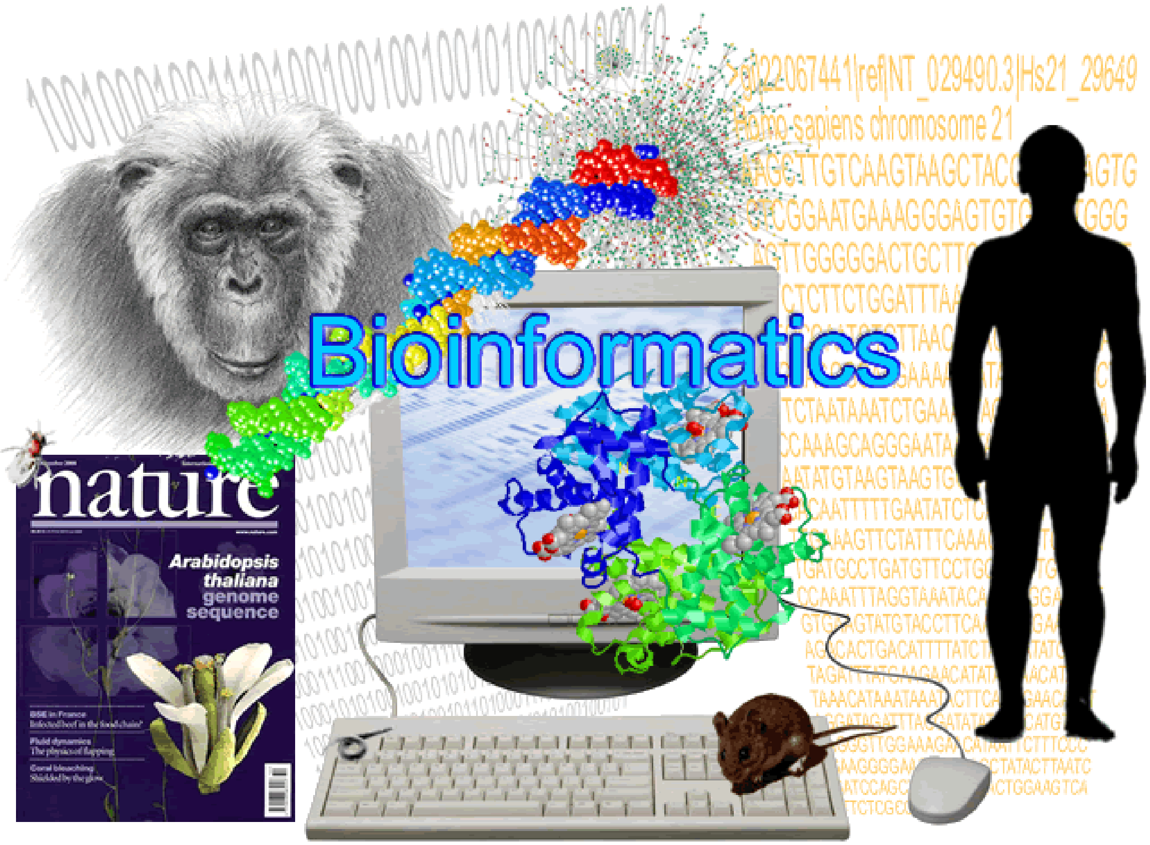
Striated Muscle Contraction (Homo sapiens)

Revision as of 23:38, 1 March 2011 by MaintBot (Talk | contribs)
(diff) ←Older revision | Current revision (diff) | Newer revision→ (diff)

Kristina Hanspers, Andra Waagmeester, Joanna Fong, Martijn van Iersel, et al.

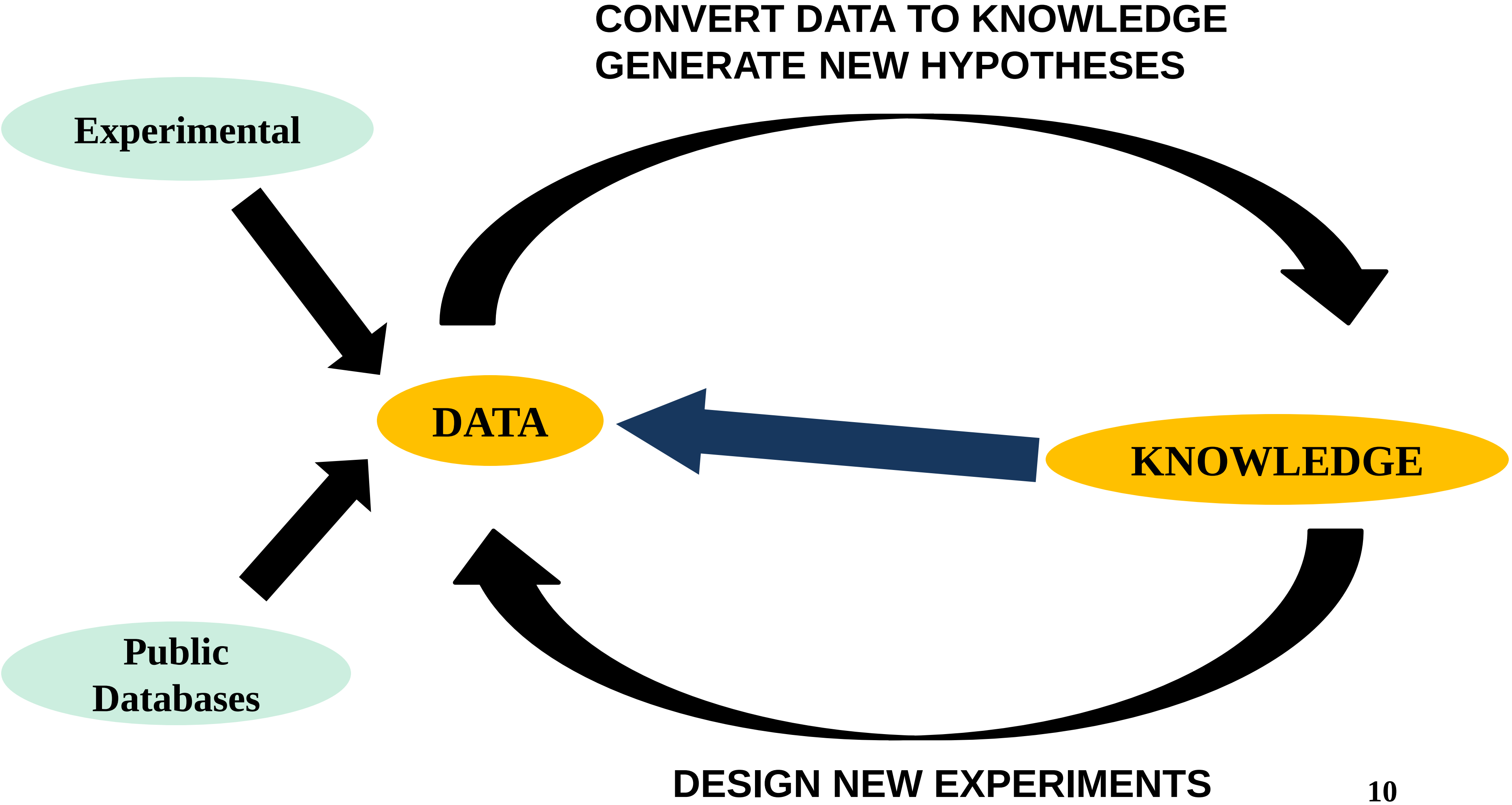
Title: Striated Muscle Contraction
Organism: Homo sapiens

What is Bioinformatics?


>gi|22067441|ref|NT_029490.3|Hs21_29649
Homo sapiens chromosome 21
Bioinformatics
nature
Arabidopsis thaliana genome sequence

Bioinformatics uses **“informatics” techniques** (from applied math, computer science, statistics, etc.) to **understand** and **organize** biological information, like genes, proteins and molecules on a **large-scale.**

Why Bioinformatics?



CONVERT DATA TO KNOWLEDGE
GENERATE NEW HYPOTHESES

Experimental

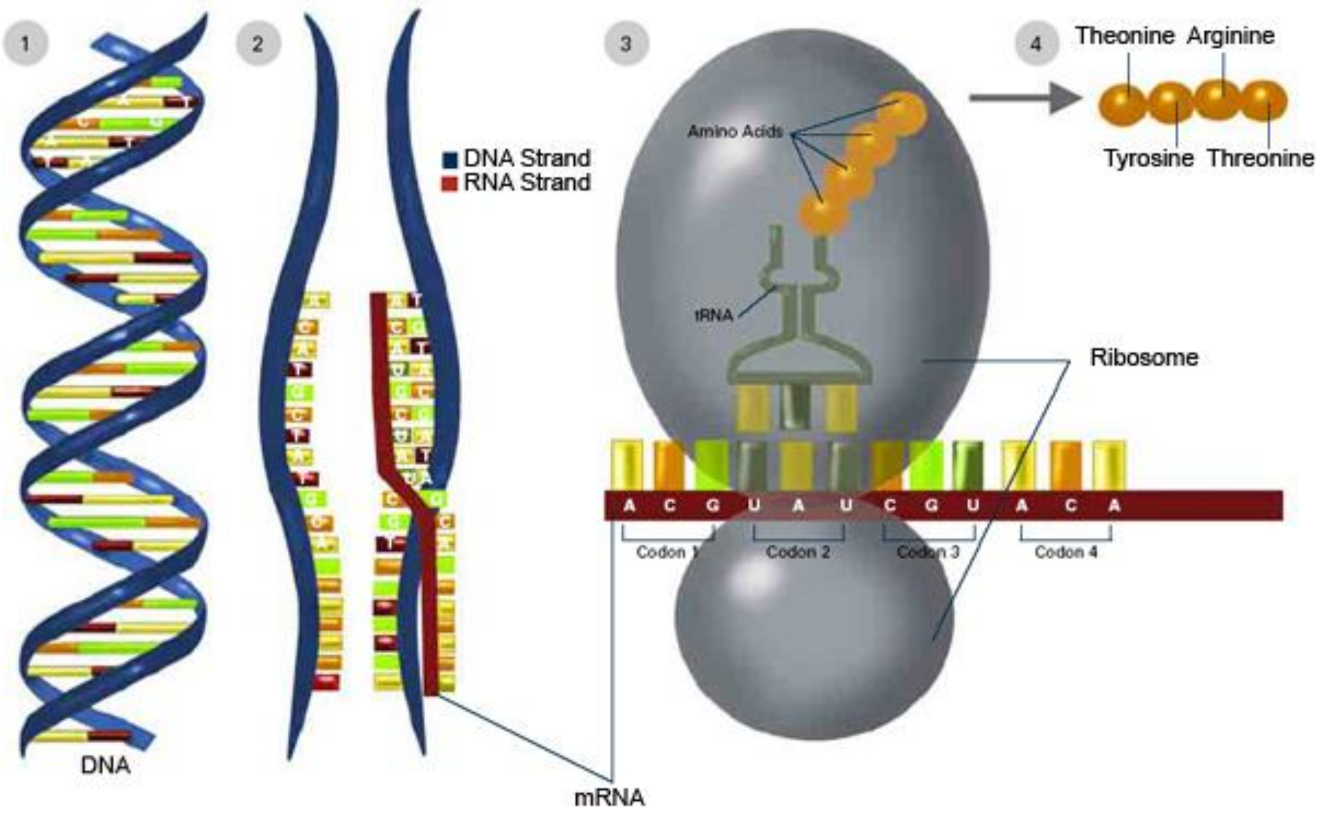
Public Databases

DATA

KNOWLEDGE

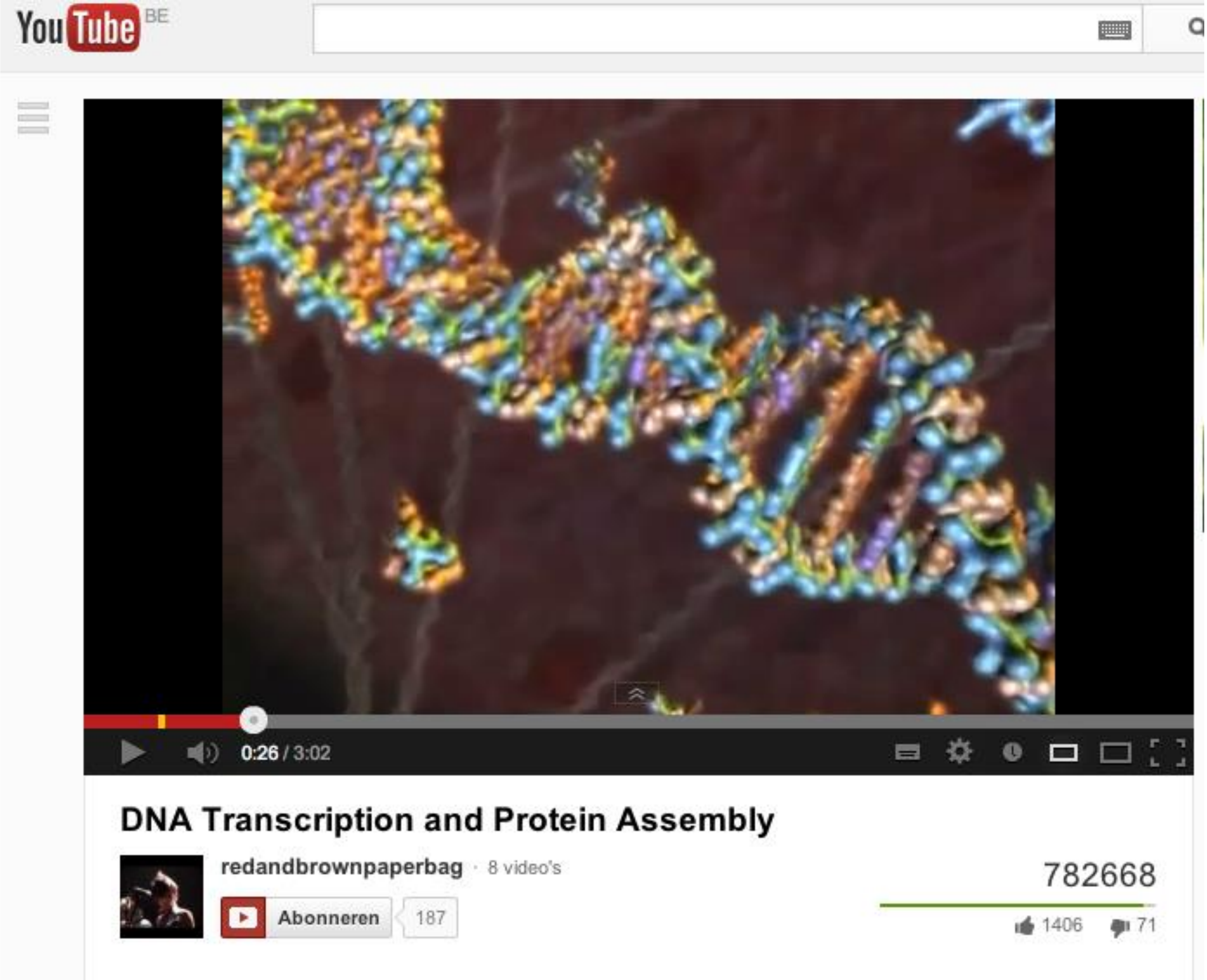
DESIGN NEW EXPERIMENTS

Central dogma of Molecular Biology



Gene (DNA) → **Transcription** → **mRNA** → **Translation** → **Protein**

Cells express **different** subset of the genes in different tissues and under different conditions

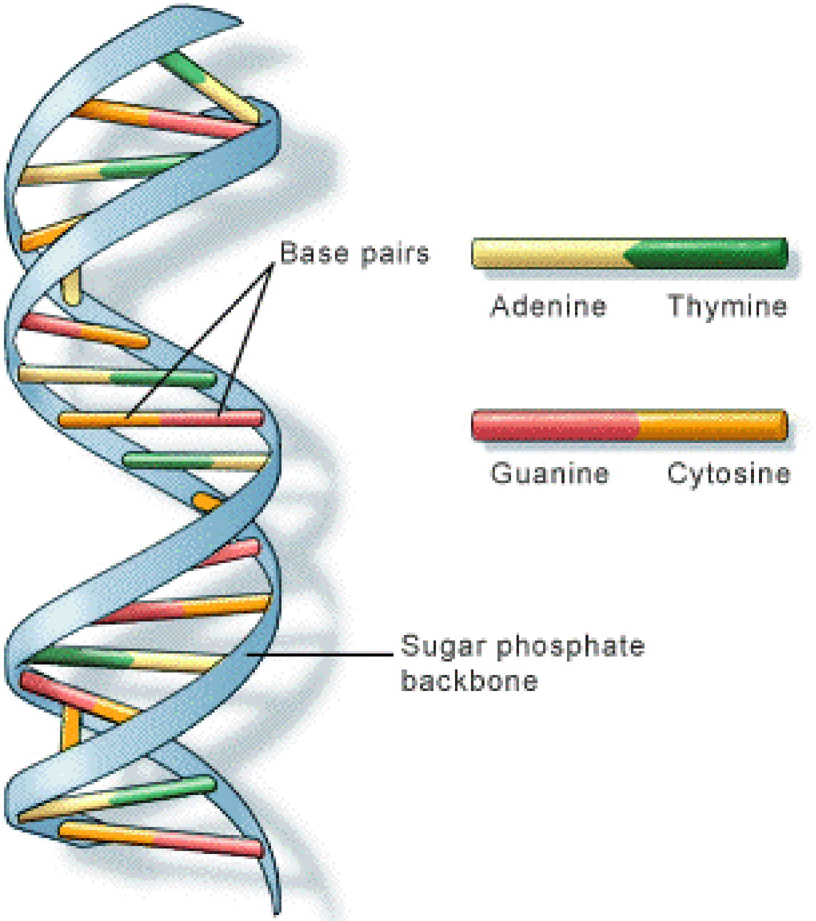


DNA Transcription and Protein Assembly

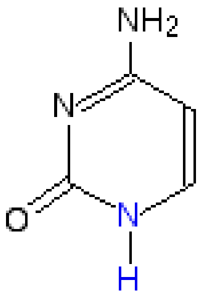
redandbrownpaperbag · 8 video's

782668

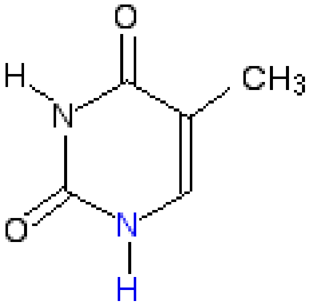
DNA



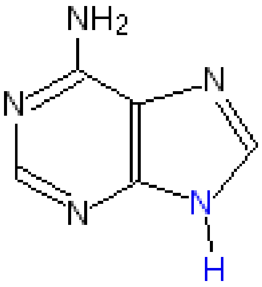
U.S. National Library of Medicine



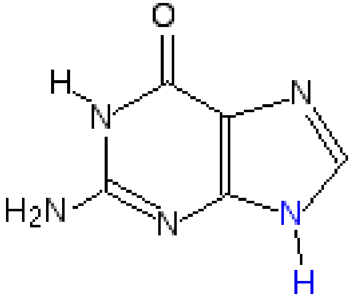
cytosine (C)



thymine (T)

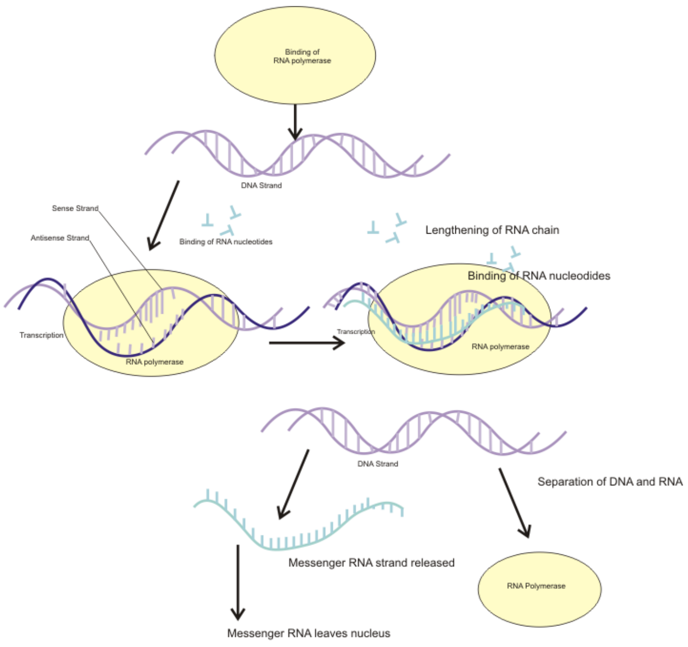


adenine (A)



guanine (G)

mRNA


Binding of RNA polymerase
DNA Strand
Sense Strand
Antisense Strand
Binding of RNA nucleotides
Lengthening of RNA chain
Binding of RNA nucleodides
Transcription
RNA polymerase
Transcription
RNA polymerase
DNA Strand
Separation of DNA and RNA
Messenger RNA strand released
RNA Polymerase
Messenger RNA leaves nucleus

Production of mRNA strand

Frank Boumphrey M.D. 2009

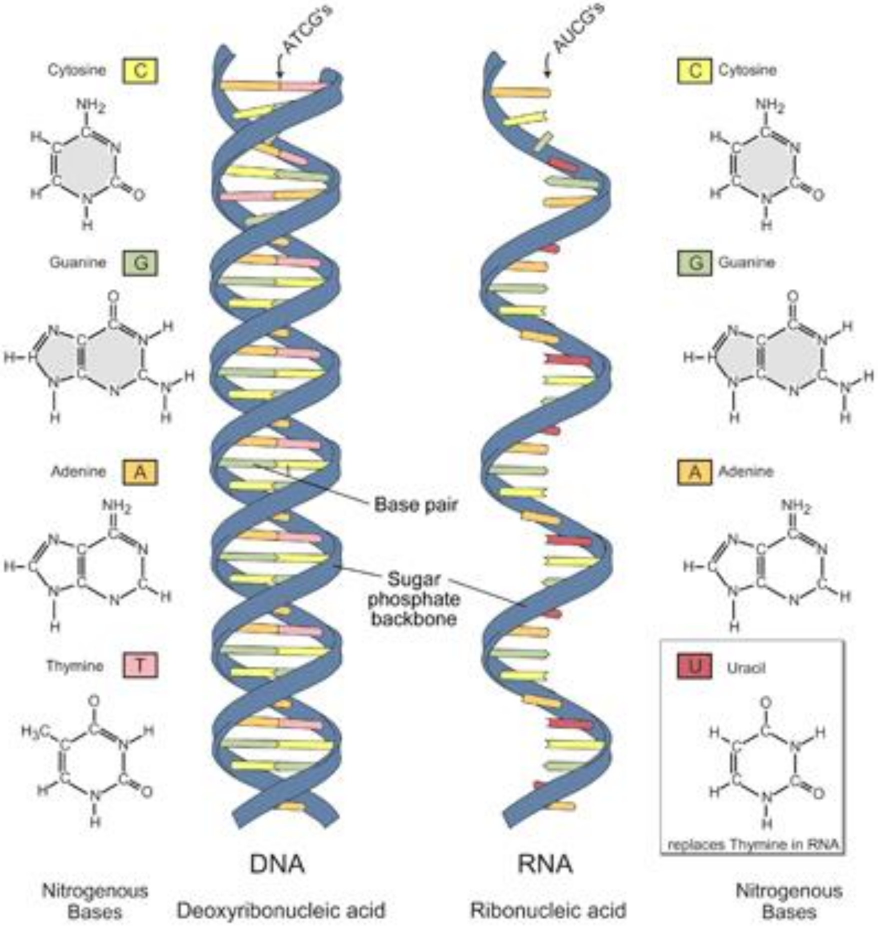
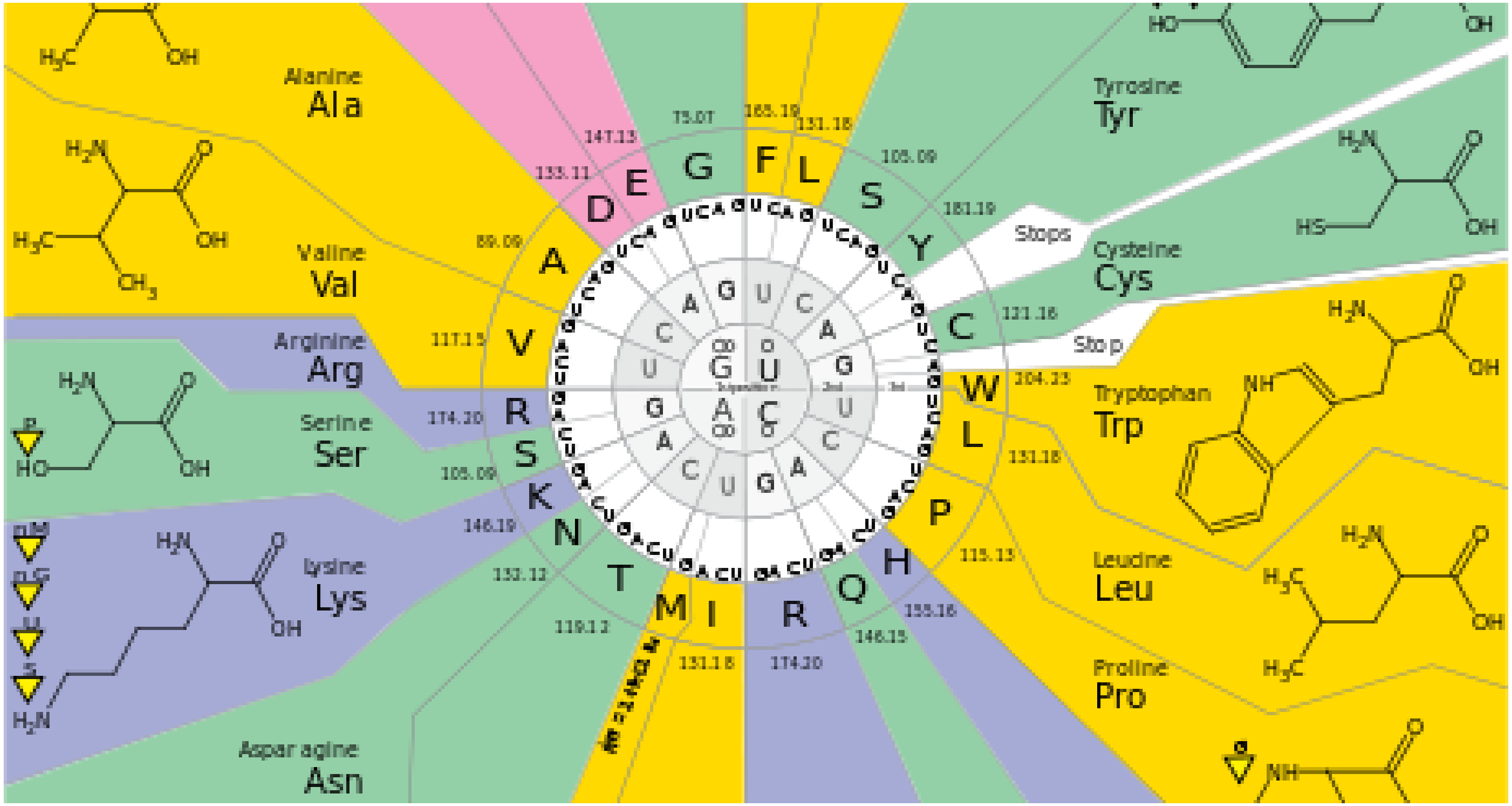
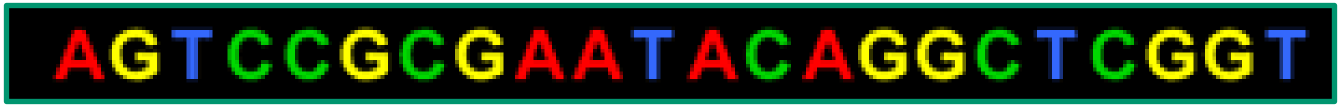


Image adapted from: National Human Genome Research Institute.

Proteins


Alanine
Ala
Valine
Val
Arginine
Arg
Serine
Ser
Lysine
Lys
Aspar agine
Asn
Tyrosine
Tyr
Stops
Cysteine
Cys
Stop
Tryptophan
Trp
Leucine
Leu
Proline
Pro
G
F
L
S
Y
C
W
L
P
H
Q
R
I
M
T
N
K
S
R
V
A
D
E
75.07
165.19
131.18
105.09
181.19
121.16
204.23
131.18
115.13
155.16
146.15
174.20
131.18
119.12
132.12
146.19
105.09
174.20
117.15
89.09
133.11
147.13

Genome Sequences / the Human Genome Project

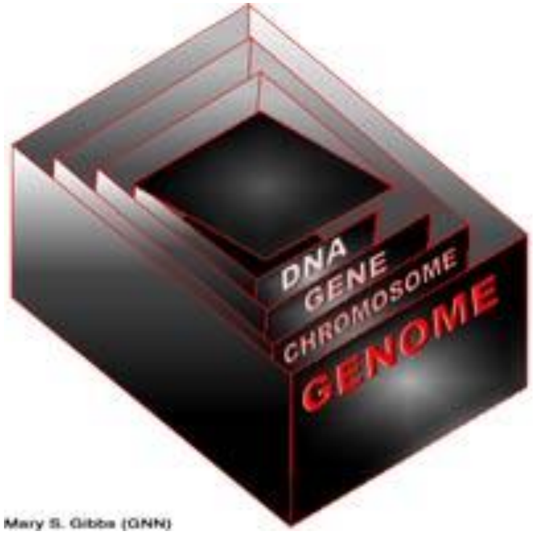


AGTCCGCGAATACAGGCTCGGT

Genomes

A genome is the collection of DNA that comprises an organism.

Today we have assembled the sequence of hundreds of genomes.

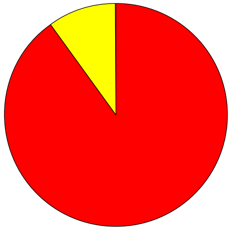
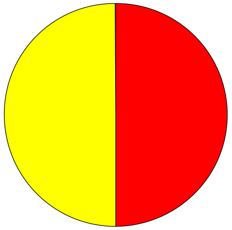
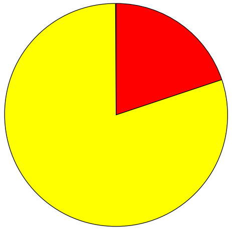
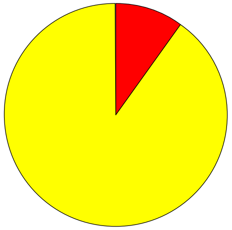
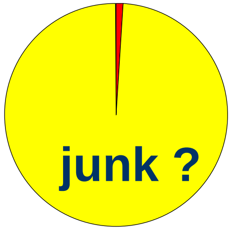


Mary S. Gibbs (GNN)

The genome is divided into chromosomes, chromosomes contain genes, and genes are made of DNA.

Each one of earth's organism has its own distinctive genome (except identical twins).

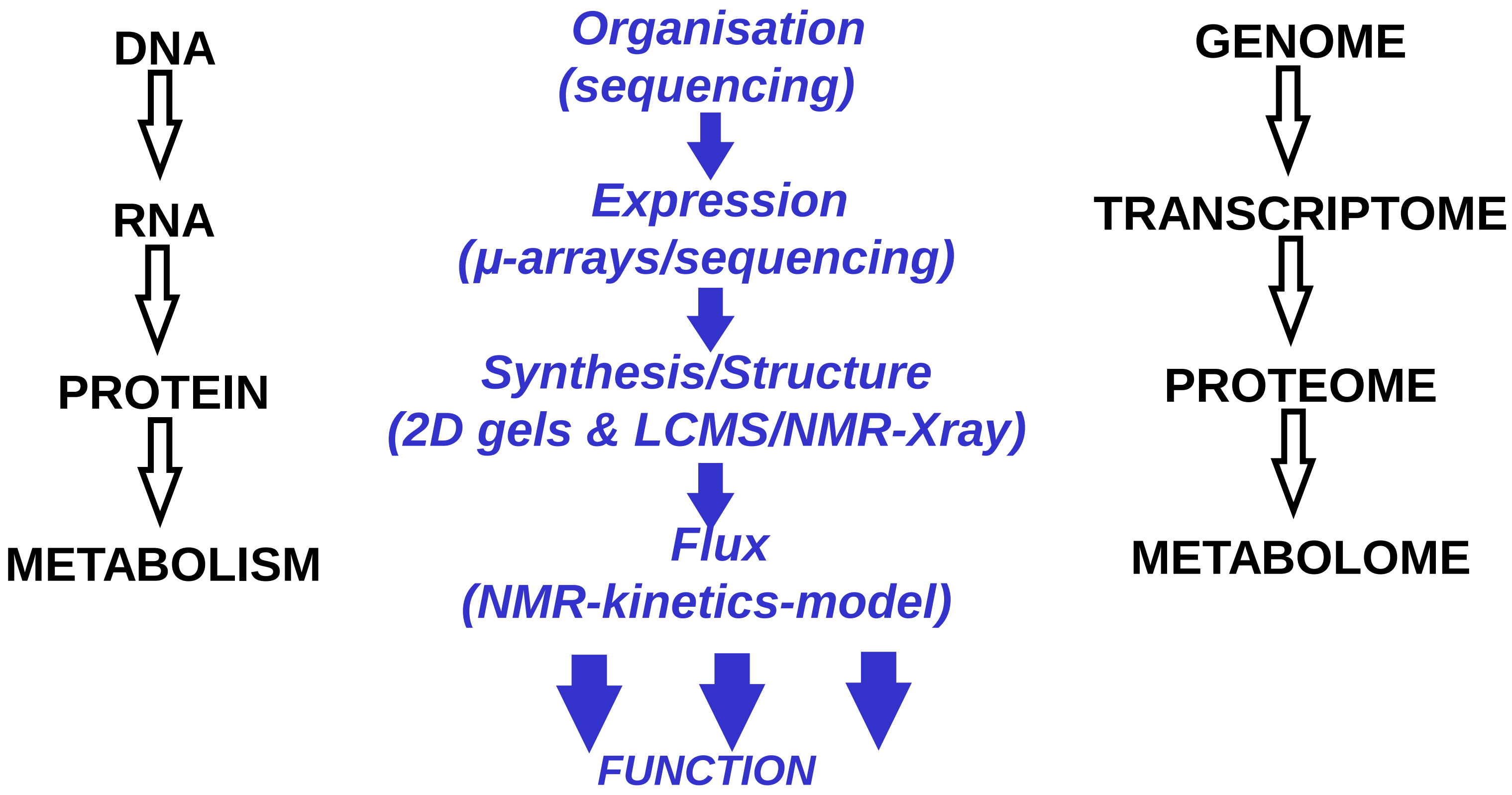
Genome content

	bacteria	yeast	worm	fly	man
Size (Mb)	2	12	97	137	3.500
% genes					
total genes	2.000	6.300	19.000	14.000	30.000 ?

Functional genomics

Single genes

All genes



DNA

RNA

PROTEIN

METABOLISM

Organisation
(sequencing)

Expression
(μ-arrays/sequencing)

Synthesis/Structure
(2D gels & LCMS/NMR-Xray)

Flux
(NMR-kinetics-model)

FUNCTION

GENOME

TRANSCRIPTOME

PROTEOME

METABOLOME

The vertebrate genomes available in Ensembl

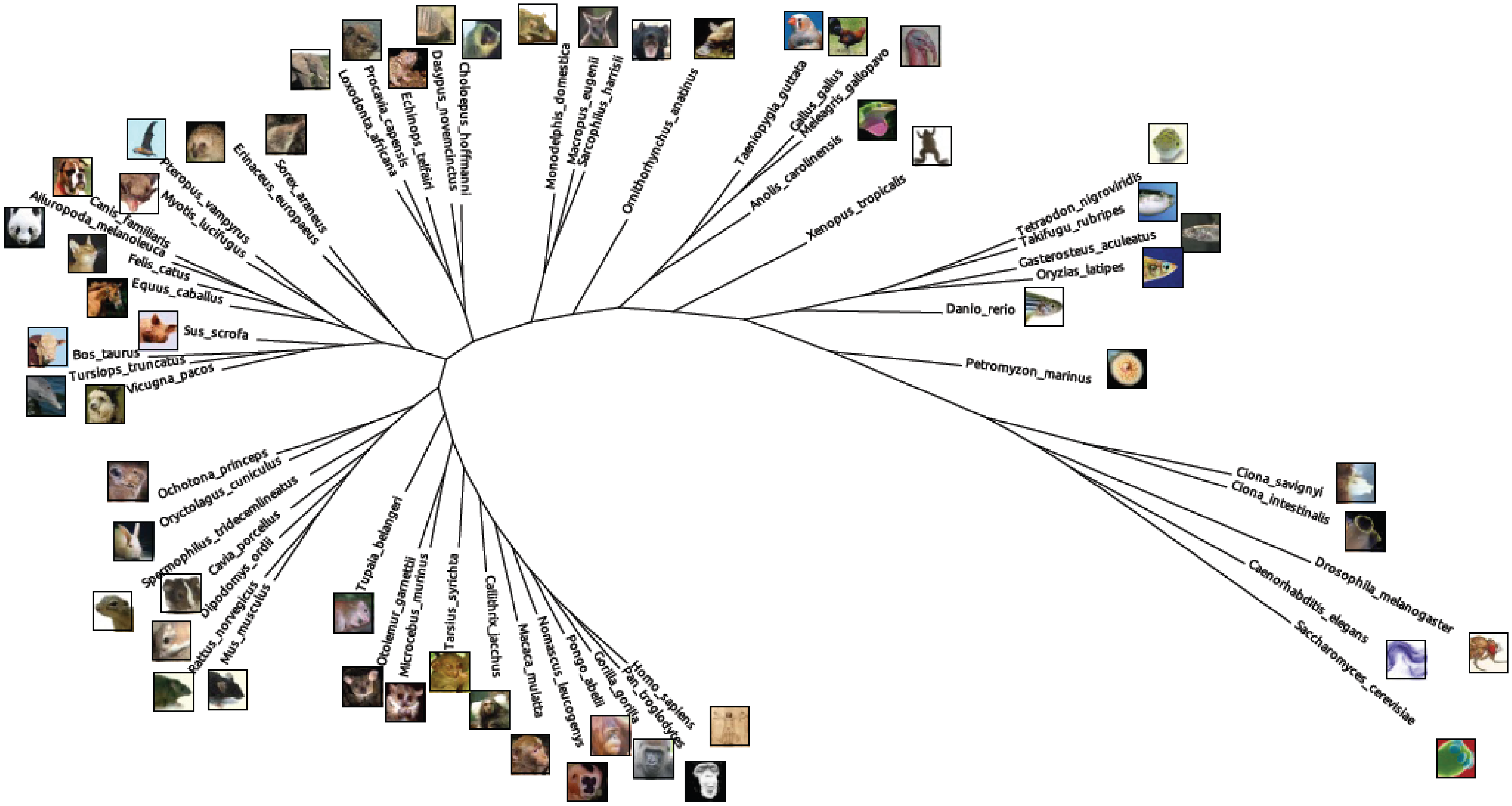


Image obtained using Dendroscope (D.H. Huson et al. "Dendroscope- An interactive viewer for large phylogenetic trees", BMC Bioinformatics 8:460, 2007)

Human Genome project

- Introduction video (http://www.youtube.com/watch?v=N4i6lYfYQzY&list=PLF0701633C91835BF&index=1)
- Strategies
- Conclusions

International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature 431, 931-945 (21 October 2004).

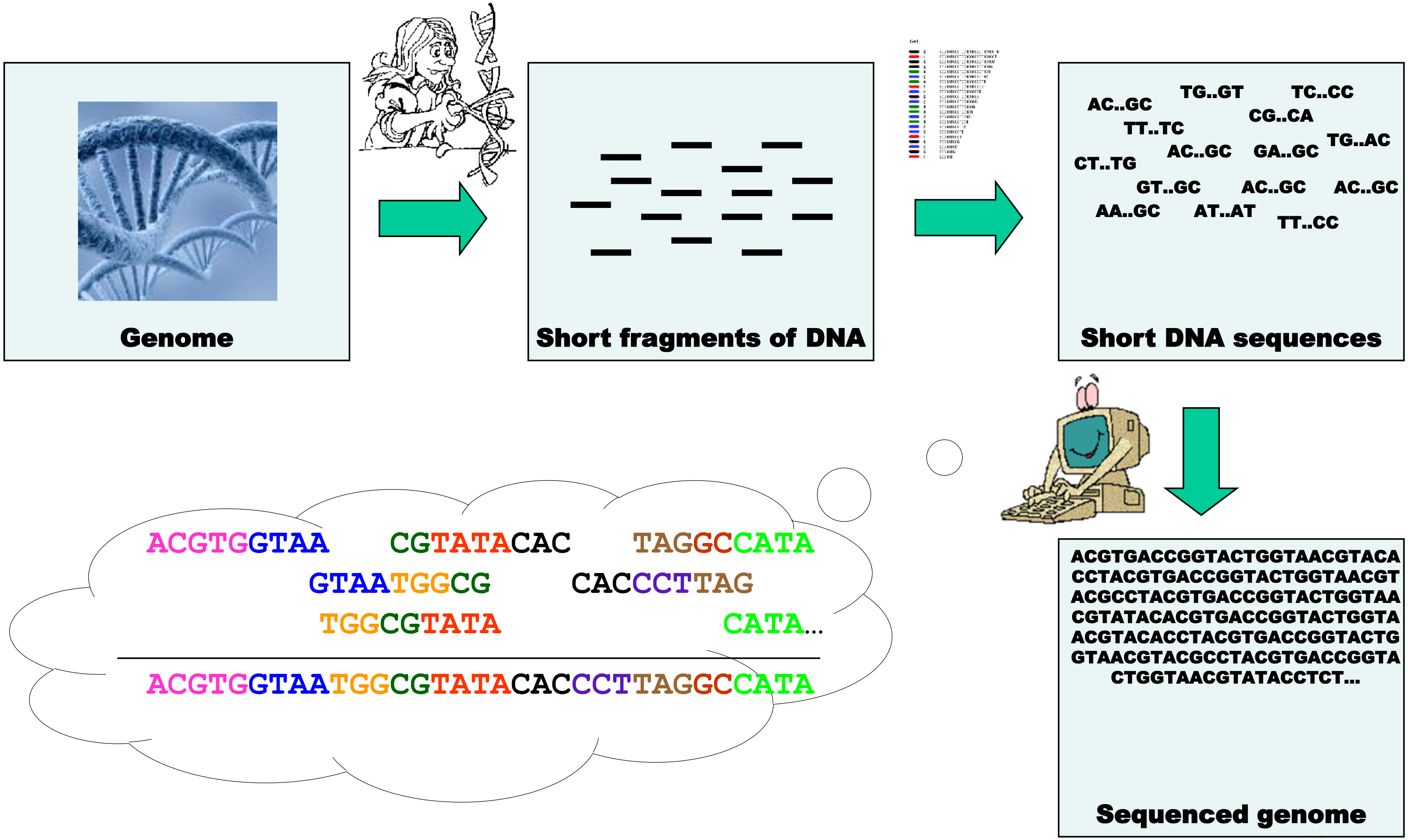
Overview of genome analysis

There are two main strategies for sequencing genomes

An approach used to decode an organism's genome by shredding it into smaller fragments of DNA which can be sequenced individually. The sequences of these fragments are then ordered, based on overlaps in the genetic code, and finally reassembled into the complete sequence.

The 'whole genome shotgun' (WGS) method is applied to the entire genome all at once, while the 'hierarchical shotgun' method is applied to large, overlapping DNA fragments of known location in the genome.

Genome sequencing



Genome

Short fragments of DNA

AC..GC TG..GT TC..CC
TT..TC CG..CA
CT..TG AC..GC GA..GC TG..AC
GT..GC AC..GC AC..GC
AA..GC AT..AT TT..CC

Short DNA sequences

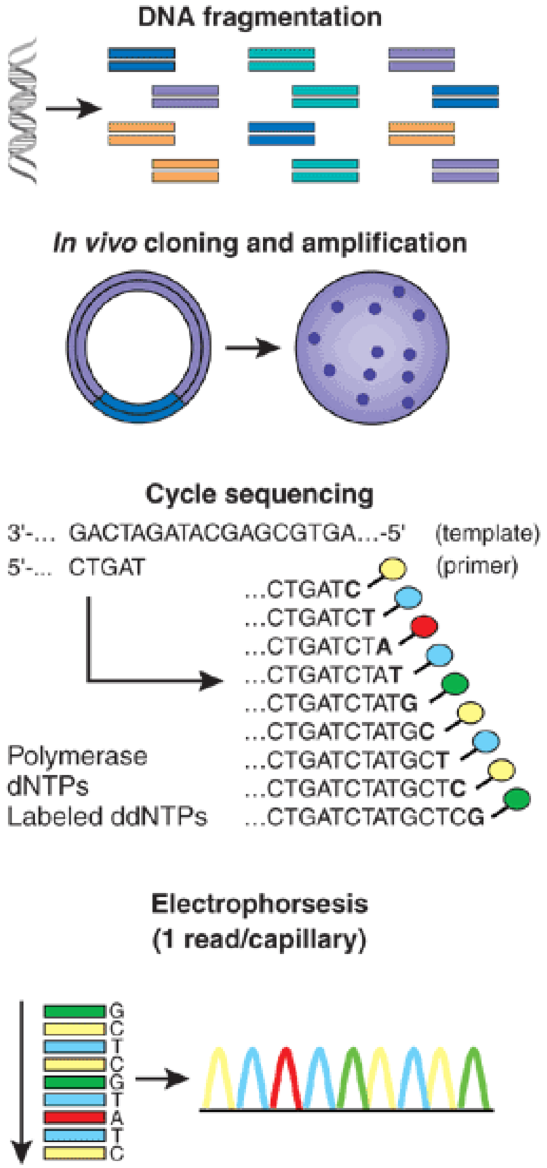
ACGTGGTAA CGTATACAC TAGGCCATA
GTAATGGCG CACCCTTAG
TGGCGTATA CATA...

ACGTGGTAATGGCGTATACACCCTTAGGCCATA

ACGTGACCGGTACTGGTAACGTACA
CCTACGTGACCGGTACTGGTAACGT
ACGCCTACGTGACCGGTACTGGTAA
CGTATACACGTGACCGGTACTGGTA
ACGTACACCTACGTGACCGGTACTG
GTAACGTACGCCTACGTGACCGGTA
CTGGTAACGTATACCTCT...

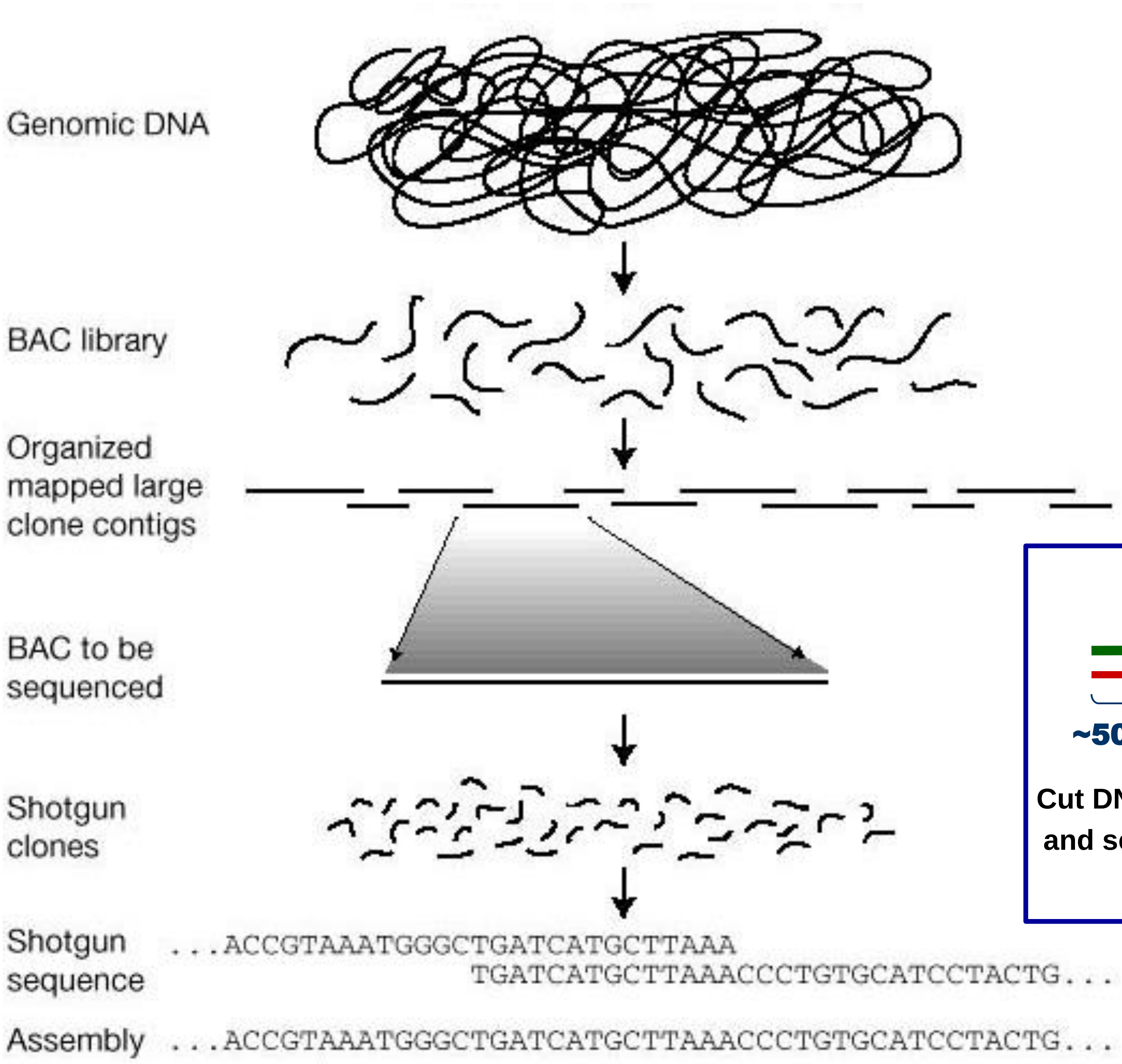
Sequenced genome

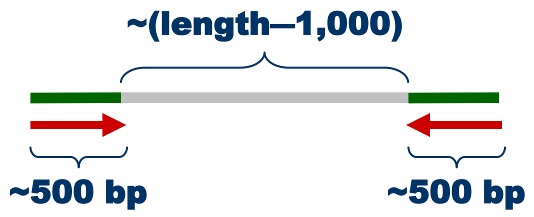
Workflow of Sanger sequencing



Jay Shendure & Hanlee Ji, Nature Biotechnology 26, 1135 - 1145 (2008)

Hierarchical shotgun sequencing

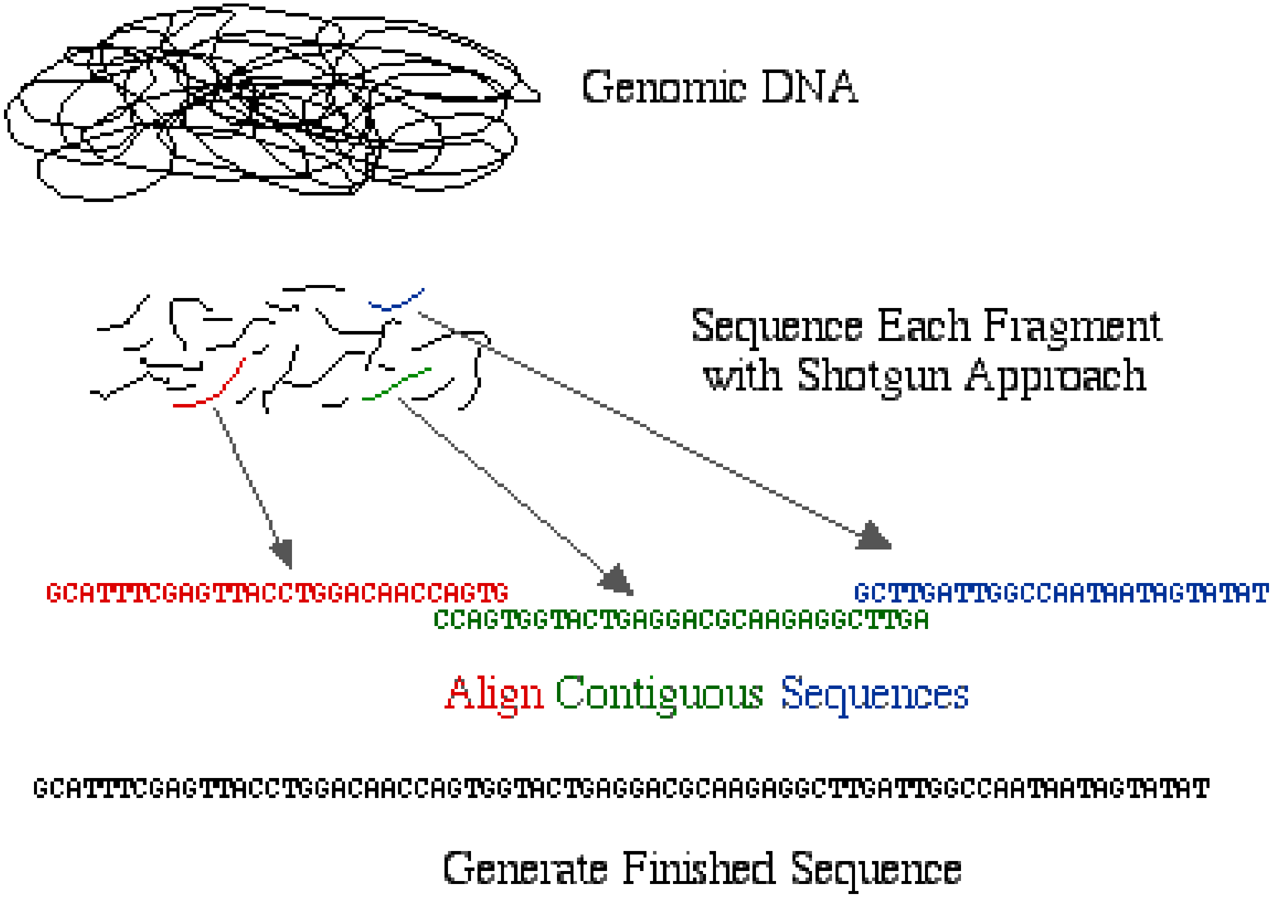

Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...



Cut DNA to larger pieces (2Kbp, 15Kbp) and sequence both ends of each piece (Fleischmann et al., 1994)

Source: IHGSC (2001)

Whole genome shotgun sequencing


Genomic DNA
Sequence Each Fragment
with Shotgun Approach
GCATTTCGAGTTACCTGGACAACCAGTG
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
GCTTGATTGGCCAATAATAGTATAT
Align Contiguous Sequences
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
Generate Finished Sequence

Source: IHGSC (2001)

When has a genome been fully sequenced?

A typical goal is to obtain five to ten-fold coverage.

Finished sequence: a clone insert is contiguously sequenced with high quality standard of error rate 0.01%. There are usually no gaps in the sequence.

Draft sequence: clone sequences may contain several regions separated by gaps. The true order and orientation of the pieces may not be known.

Main conclusions of human genome project (1)

1. We have about the same number of genes as fish and plants, and not that many more genes than worms and flies – 20,000-30,000

2. The human proteome is far more complex than the set of proteins encoded by invertebrate genomes – alternative splicing

3. Hundreds of human genes were acquired from bacteria by lateral gene transfer

4. 98% of the genome does not code for genes and >50% of the genome consists of repetitive DNA

5. Segmental duplication is a frequent occurrence in the human genome

Main conclusions of human genome project (2)

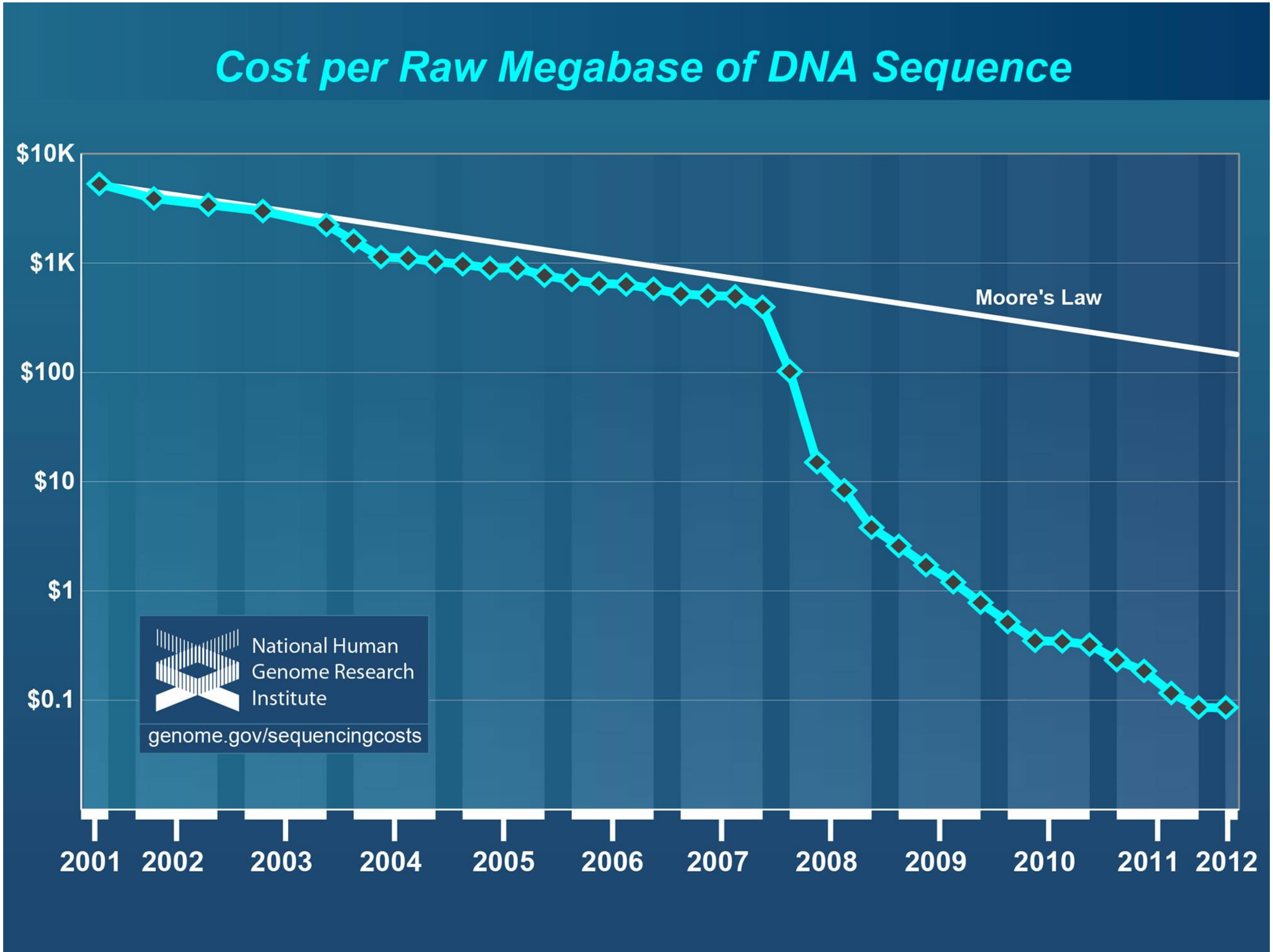
6. There are 300,000 Alu repeats in the human genome
   - These are about 300 base pairs and contain an AluI restriction enzyme site.
   - They occupy 3% of the genome and may confer some benefit

7. The mutation rate is about twice as high in male meiosis than female meiosis; most mutation probably occurs in males

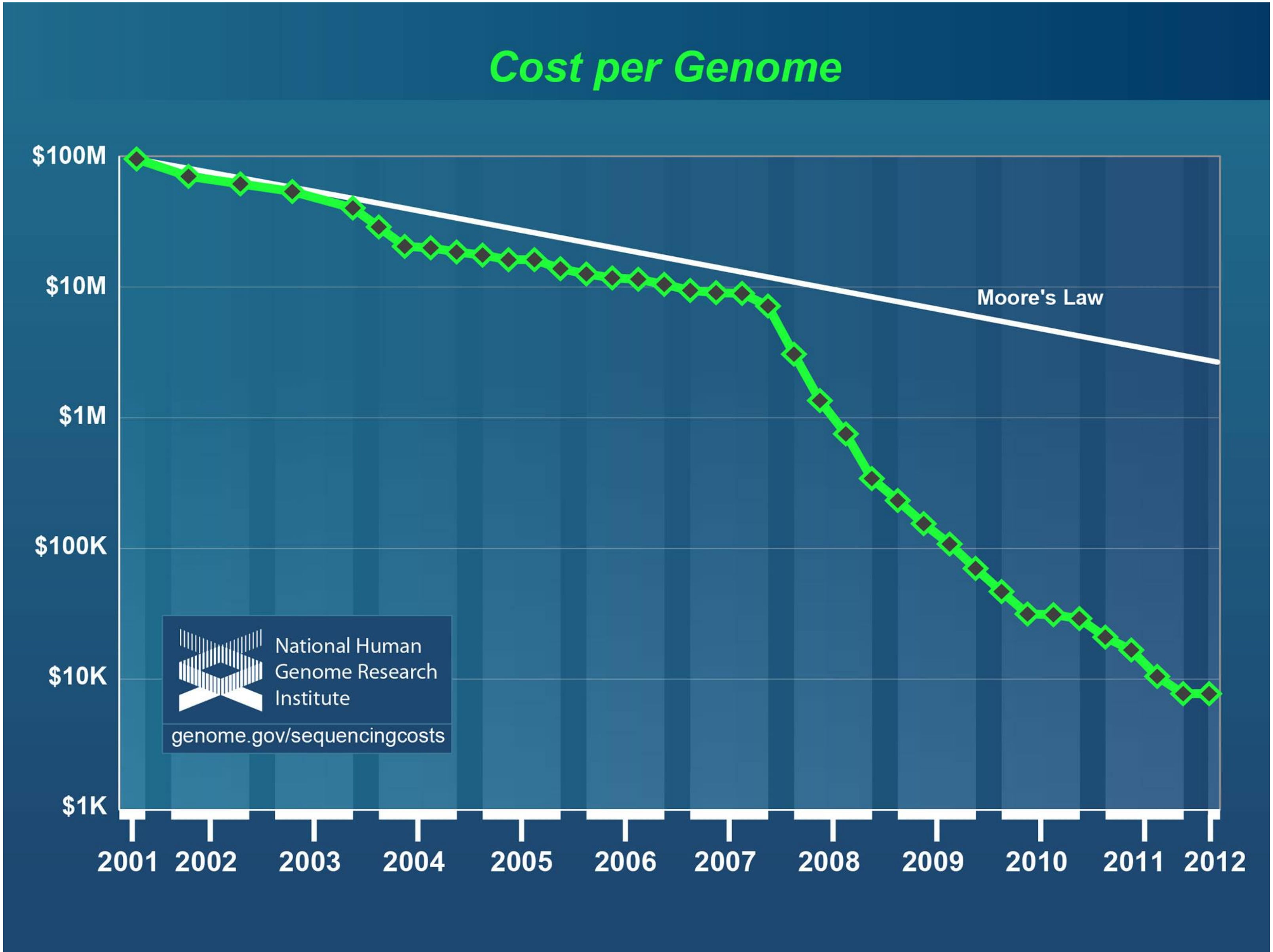
8. 1.5 – 2 million single base pair changes or single nucleotide polymorphisms (SNPs) were originally identified.
   - Currently, dbSNP at NCBI over 10 million human SNPs
   - Half of these have been validated
   - A SNP occurs every 100 to 300 base pairs
   - Fewer than 1% of SNPs alter protein sequence

9. Noncoding RNAs are also important (for example miRNAs)

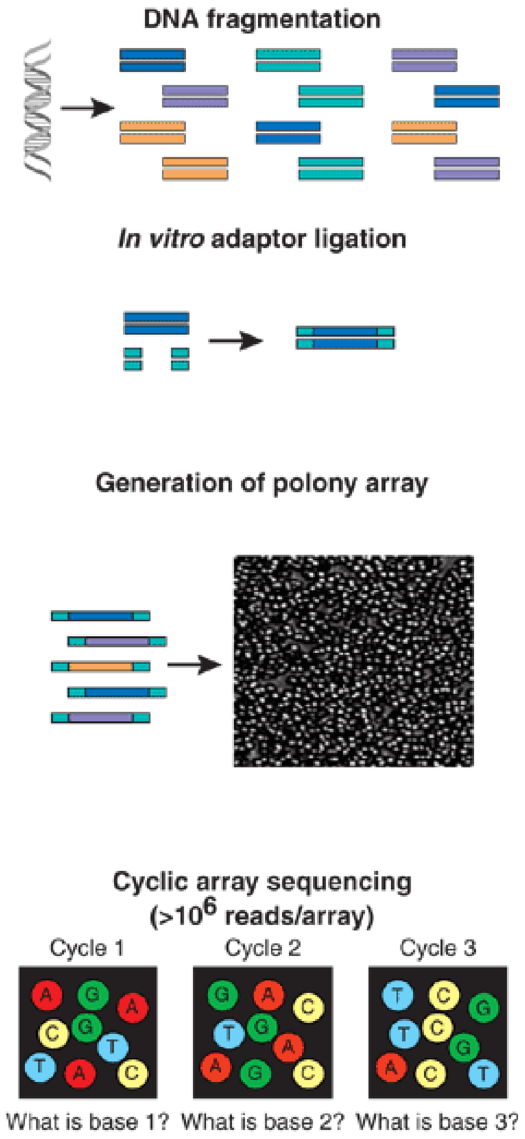
Sequencing the Human Genome



Sequencing the Human Genome



Nowadays: Work flow second-generation sequencing



Jay Shendure & Hanlee Ji, Nature Biotechnology 26, 1135 - 1145 (2008)

Gateways to the genome sequences

Genome browsers: gateways to the genome sequences

- Over the last few decades a gigantic amount of information on DNA sequences, gene locations, gene transcripts, protein functions and so on has been gathered

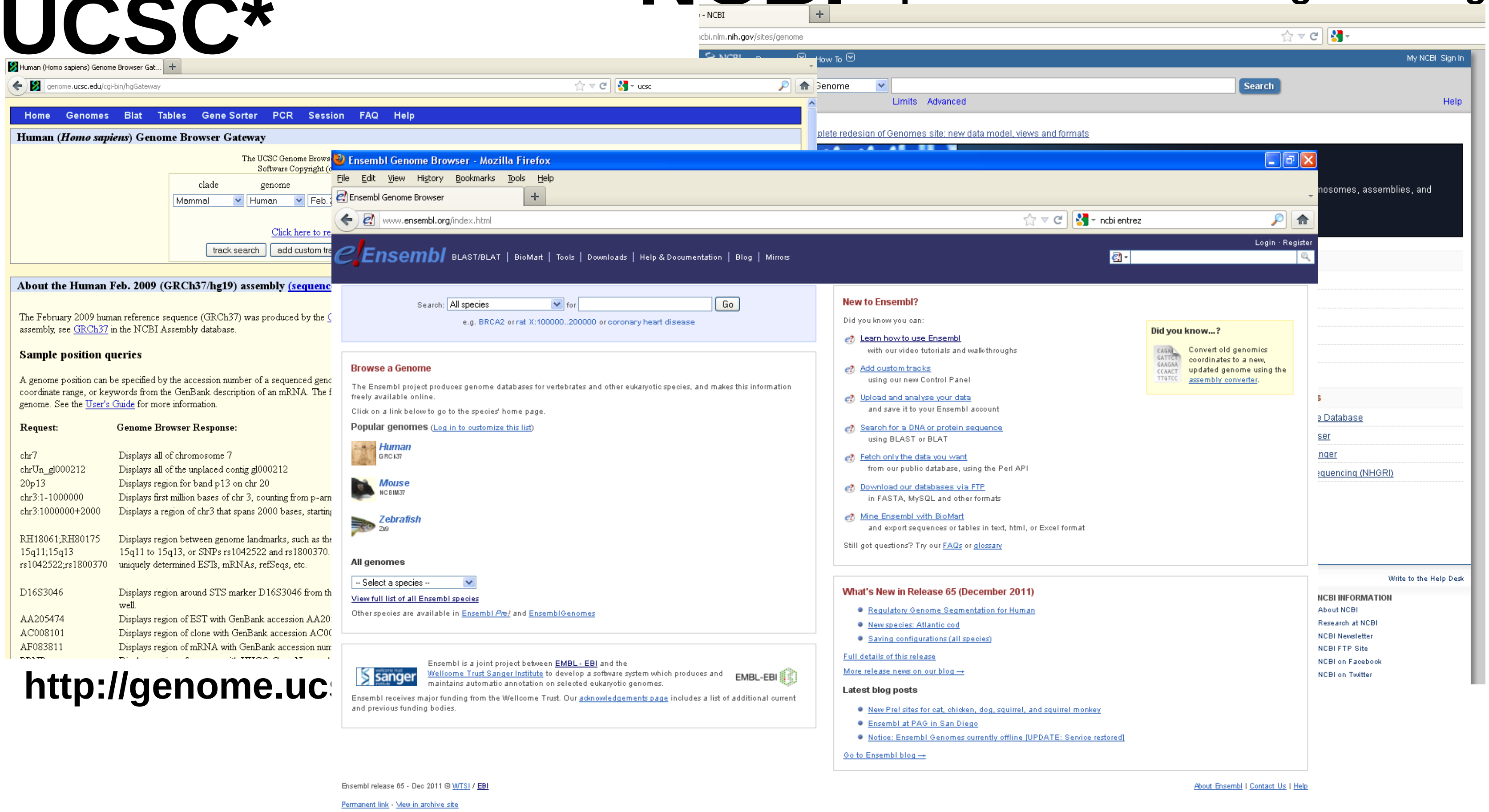
- Now we will discuss several websites that provide all this information collection, and that you will use in the afternoon session

- They all contain essentially the same information, but have a differente interface, look-and-feel, viewing options

Genome Browsers

UCSC*

NCBI **http://www.ncbi.nlm.nih.gov/sites/genome**



Human (Homo sapiens) Genome Browser Gat...

genome.ucsc.edu/cgi-bin/hgGateway

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Brows
Software Copyright (

clade: Mammal — genome: Human — Feb.

Click here to re

track search — add custom tra

About the Human Feb. 2009 (GRCh37/hg19) assembly (sequenc

The February 2009 human reference sequence (GRCh37) was produced by the G
assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genc
coordinate range, or keywords from the GenBank description of an mRNA. The f
genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting
RH18061;RH80175	Displays region between genome landmarks, such as the
15q11;15q13	15q11 to 15q13, or SNPs rs1042522 and rs1800370.
rs1042522;rs1800370	uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046	Displays region around STS marker D16S3046 from th well.
AA205474	Displays region of EST with GenBank accession AA20
AC008101	Displays region of clone with GenBank accession AC00
AF083811	Displays region of mRNA with GenBank accession num

- NCBI

ncbi.nlm.nih.gov/sites/genome

How To — My NCBI Sign In

Genome — Search

Limits Advanced — Help

plete redesign of Genomes site: new data model, views and formats

nosomes, assemblies, and

e Database

ser

nger

equencing (NHGRI)

Write to the Help Desk

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI Newsletter
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter

Ensembl Genome Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Ensembl Genome Browser

www.ensembl.org/index.html — ncbi entrez

Login · Register

Ensembl BLAST/BLAST | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: All species for — Go
e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

Human GRCh37
Mouse NCBIM37
Zebrafish Zv9

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl Pre! and EnsemblGenomes

Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

New to Ensembl?

Did you know you can:

- Learn how to use Ensembl with our video tutorials and walk-throughs
- Add custom tracks using our new Control Panel
- Upload and analyse your data and save it to your Ensembl account
- Search for a DNA or protein sequence using BLAST or BLAT
- Fetch only the data you want from our public database, using the Perl API
- Download our databases via FTP in FASTA, MySQL and other formats
- Mine Ensembl with BioMart and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs or glossary

Did you know...?

Convert old genomics coordinates to a new, updated genome using the assembly converter.

What's New in Release 65 (December 2011)

- Regulatory Genome Segmentation for Human
- New species: Atlantic cod
- Saving configurations (all species)

Full details of this release

More release news on our blog →

Latest blog posts

- New Pre! sites for cat, chicken, dog, squirrel, and squirrel monkey
- Ensembl at PAG in San Diego
- Notice: Ensembl Genomes currently offline [UPDATE: Service restored]

Go to Ensembl blog →

Ensembl release 65 - Dec 2011 © WTSI / EBI — About Ensembl | Contact Us | Help

Permanent link - View in archive site

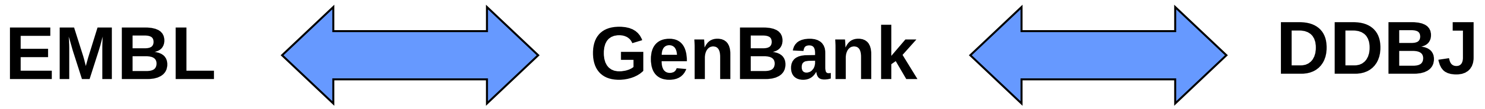
http://genome.uc

Ensembl **http://www.ensembl.org/**

*** We will use the UCSC browser later during the course**

Nucleotide databases

The underlying raw DNA sequences are identical



EMBL

Housed at EBI

European Bioinformatics Institute

www.ebi.ac.uk/embl/

GenBank

Housed at NCBI

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov/Genbank/

DDBJ

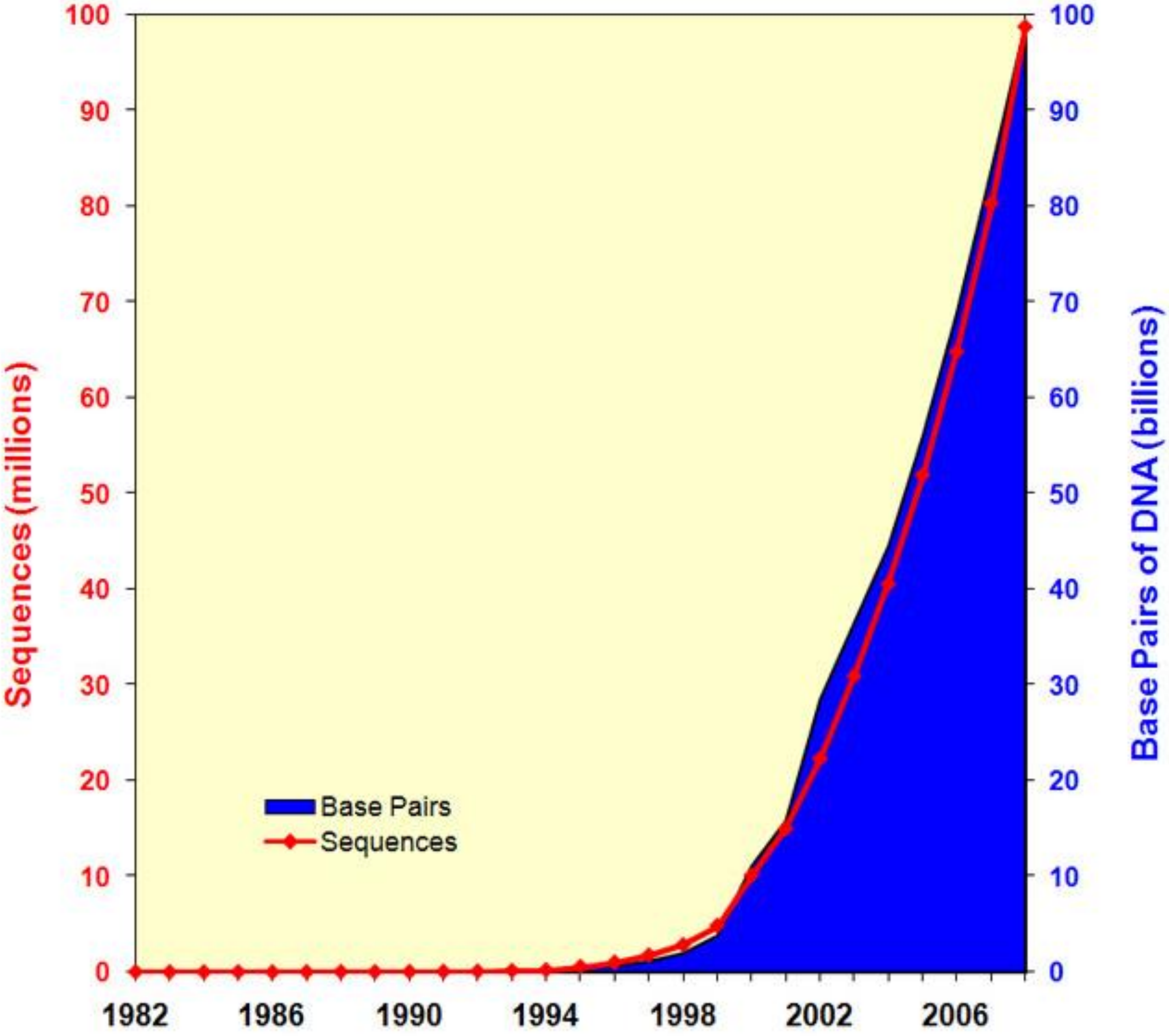
Housed in Japan

DNA Data Bank of Japan

www.ddbj.nig.ac.jp/

Hundreds of thousands of species are represented

Growth of GenBank (1982-2008)



NCBI nucleotide databases

- GenBank
  - Individual submissions
  - Bulk submissions (Genome centers)
    - High throughput sequencing (DNA)
    - Expressed Sequence Tags (mRNA)

- RefSeq
  - Curated subset of GenBank
  - “Reference” sequence
  - Single sequence per locus / molecule

Protein sequence databases

- NCBI
  - **RefSeq** and **Protein**
- EBI
  - **Swiss-Prot** and **TrEMBL** → **UniProt**
- **Translated** from nucleotide sequence
- **Curated**
- **Combined**

Accession numbers (Identifiers)

Label to unambiguously identify a sequence

Examples (all for retinol-binding protein, RBP4):

DNA	X02775	GenBank genomic DNA sequence
	NT_030059	Genomic contig
	Rs7079946	dbSNP (single nucleotide polymorphism)
RNA	N91759.1	An expressed sequence tag (1 of 170)
	NM_006744	RefSeq DNA sequence (from a transcript)
protein	NP_007635	RefSeq protein
	AAC02945	GenBank protein
	Q28369	UniProt protein
	1KT7	Protein Data Bank structure record

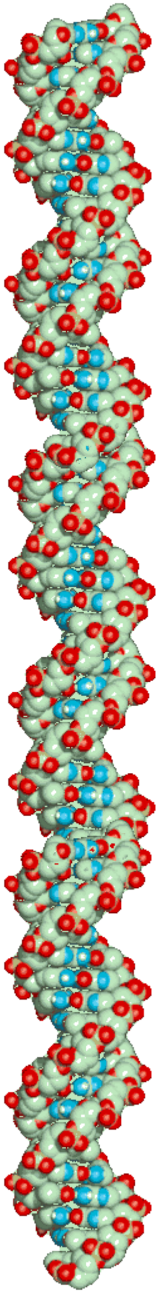
From Sequence to Genes: where are the genes?

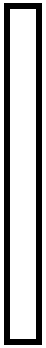
- Gene prediction
  - Extrinsic
    - Search for genes based on observed mRNA / Protein sequences
    - UniGene
  - Ab initio
    - Predict genes based on genomic sequence alone
    - Promoter sequence
    - Poly(A) tail binding sites, CG content (higher in genes), splicing sites

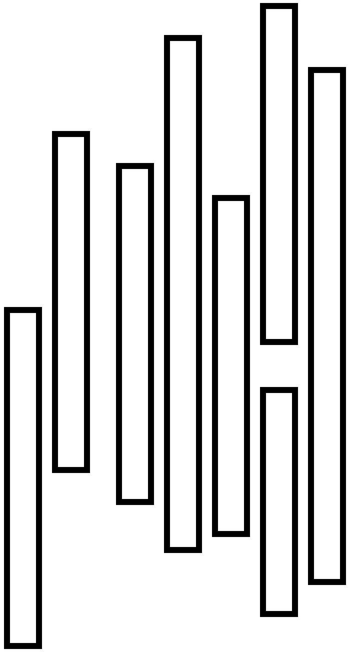
UniGene

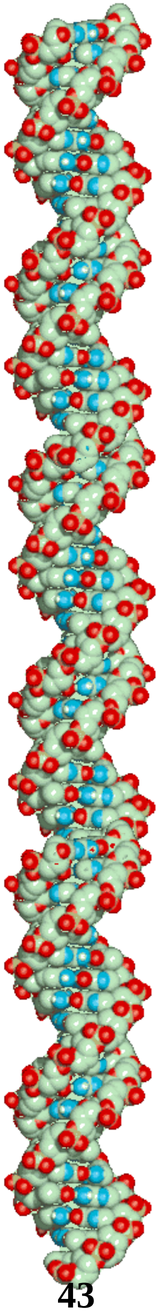
- Predict genes based on ESTs
- EST:
  - DNA sequence corresponding to mRNA from expressed gene
  - ~500 base pairs long
  - Sequenced from a cDNA library
- Cluster ESTs from many cDNA libraries to predict distinct genes

EST clusters





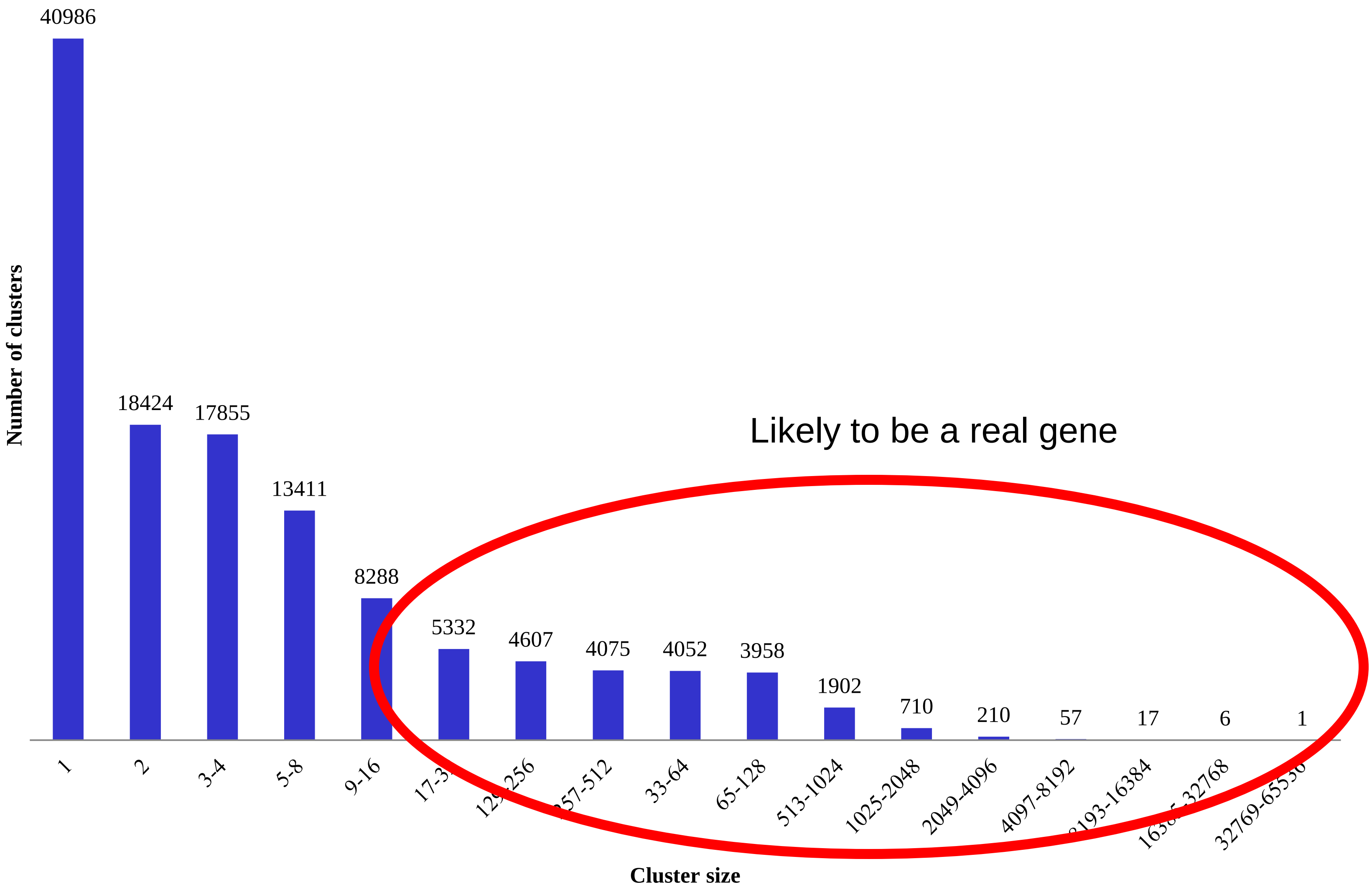




This is a gene with 1 EST associated; the cluster size is 1

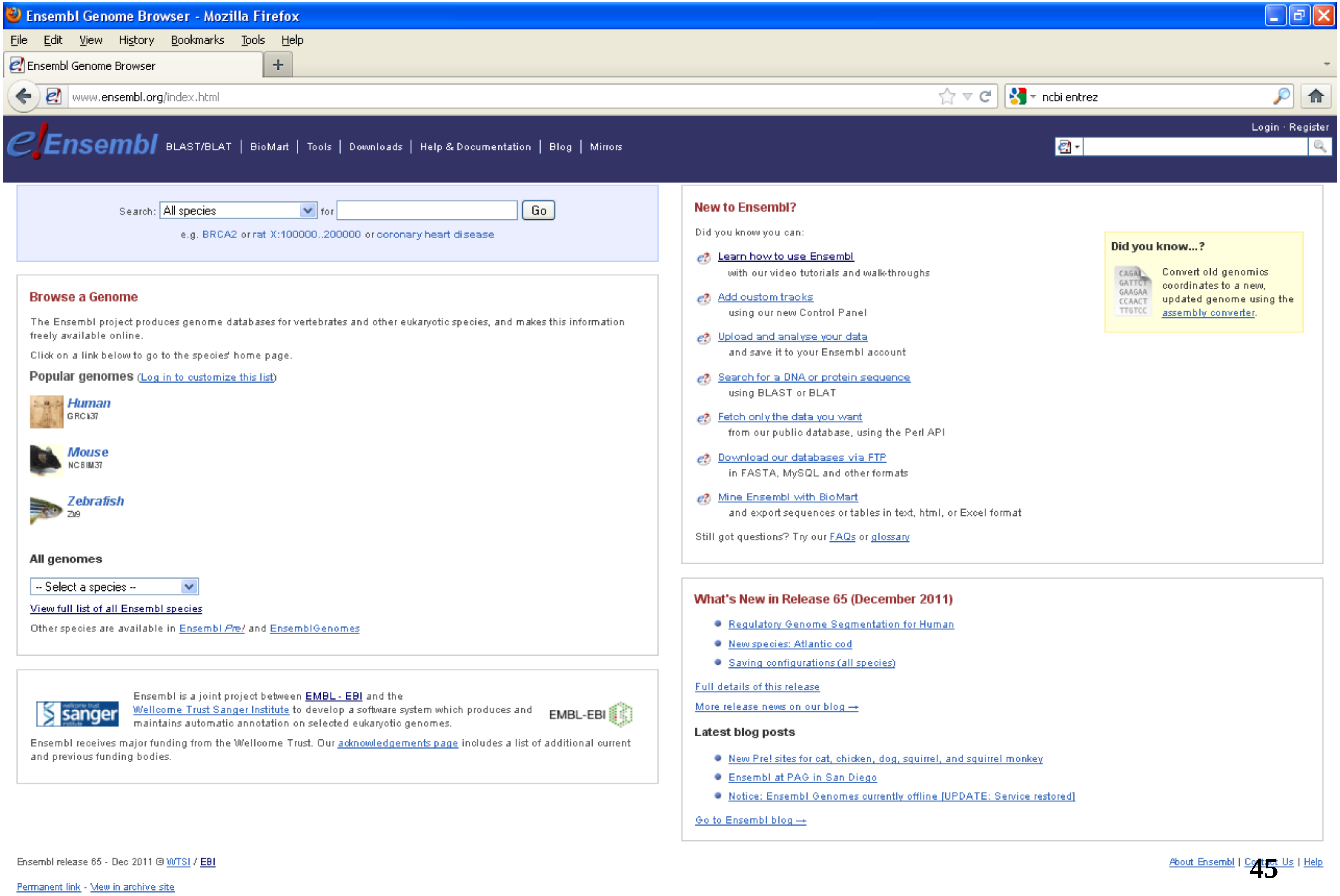
This is a gene with 10 ESTs associated; the cluster size is 10

UniGene clusters



Likely to be a real gene

Ensembl website (1)


Ensembl Genome Browser - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Ensembl Genome Browser
www.ensembl.org/index.html
ncbi entrez
Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login · Register

Search: All species for Go

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

Human GRCh37

Mouse NCBIM37

Zebrafish Zv9

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

New to Ensembl?

Did you know you can:

- Learn how to use Ensembl with our video tutorials and walk-throughs
- Add custom tracks using our new Control Panel
- Upload and analyse your data and save it to your Ensembl account
- Search for a DNA or protein sequence using BLAST or BLAT
- Fetch only the data you want from our public database, using the Perl API
- Download our databases via FTP in FASTA, MySQL and other formats
- Mine Ensembl with BioMart and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs or glossary

Did you know...?

Convert old genomics coordinates to a new, updated genome using the assembly converter.

What's New in Release 65 (December 2011)

- Regulatory Genome Segmentation for Human
- New species: Atlantic cod
- Saving configurations (all species)

Full details of this release

More release news on our blog →

Latest blog posts

- New Pre! sites for cat, chicken, dog, squirrel, and squirrel monkey
- Ensembl at PAG in San Diego
- Notice: Ensembl Genomes currently offline [UPDATE: Service restored]

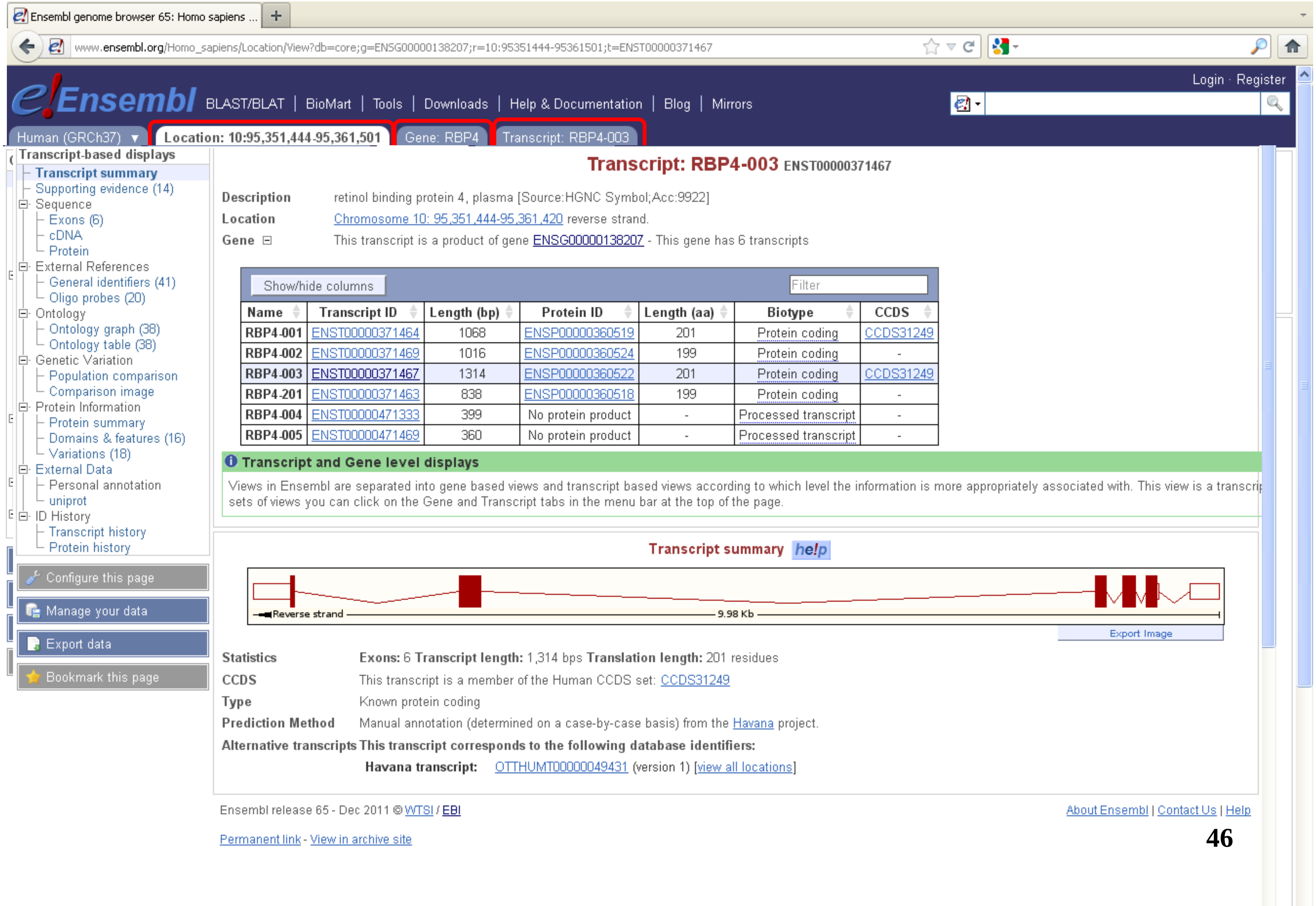
Go to Ensembl blog →

Ensembl release 65 - Dec 2011 © WTSI / EBI

Permanent link - View in archive site

About Ensembl | Contact Us | Help

Ensembl website (2)



Transcript: RBP4-003 ENST00000371467

Description retinol binding protein 4, plasma [Source:HGNC Symbol;Acc:9922]

Location Chromosome 10: 95,351,444-95,361,420 reverse strand.

Gene This transcript is a product of gene ENSG00000138207 - This gene has 6 transcripts

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RBP4-001	ENST00000371464	1068	ENSP00000360519	201	Protein coding	CCDS31249
RBP4-002	ENST00000371469	1016	ENSP00000360524	199	Protein coding	-
RBP4-003	ENST00000371467	1314	ENSP00000360522	201	Protein coding	CCDS31249
RBP4-201	ENST00000371463	838	ENSP00000360518	199	Protein coding	-
RBP4-004	ENST00000471333	399	No protein product	-	Processed transcript	-
RBP4-005	ENST00000471469	360	No protein product	-	Processed transcript	-

Transcript and Gene level displays

Views in Ensembl are separated into gene based views and transcript based views according to which level the information is more appropriately associated with. This view is a transcri sets of views you can click on the Gene and Transcript tabs in the menu bar at the top of the page.

Transcript summary he!p

Statistics **Exons:** 6 **Transcript length:** 1,314 bps **Translation length:** 201 residues

CCDS This transcript is a member of the Human CCDS set: CCDS31249

Type Known protein coding

Prediction Method Manual annotation (determined on a case-by-case basis) from the Havana project.

Alternative transcripts **This transcript corresponds to the following database identifiers:**

Havana transcript: OTTHUMT00000049431 (version 1) [view all locations]

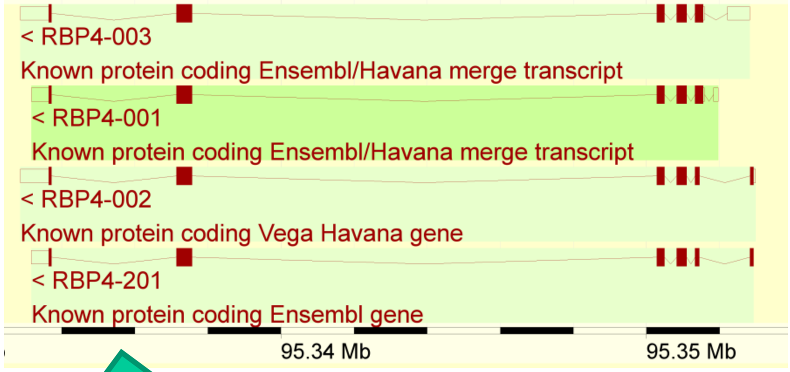
Ensembl release 65 - Dec 2011 © WTSI / EBI

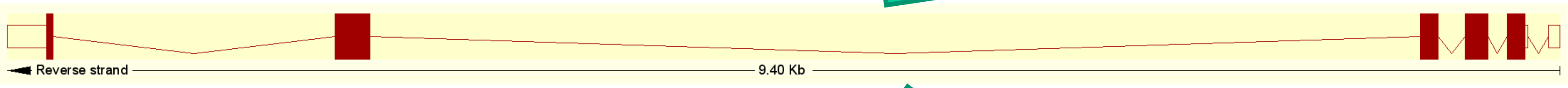
About Ensembl | Contact Us | Help

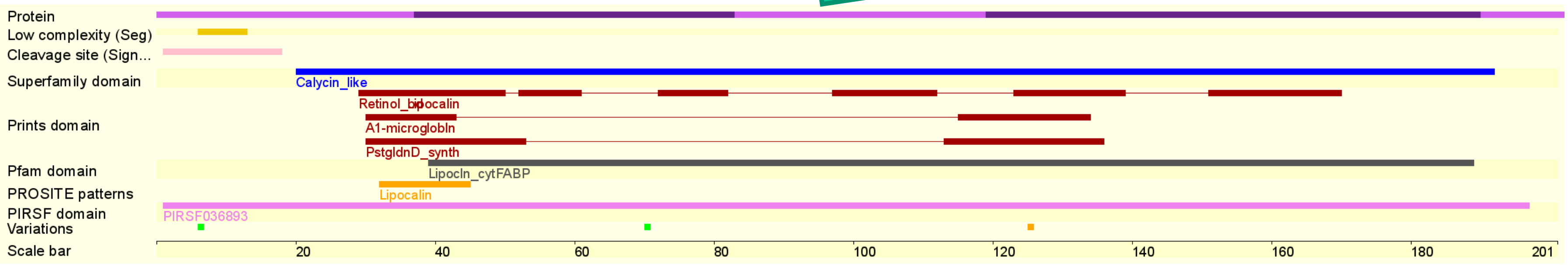
Permanent link - View in archive site

Ensembl identifiers

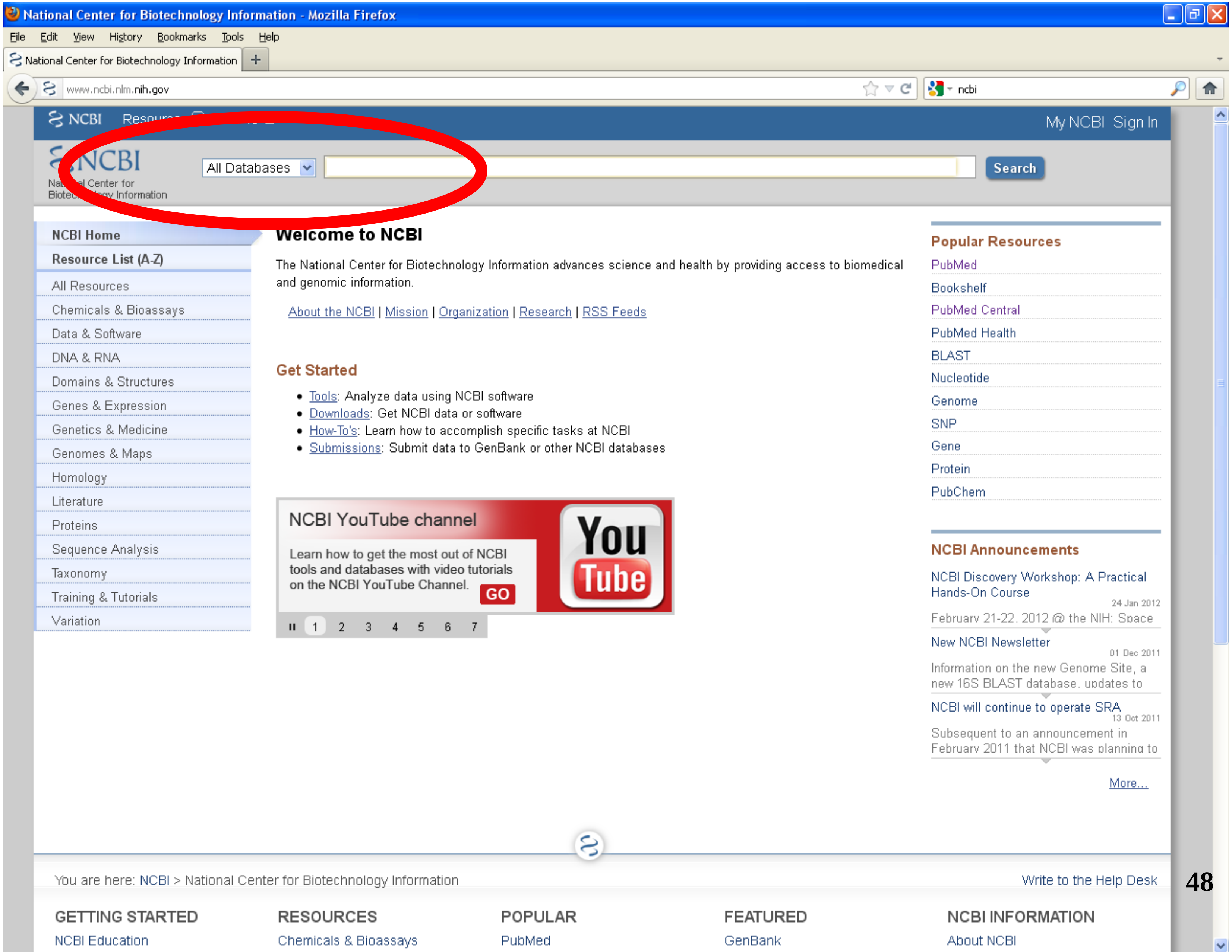
- Gene: ENSG…
- Transcript: ENST…
- Protein: ENSP…


< RBP4-003
Known protein coding Ensembl/Havana merge transcript
< RBP4-001
Known protein coding Ensembl/Havana merge transcript
< RBP4-002
Known protein coding Vega Havana gene
< RBP4-201
Known protein coding Ensembl gene
95.34 Mb
95.35 Mb





NCBI website (1)


National Center for Biotechnology Information - Mozilla Firefox
File Edit View History Bookmarks Tools Help
National Center for Biotechnology Information
www.ncbi.nlm.nih.gov
ncbi
NCBI Resour
My NCBI Sign In
NCBI
Nat al Center for
Biotec ogy Information
All Databases
Search

NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

NCBI YouTube channel

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

NCBI Discovery Workshop: A Practical Hands-On Course 24 Jan 2012

February 21-22, 2012 @ the NIH: Space

New NCBI Newsletter 01 Dec 2011

Information on the new Genome Site, a new 16S BLAST database, updates to

NCBI will continue to operate SRA 13 Oct 2011

Subsequent to an announcement in February 2011 that NCBI was planning to

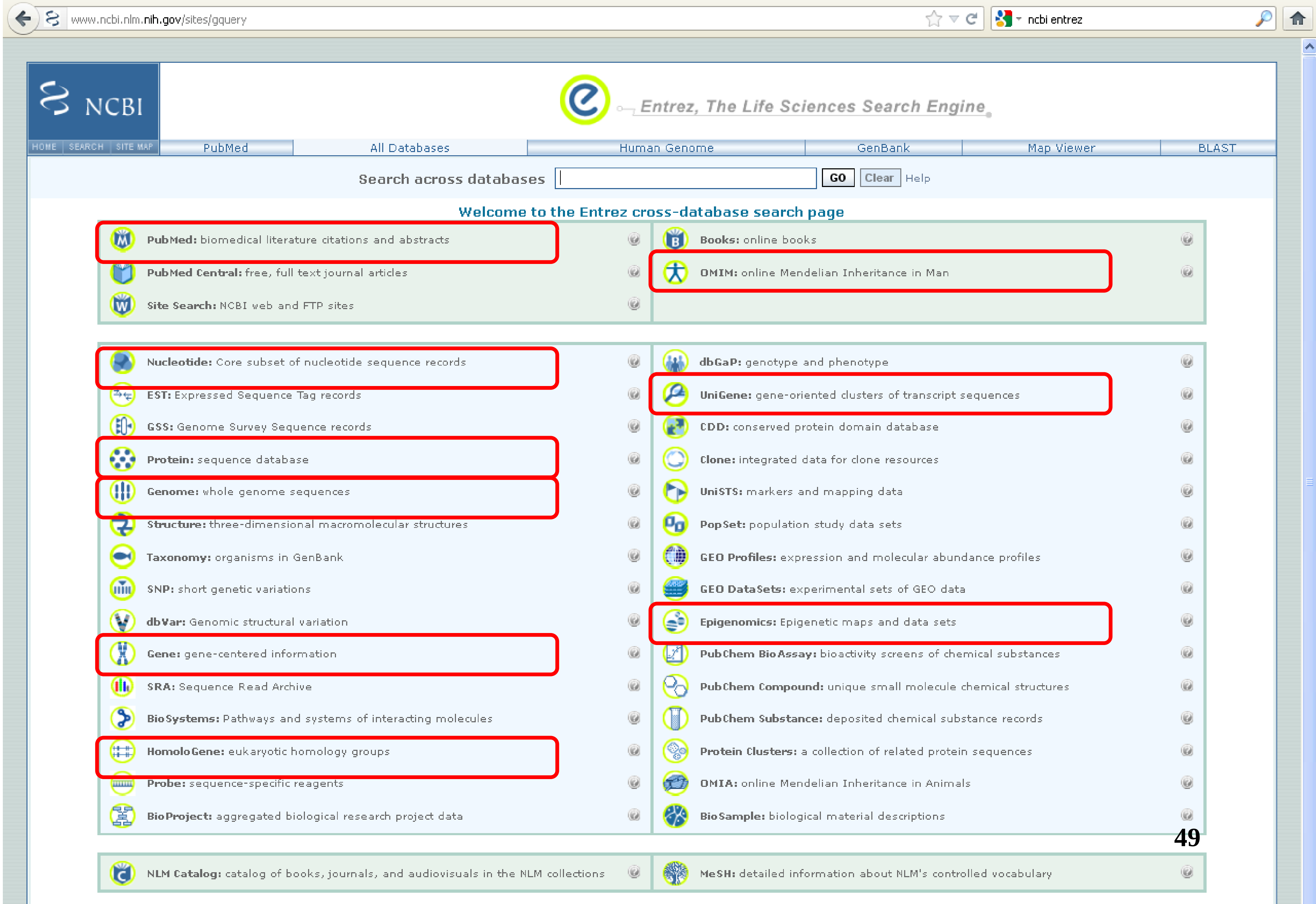
More...

You are here: NCBI > National Center for Biotechnology Information

Write to the Help Desk

GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
NCBI Education	Chemicals & Bioassays	PubMed	GenBank	About NCBI

NCBI website (2)



Entrez, The Life Sciences Search Engine

Search across databases

Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts

PubMed Central: free, full text journal articles

Site Search: NCBI web and FTP sites

Books: online books

OMIM: online Mendelian Inheritance in Man

Nucleotide: Core subset of nucleotide sequence records

EST: Expressed Sequence Tag records

GSS: Genome Survey Sequence records

Protein: sequence database

Genome: whole genome sequences

Structure: three-dimensional macromolecular structures

Taxonomy: organisms in GenBank

SNP: short genetic variations

dbVar: Genomic structural variation

Gene: gene-centered information

SRA: Sequence Read Archive

BioSystems: Pathways and systems of interacting molecules

HomoloGene: eukaryotic homology groups

Probe: sequence-specific reagents

BioProject: aggregated biological research project data

dbGaP: genotype and phenotype

UniGene: gene-oriented clusters of transcript sequences

CDD: conserved protein domain database

Clone: integrated data for clone resources

UniSTS: markers and mapping data

PopSet: population study data sets

GEO Profiles: expression and molecular abundance profiles

GEO DataSets: experimental sets of GEO data

Epigenomics: Epigenetic maps and data sets

PubChem BioAssay: bioactivity screens of chemical substances

PubChem Compound: unique small molecule chemical structures

PubChem Substance: deposited chemical substance records

Protein Clusters: a collection of related protein sequences

OMIA: online Mendelian Inheritance in Animals

BioSample: biological material descriptions

NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections

MeSH: detailed information about NLM's controlled vocabulary

NCBI identifiers

- RefSeq:
  - Chromosome: NC_
  - mRNA: NM_
  - Protein: NP_
- Genbank:
  - Many types of IDs
- Entrez gene ID:
  - Number
- OMIM ID:
  - Number
- Pubmed ID:
  - Number
- UniGene ID:
  - Abbreviation of species.number (e.g. Hs.50223)