Chapter 8:
DNA: The eukaryotic chromosome

Learning objectives

Upon completing this chapter you should be able to:

- define features of eukaryotic genomes such as the *C* value;
- define five major types of repetitive DNA and bioinformatics resources to study them;
- describe eukaryotic genes;
- explain several categories of regulatory regions;
- use bioinformatics tools to compare eukaryotic DNA;
- define single-nucleotide polymorphisms (SNPs) and analyze SNP data; and
- compare and contrast methods to measure chromosomal change.

Outline

Introduction

General features of eukaryotic genomes and chromosomes

- *C* value paradox; organization; genome browsers
- Analysis of chromosomes using BioMart and biomaRt
- ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

- Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

- Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

- Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

Variation in chromosomal DNA

- Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

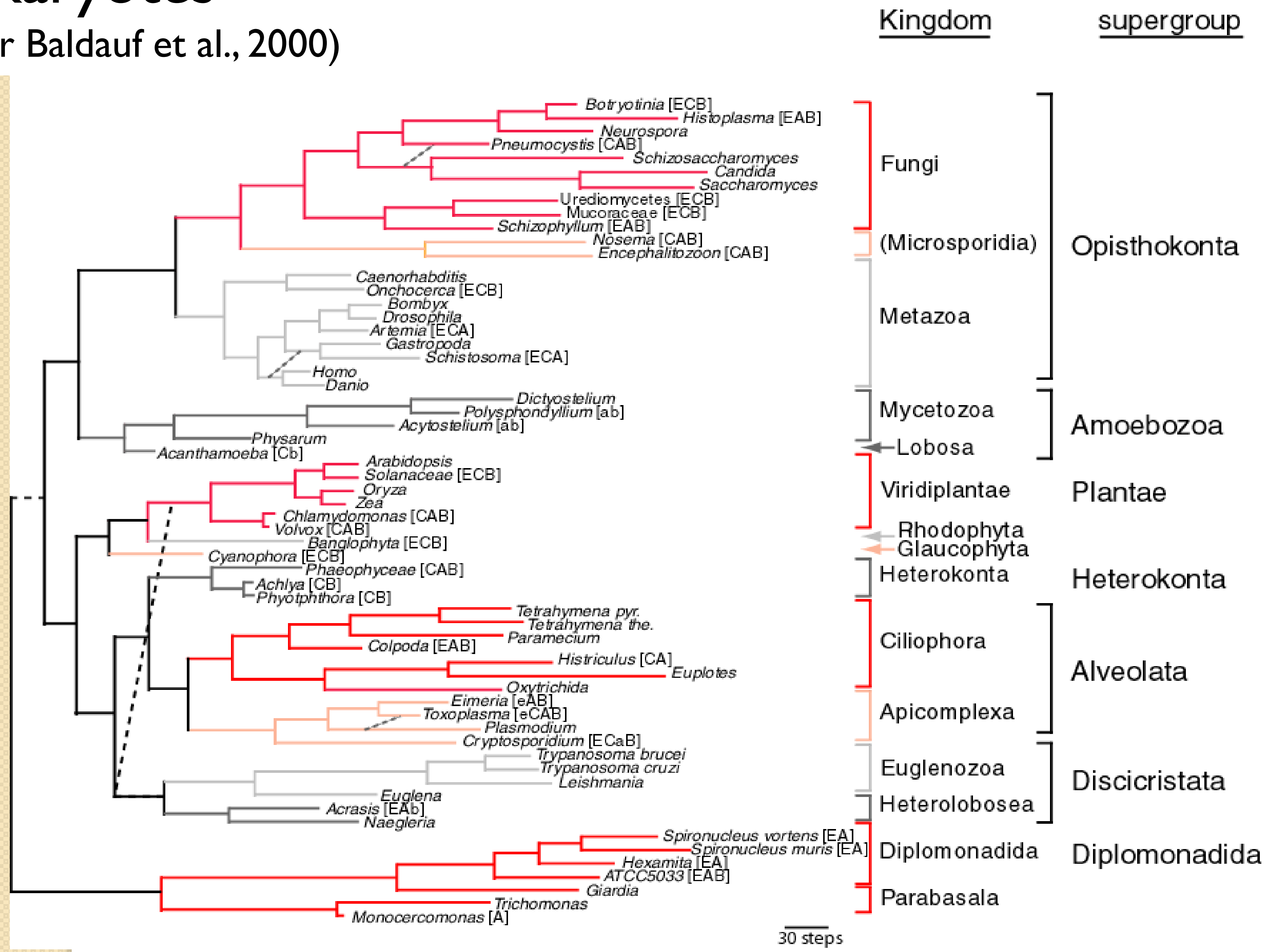
Introduction to the eukaryotes

Eukaryotes are single-celled or multicellular organisms that are distinguished from prokaryotes by the presence of a membrane-bound nucleus, an extensive system of intracellular organelles, and a cytoskeleton.

We will explore the eukaryotes using a phylogenetic tree by Baldauf et al. (Science, 2000). This tree was made by concatenating four protein sequences: elongation factor 1a, actin, α-tubulin, and β-tubulin.

Eukaryotes

(after Baldauf et al., 2000)



General features of the eukaryotes

- Some of the general features of eukaryotes that distinguish them from prokaryotes (bacteria and archaea) are:

• Eukaryotes include many multicellular organisms, in addition to unicellular organisms.
• Eukaryotes have [1] a membrane-bound nucleus, [2] intracellular organelles, and [3] a cytoskeleton
• Most eukaryotes undergo sexual reproduction
• The genome size of eukaryotes spans a wider range than that of most prokaryotes
• Eukaryotic genomes have a lower density of genes
• Prokaryotes are haploid; eukaryotes have varying ploidy
• Eukaryotic genomes tend to be organized into linear chromosomes with a centromere and telomeres.

Questions about eukaryotic chromosomes

What are the sizes of eukaryotic genomes, and how are they organized into chromosomes?

What are the types of repetitive DNA elements? What are their properties and amounts?

What are the types of genes? How can they be identified?

What is the mutation rate across the genome; what are the selective forces affecting genome evolution?

What is the spectrum of variation between species (comparative genomics) and within species?

Features of bacterial and eukaryotic genomes

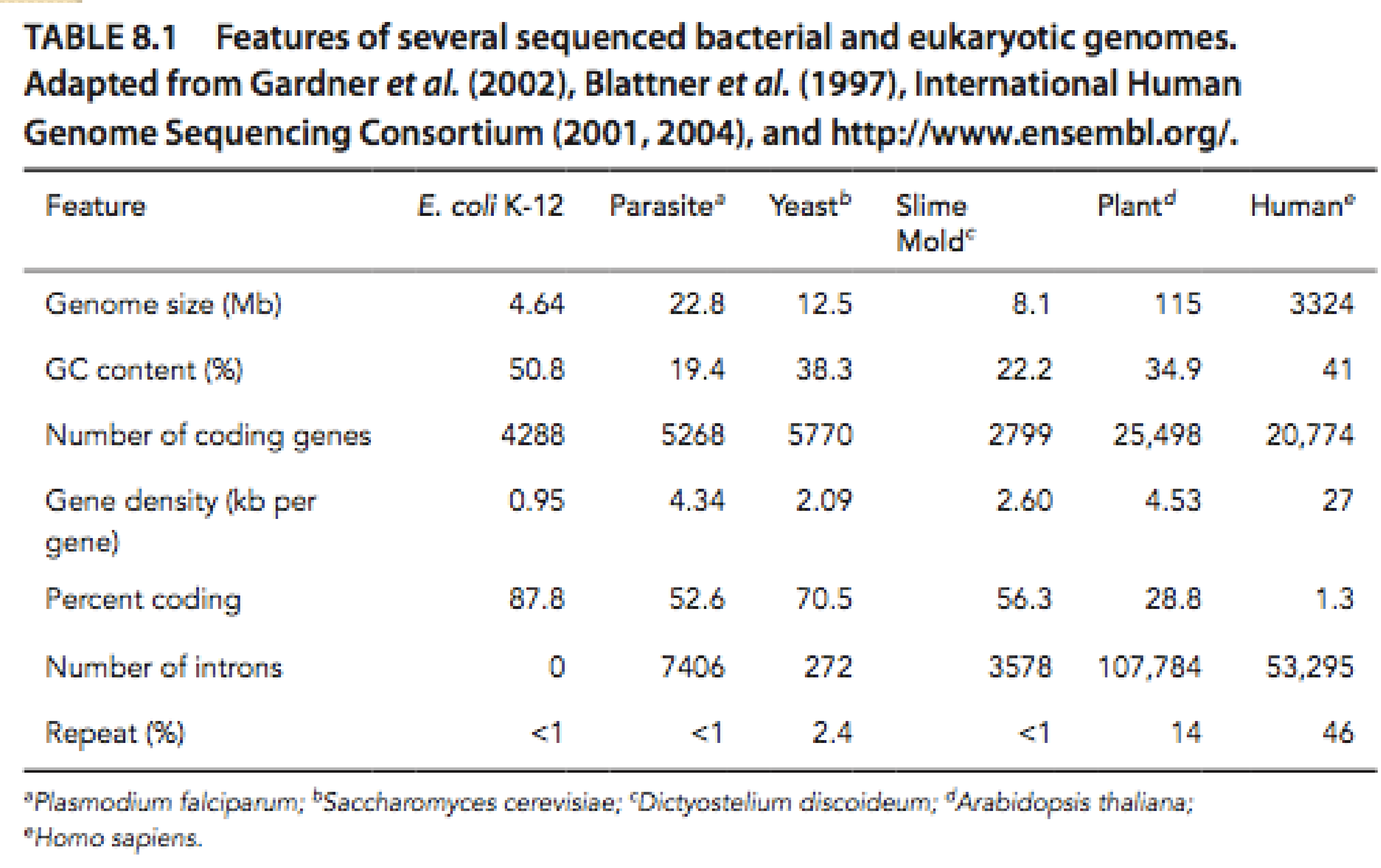


TABLE 8.1 Features of several sequenced bacterial and eukaryotic genomes. Adapted from Gardner *et al.* (2002), Blattner *et al.* (1997), International Human Genome Sequencing Consortium (2001, 2004), and http://www.ensembl.org/.

Feature	*E. coli* K-12	Parasite[a]	Yeast[b]	Slime Mold[c]	Plant[d]	Human[e]
Genome size (Mb)	4.64	22.8	12.5	8.1	115	3324
GC content (%)	50.8	19.4	38.3	22.2	34.9	41
Number of coding genes	4288	5268	5770	2799	25,498	20,774
Gene density (kb per gene)	0.95	4.34	2.09	2.60	4.53	27
Percent coding	87.8	52.6	70.5	56.3	28.8	1.3
Number of introns	0	7406	272	3578	107,784	53,295
Repeat (%)	<1	<1	2.4	<1	14	46

[a] *Plasmodium falciparum;* [b] *Saccharomyces cerevisiae;* [c] *Dictyostelium discoideum;* [d] *Arabidopsis thaliana;* [e] *Homo sapiens.*

Outline

Introduction

General features of eukaryotic genomes and chromosomes

- *C* value paradox; organization; genome browsers
- Analysis of chromosomes using BioMart and biomaRt
- ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

- Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

- Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

- Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

Variation in chromosomal DNA

- Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

C value paradox:
why eukaryotic genome sizes vary

The haploid genome size of eukaryotes, called the *C* value, varies enormously.

Small genomes include:
Encephalitozoon cuniculi (2.9 Mb)
A variety of fungi (10-40 Mb)
Takifugu rubripes (pufferfish)(365 Mb)(same number of genes as other fish or as the human genome, but 1/8th the size)

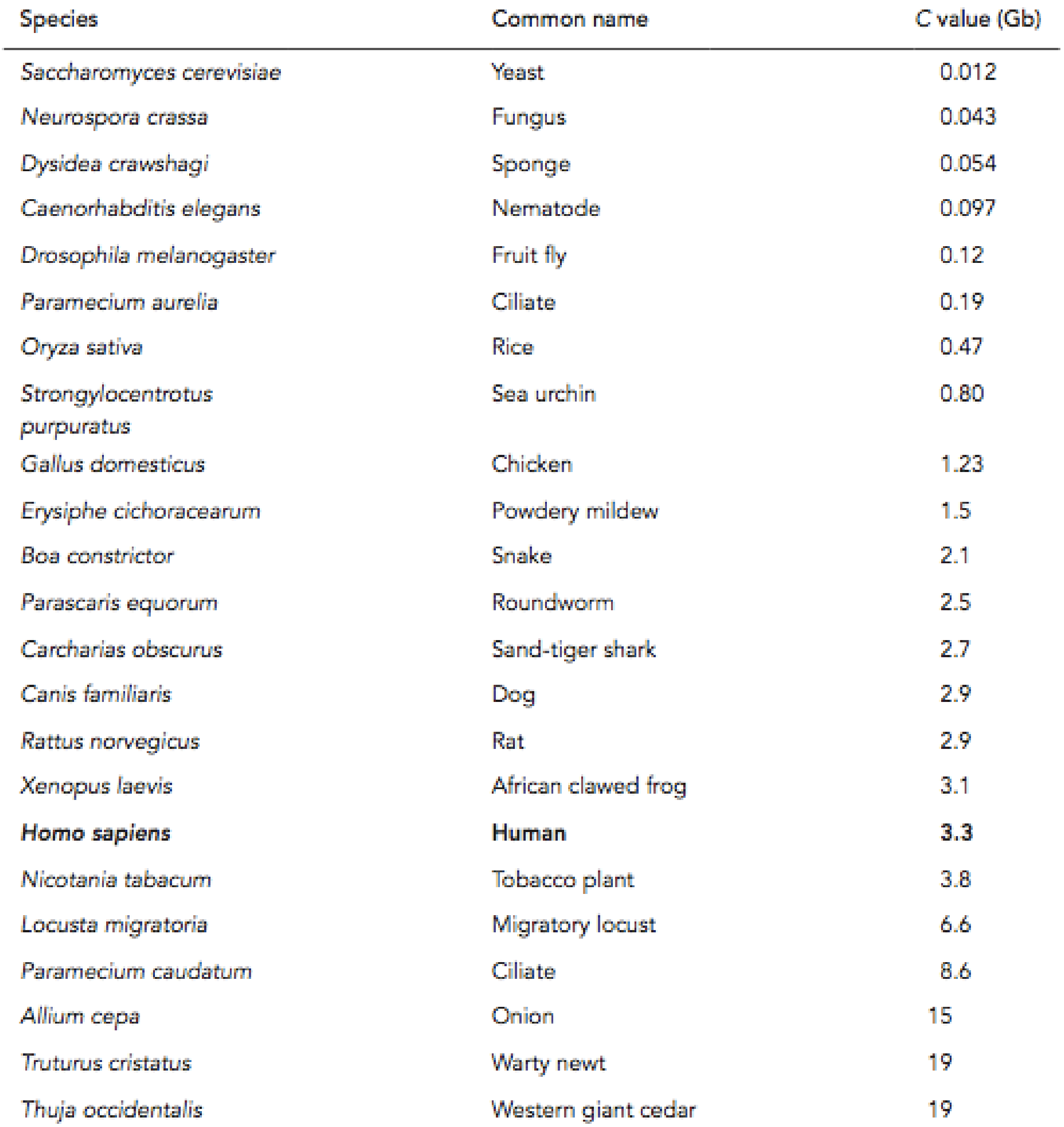
Large genomes include:
Pinus resinosa (Canadian red pine)(68 Gb)
Protopterus aethiopicus (Marbled lungfish)(140 Gb)
Amoeba dubia (amoeba)(690 Gb)

C value paradox: why eukaryotic genome sizes vary

The range in *C* values does not correlate well with the complexity of the organism. This phenomenon is called the *C* value paradox.

The solution to this "paradox" is that genomes are filled with variable amounts of large tracts of noncoding, often repetitive DNA sequences.

Genome size (*C* value) for various eukaryotic species



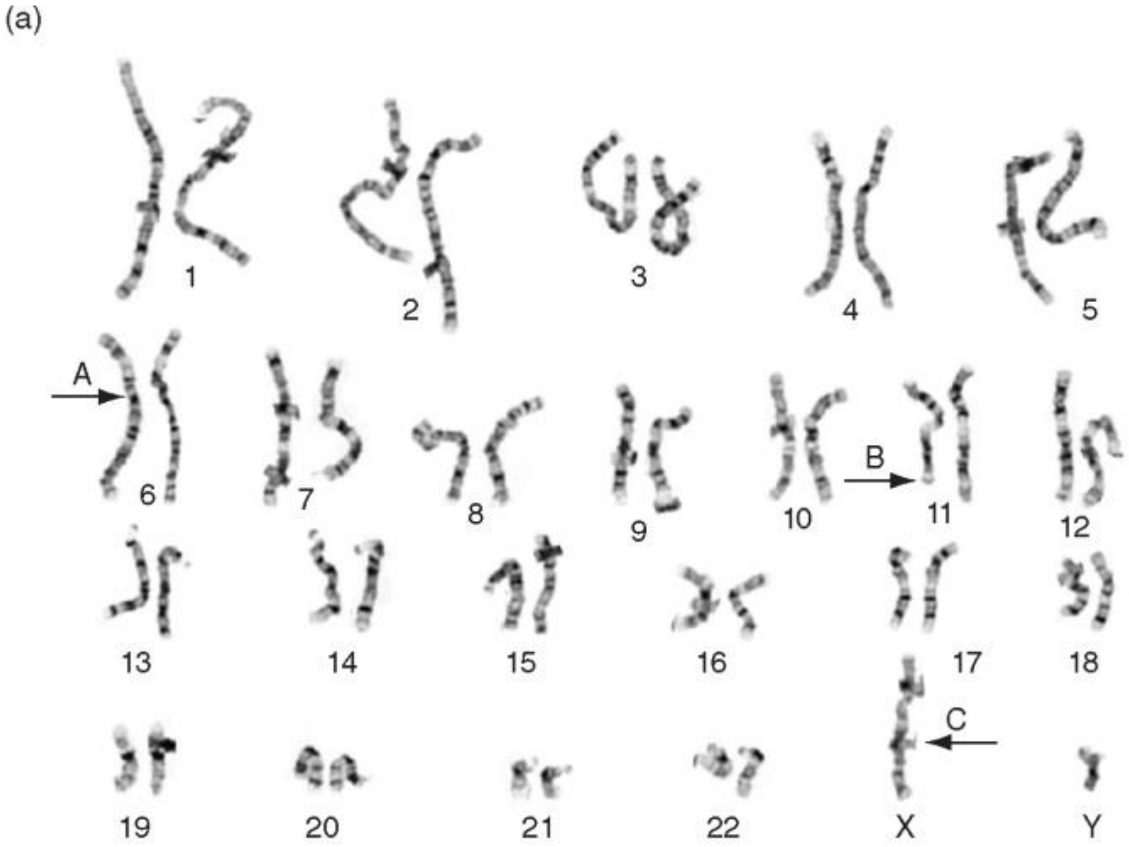
Species	Common name	C value (Gb)
Saccharomyces cerevisiae	Yeast	0.012
Neurospora crassa	Fungus	0.043
Dysidea crawshagi	Sponge	0.054
Caenorhabditis elegans	Nematode	0.097
Drosophila melanogaster	Fruit fly	0.12
Paramecium aurelia	Ciliate	0.19
Oryza sativa	Rice	0.47
Strongylocentrotus purpuratus	Sea urchin	0.80
Gallus domesticus	Chicken	1.23
Erysiphe cichoracearum	Powdery mildew	1.5
Boa constrictor	Snake	2.1
Parascaris equorum	Roundworm	2.5
Carcharias obscurus	Sand-tiger shark	2.7
Canis familiaris	Dog	2.9
Rattus norvegicus	Rat	2.9
Xenopus laevis	African clawed frog	3.1
Homo sapiens	**Human**	**3.3**
Nicotania tabacum	Tobacco plant	3.8
Locusta migratoria	Migratory locust	6.6
Paramecium caudatum	Ciliate	8.6
Allium cepa	Onion	15
Truturus cristatus	Warty newt	19
Thuja occidentalis	Western giant cedar	19

Eukaryotic genomes are organized into chromosomes

Genomic DNA is organized in chromosomes. The diploid number of chromosomes is constant in each species (e.g. 46 in human). Chromosomes are distinguished by a centromere and telomeres.

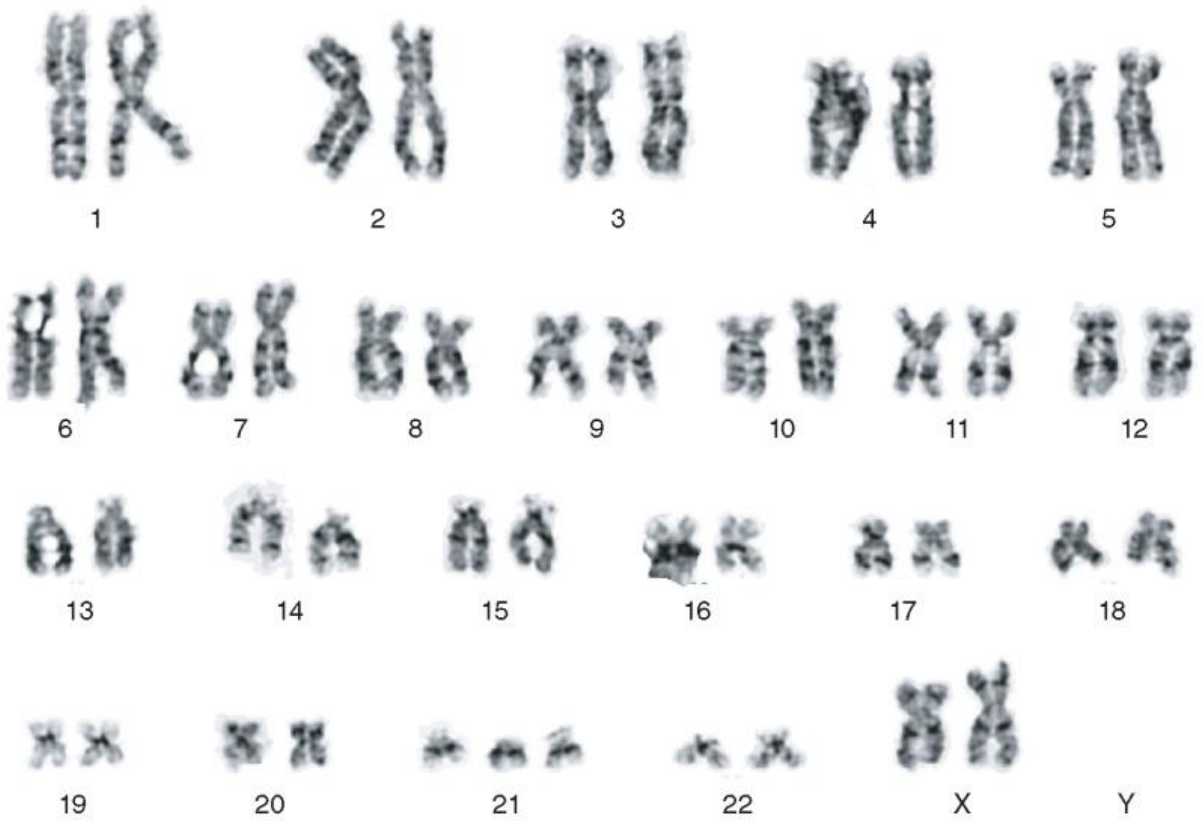
The chromosomes are routinely visualized by karyotyping (imaging the chromosomes during metaphase, when each chromosome is a pair of sister chromatids).

Human karyotypes: boy with deletion on 11q

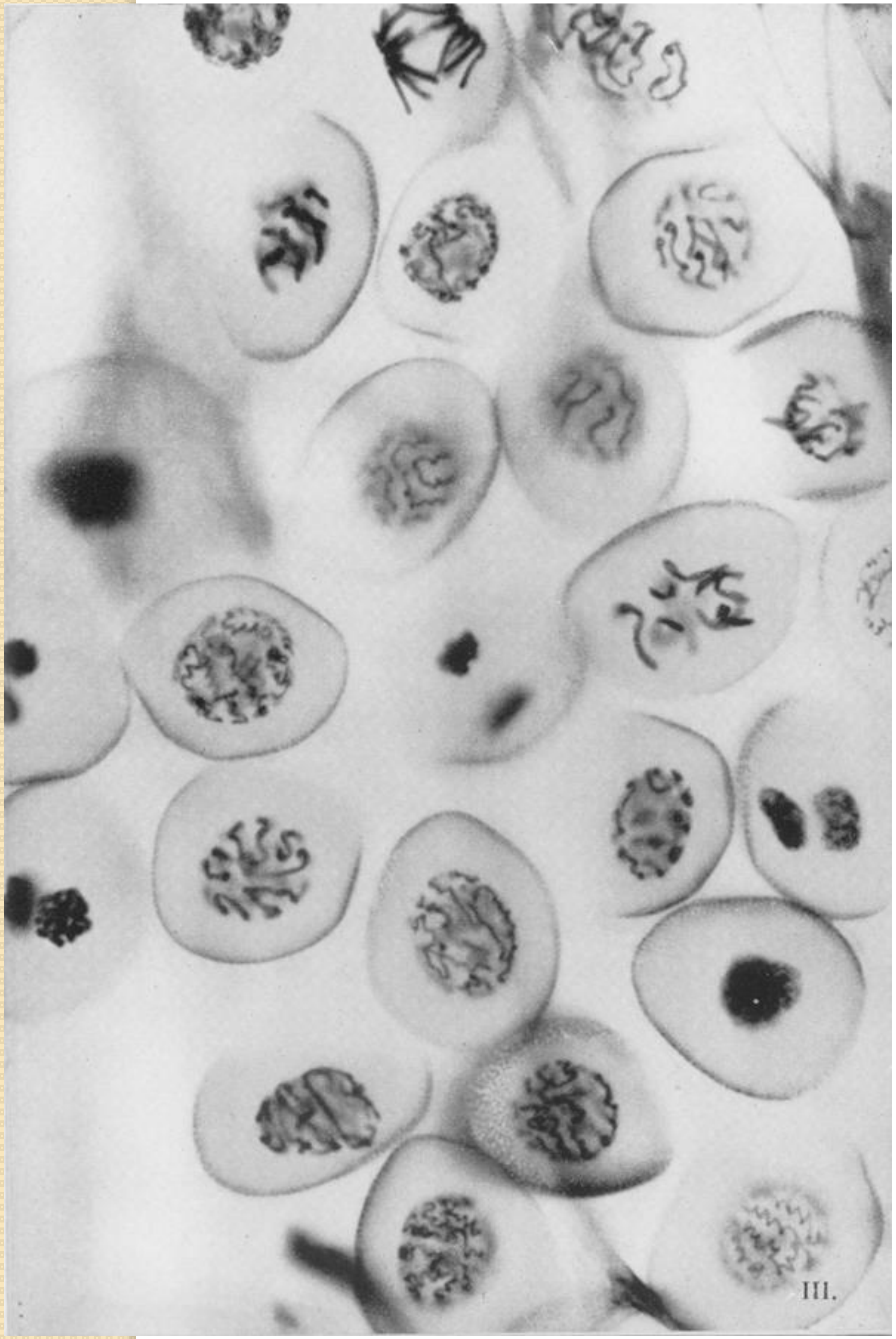


Arrows A, C mark examples of centromeres
p = short arm (“petit”)
q = long arm (letter after p)

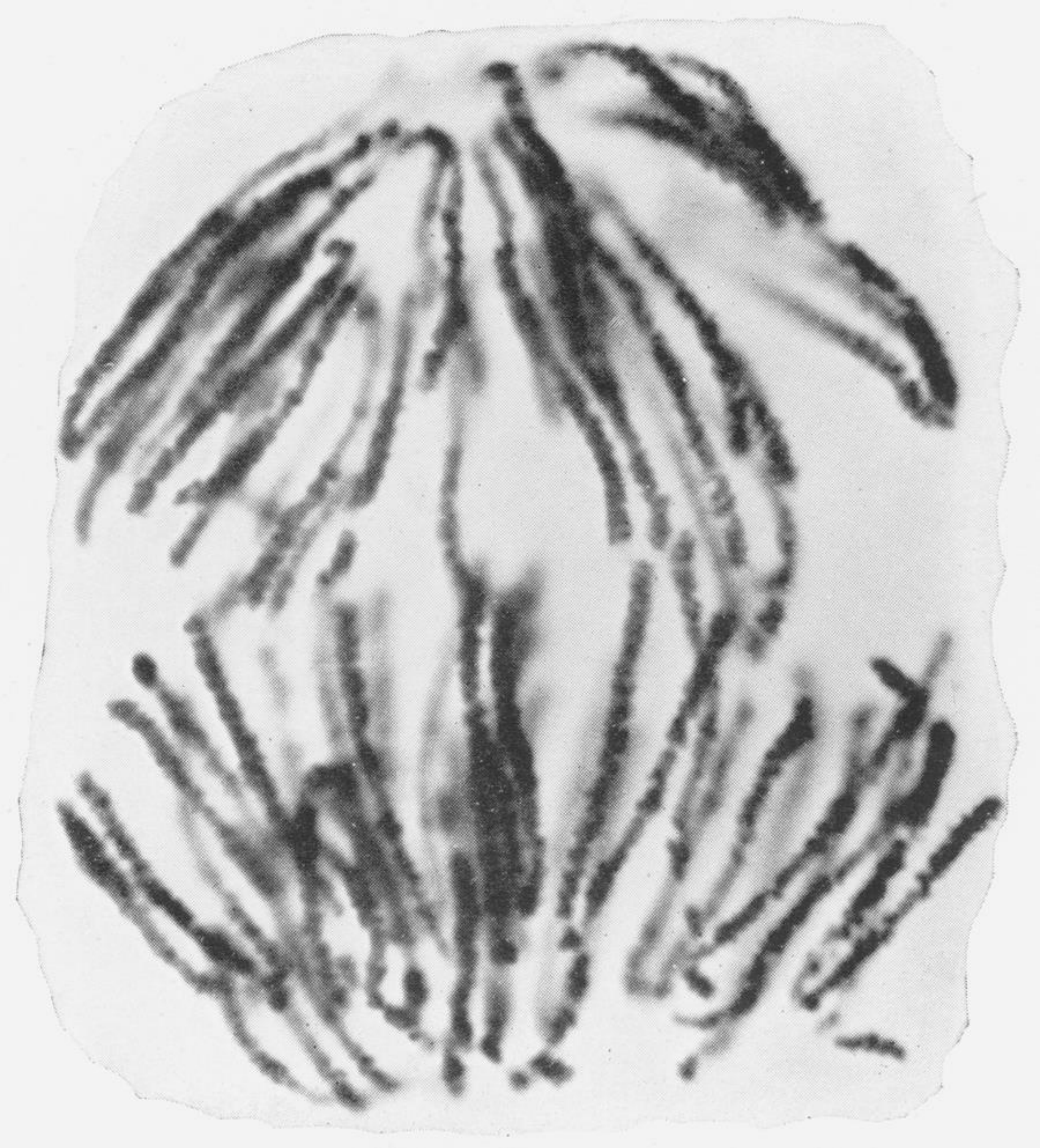
Human karyotypes: girl with trisomy 21



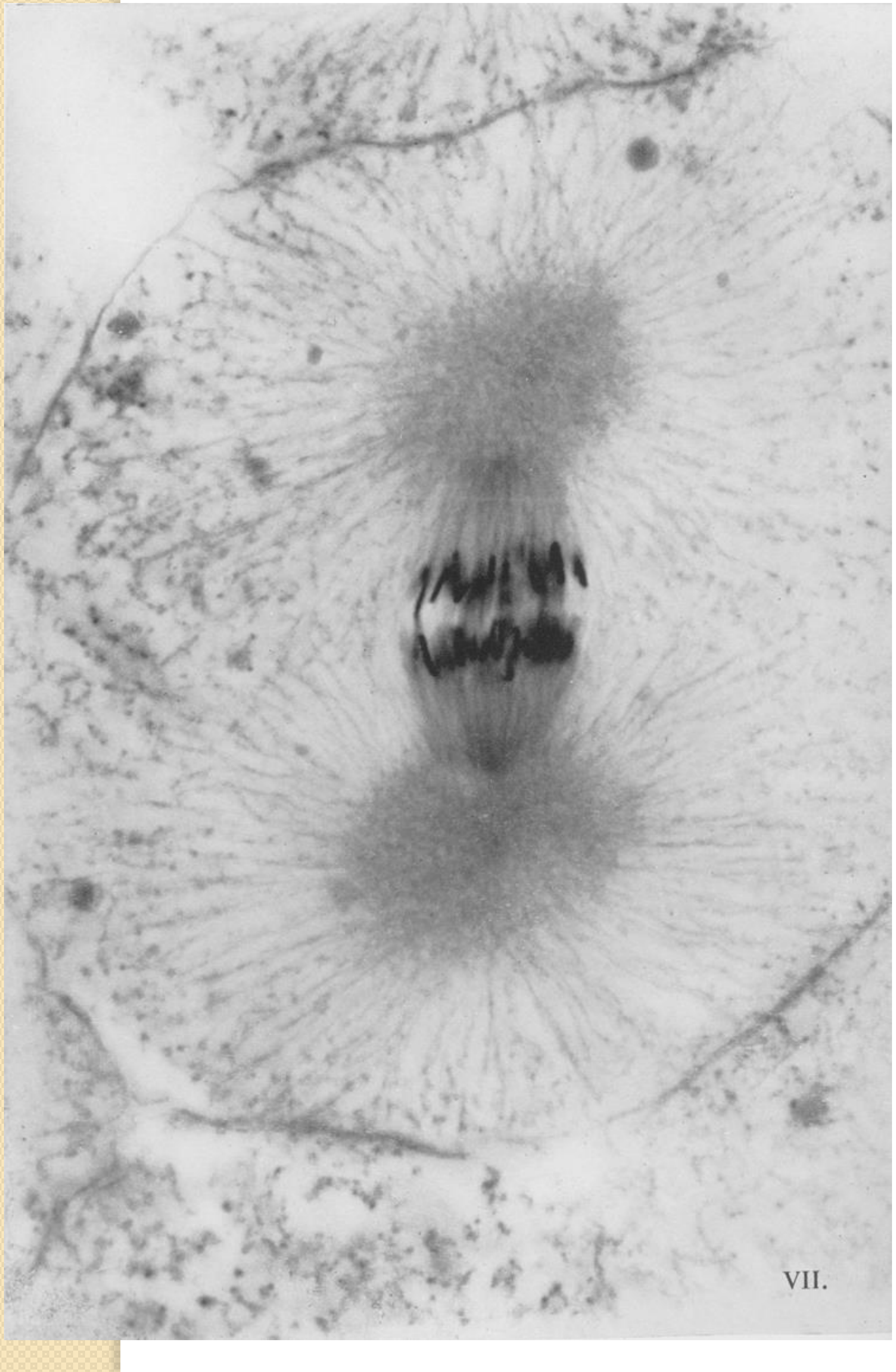
Note three copies of chromosome 21.



Mitosis in *Paris quadrifolia*, Liliaceae, showing all stages from prophase to telophase. *n* = 10 (Darlington).



Root tip squashes showing anaphase separation. *Fritillaria pudica*, 3*x* = 39, spiral structure of chromatids revealed by pressure after cold treatment. Darlington.



Cleavage mitosis in the teleostean fish, *Coregonus clupeoides*, in the middle of anaphase. Spindle structure revealed by slow fixation. Darlington.

The eukaryotic chromosome: the centromere

The centromere is a primary constriction where the chromosome attaches to the spindle fibers; here the boundary between sister chromatids is not clear. It may be in the middle (metacentric) or the end (acrocentric).

If a chromosome has two centromeres spaced apart (dicentric) then at anaphase there is a 50% chance that a single chromatid would be pulled to opposite poles of the mitotic spindle. This would result in a bridge formation and chromosome breakage.

The eukaryotic chromosome: the centromere

The short arm of the acrocentric autosomes has a secondary constriction usually containing a nucleolar organizer. This contains the genes for 18S and 28S ribosomal RNA.

The eukaryotic chromosome: the telomere

The telomere is a region of highly repetitive DNA at either end of a linear chromosome. Telomeres include nucleoprotein complexes that function in the protection, replication, and stabilization of chromosome ends. Telomeres of many eukaryotes have tandemly repeated DNA sequences.

Outline

Introduction

General features of eukaryotic genomes and chromosomes

- *C* value paradox; organization; genome browsers
- Analysis of chromosomes using BioMart and biomaRt
- ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

- Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

- Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

- Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

Variation in chromosomal DNA

- Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

Three main genome browsers

There are three principal genome browsers for eukaryotes:

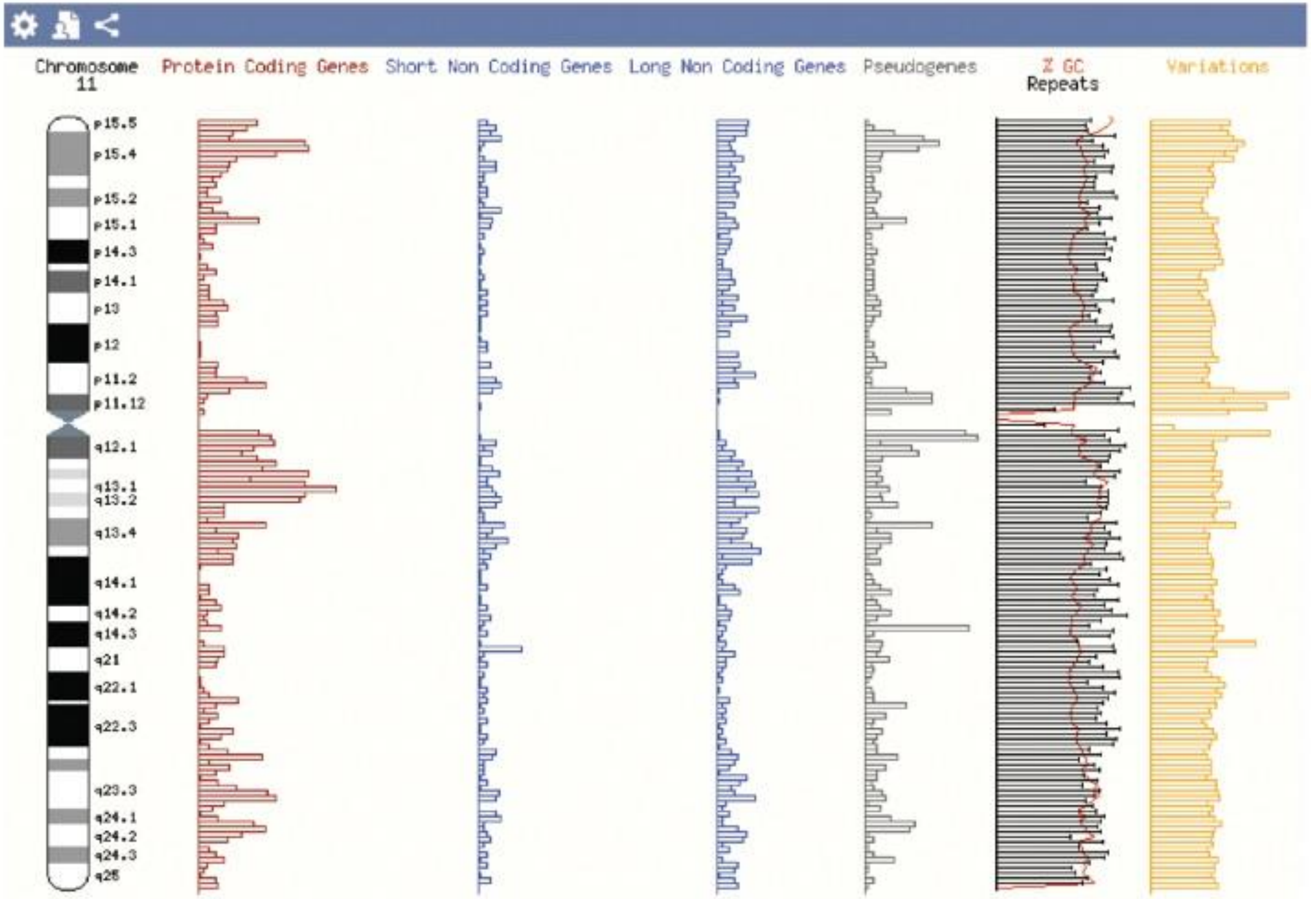
(1) NCBI offers Map Viewer

(2) Ensembl (www.ensembl.org) offers browsers for dozens of genomes

(3) UCSC (http://genome.ucsc.edu) offers genome and table browsers for dozens of organisms. We will focus on this browser.

Ensembl browser: view of human chromosome 11

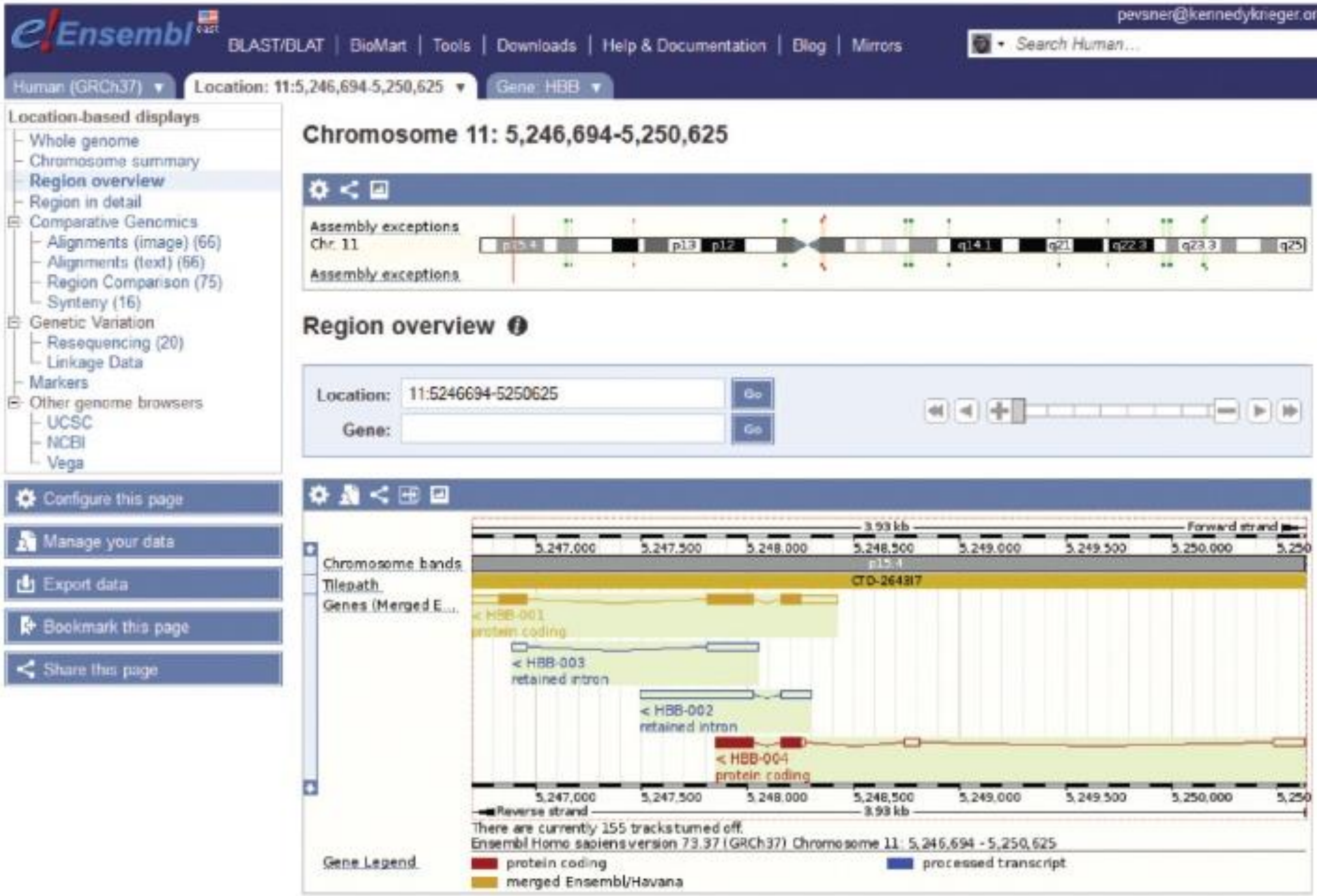
(a) Ensembl: chromosome summary



Chromosome summary view includes many configuration options.

Ensembl browser: view of human chromosome 11

(b) Ensembl: Region overview



Chromosome 11: 5,246,694-5,250,625

Location: 11:5246694-5250625

Gene:

There are currently 155 tracks turned off.

Ensembl Homo sapiens version 73.37 (GRCh37) Chromosome 11: 5,246,694 - 5,250,625

Gene Legend

🟥 protein coding

🟧 merged Ensembl/Havana

🟦 processed transcript

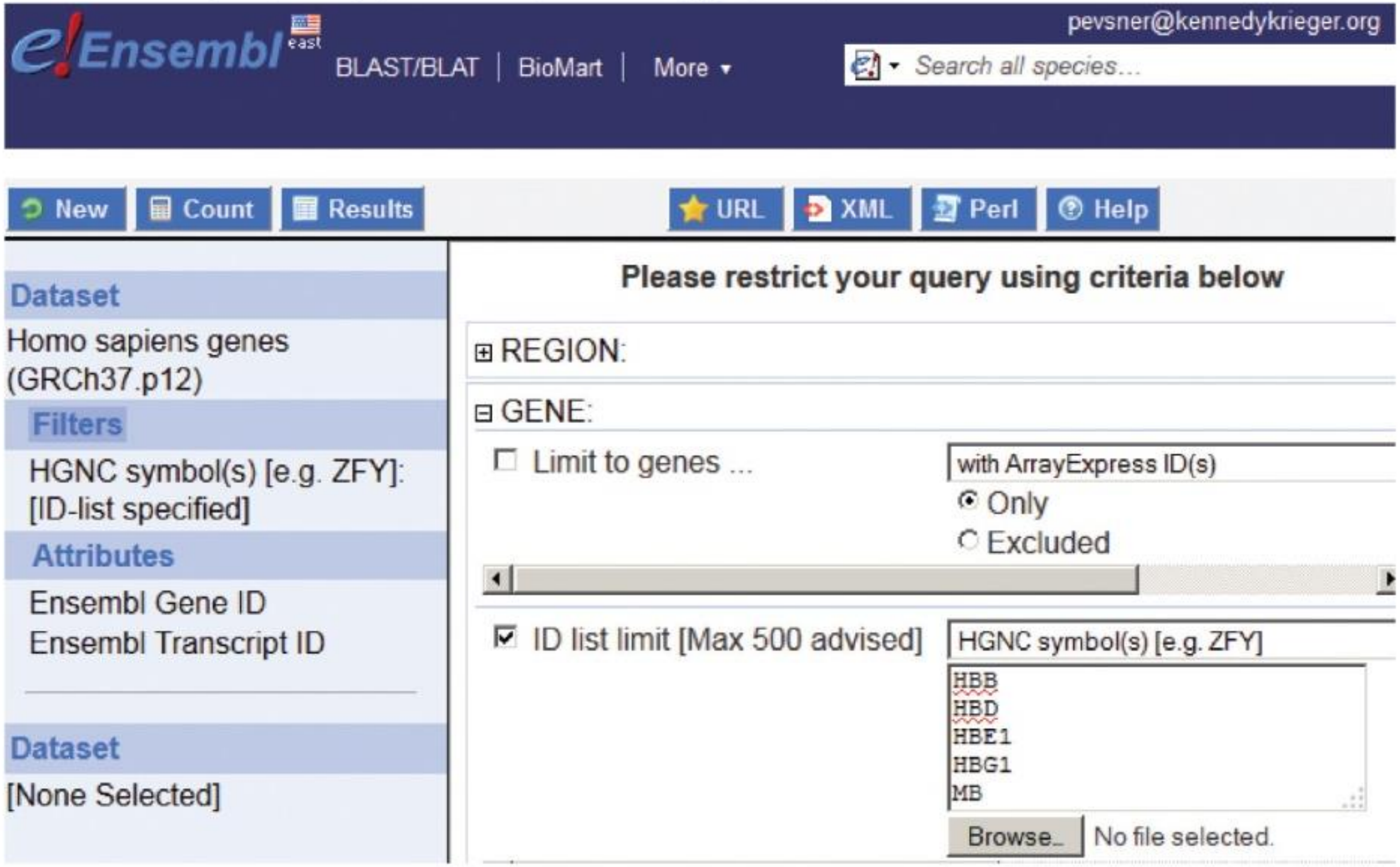
Region overview includes an ideogram (representation of a chromosome) with a red bar including the location of *HBB*. Hundreds of tracks may be added (gear-shaped link).

Outline

- Introduction
- General features of eukaryotic genomes and chromosomes
    - *C* value paradox; organization; genome browsers
    - Analysis of chromosomes using BioMart and biomaRt
    - ENCODE Project; critiques of ENCODE
- Repetitive DNA content of eukaryotic genomes
    - Noncoding and repetitive DNA sequences
- Gene content of eukaryotic chromosomes
    - Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE
- Regulatory regions of eukaryotic chromosomes
    - Databases of regulatory factors; ultraconserved elements; nonconserved elements
- Comparison of eukaryotic DNA
- Variation in chromosomal DNA
    - Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation
- Techniques to measure chromosomal change
- Perspective

Biomart service (Ensembl): query many databases

(a) BioMart at Ensembl: specify filters (input for which you want to apply queries)



Please restrict your query using criteria below

Dataset

Homo sapiens genes (GRCh37.p12)

Filters

HGNC symbol(s) [e.g. ZFY]: [ID-list specified]

Attributes

Ensembl Gene ID

Ensembl Transcript ID

Dataset

[None Selected]

⊞ REGION:

⊟ GENE:

□ Limit to genes ... with ArrayExpress ID(s)

● Only

○ Excluded

☑ ID list limit [Max 500 advised] HGNC symbol(s) [e.g. ZFY]

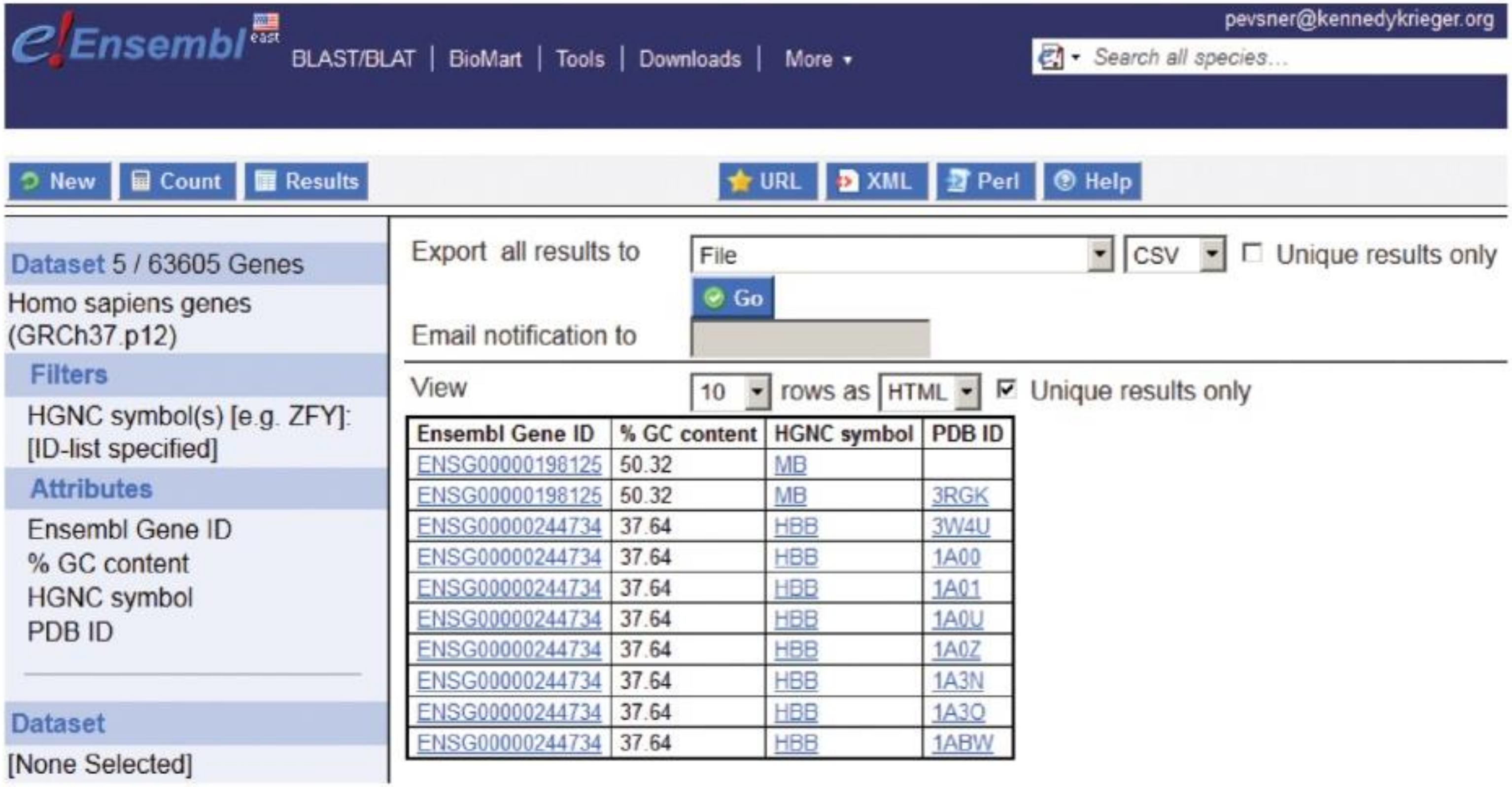
```
HBB
HBD
HBE1
HBG1
MB
```

Browse... No file selected.

Select a dataset (e.g. human genes), filters (e.g. chromosomal regions), and attributes (thousands are available). Click results. Here we ask for information (attributes) about a set of genes (given by a list of gene symbols in the box to the right).

Biomart service (Ensembl)

(b) BioMart output


Ensembl east
BLAST/BLAT | BioMart | Tools | Downloads | More ▾
pevsner@kennedykrieger.org
Search all species…
New
Count
Results
URL
XML
Perl
Help
Go

Dataset 5 / 63605 Genes

Homo sapiens genes (GRCh37.p12)

Filters

HGNC symbol(s) [e.g. ZFY]: [ID-list specified]

Attributes

Ensembl Gene ID
% GC content
HGNC symbol
PDB ID

Dataset

[None Selected]

Export all results to: File | CSV | □ Unique results only

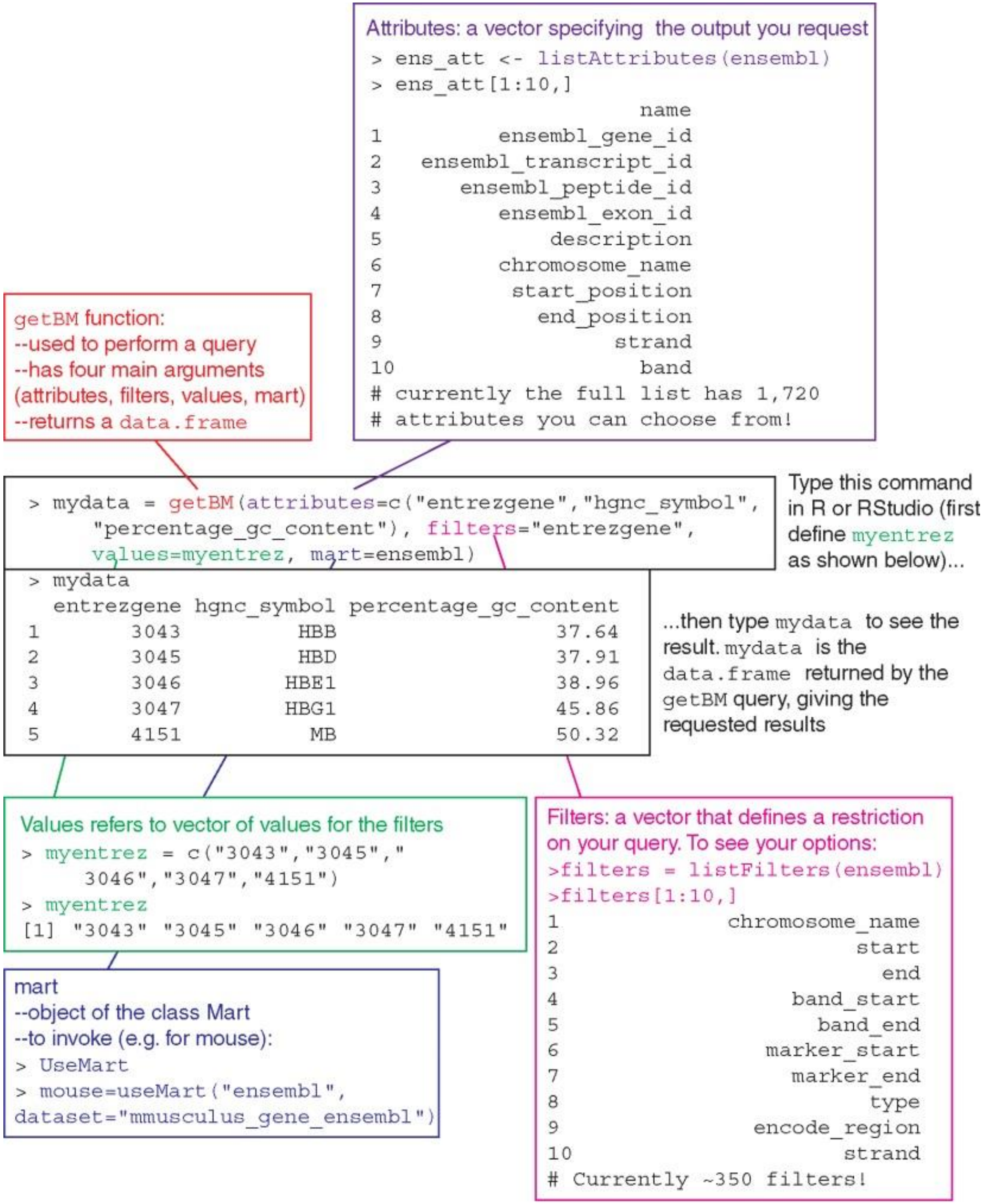
Email notification to

View 10 rows as HTML | ☑ Unique results only

Ensembl Gene ID	% GC content	HGNC symbol	PDB ID
ENSG00000198125	50.32	MB	
ENSG00000198125	50.32	MB	3RGK
ENSG00000244734	37.64	HBB	3W4U
ENSG00000244734	37.64	HBB	1A00
ENSG00000244734	37.64	HBB	1A01
ENSG00000244734	37.64	HBB	1A0U
ENSG00000244734	37.64	HBB	1A0Z
ENSG00000244734	37.64	HBB	1A3N
ENSG00000244734	37.64	HBB	1A3O
ENSG00000244734	37.64	HBB	1ABW

Output options include CSV or other text files. In this example we get the Ensembl Gene ID, GC content, official HGNC symbol, and Protein Data Bank (PDB) links for a group of globin genes.

biomaRt
R package



Attributes: a vector specifying the output you request

```
> ens_att <- listAttributes(ensembl)
> ens_att[1:10,]
                     name
1         ensembl_gene_id
2   ensembl_transcript_id
3      ensembl_peptide_id
4         ensembl_exon_id
5             description
6         chromosome_name
7          start_position
8            end_position
9                  strand
10                   band
# currently the full list has 1,720
# attributes you can choose from!
```

getBM function:
--used to perform a query
--has four main arguments (attributes, filters, values, mart)
--returns a data.frame

```
> mydata = getBM(attributes=c("entrezgene","hgnc_symbol",
     "percentage_gc_content"), filters="entrezgene",
     values=myentrez, mart=ensembl)
```

Type this command in R or RStudio (first define myentrez as shown below)...

```
> mydata
  entrezgene hgnc_symbol percentage_gc_content
1       3043         HBB                 37.64
2       3045         HBD                 37.91
3       3046        HBE1                 38.96
4       3047        HBG1                 45.86
5       4151          MB                 50.32
```

...then type mydata to see the result. mydata is the data.frame returned by the getBM query, giving the requested results

Values refers to vector of values for the filters

```
> myentrez = c("3043","3045","
     3046","3047","4151")
> myentrez
[1] "3043" "3045" "3046" "3047" "4151"
```

Filters: a vector that defines a restriction on your query. To see your options:

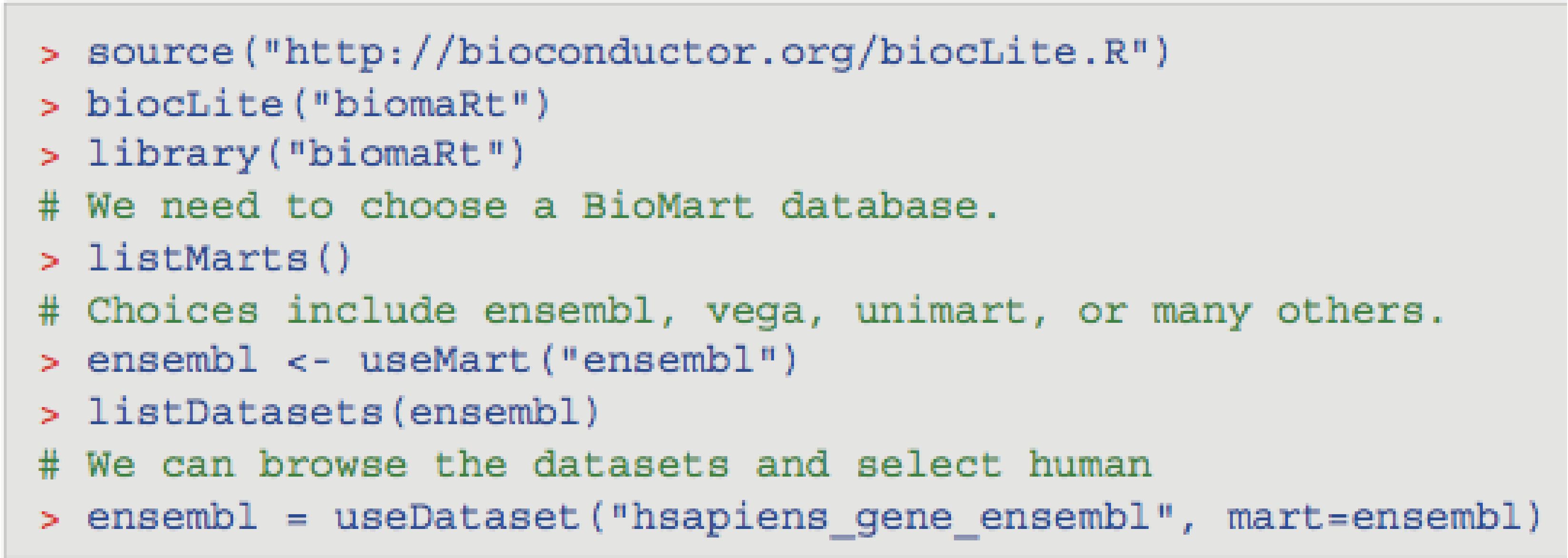
```
>filters = listFilters(ensembl)
>filters[1:10,]
1              chromosome_name
2                        start
3                          end
4                   band_start
5                     band_end
6                 marker_start
7                   marker_end
8                         type
9                encode_region
10                      strand
# Currently ~350 filters!
```

mart
--object of the class Mart
--to invoke (e.g. for mouse):

```
> UseMart
> mouse=useMart("ensembl",
dataset="mmusculus_gene_ensembl")
```

biomaRt R package example 1:

Given NCBI gene identifiers for five globins, what are the official (HGNC) gene symbols and the GC content?



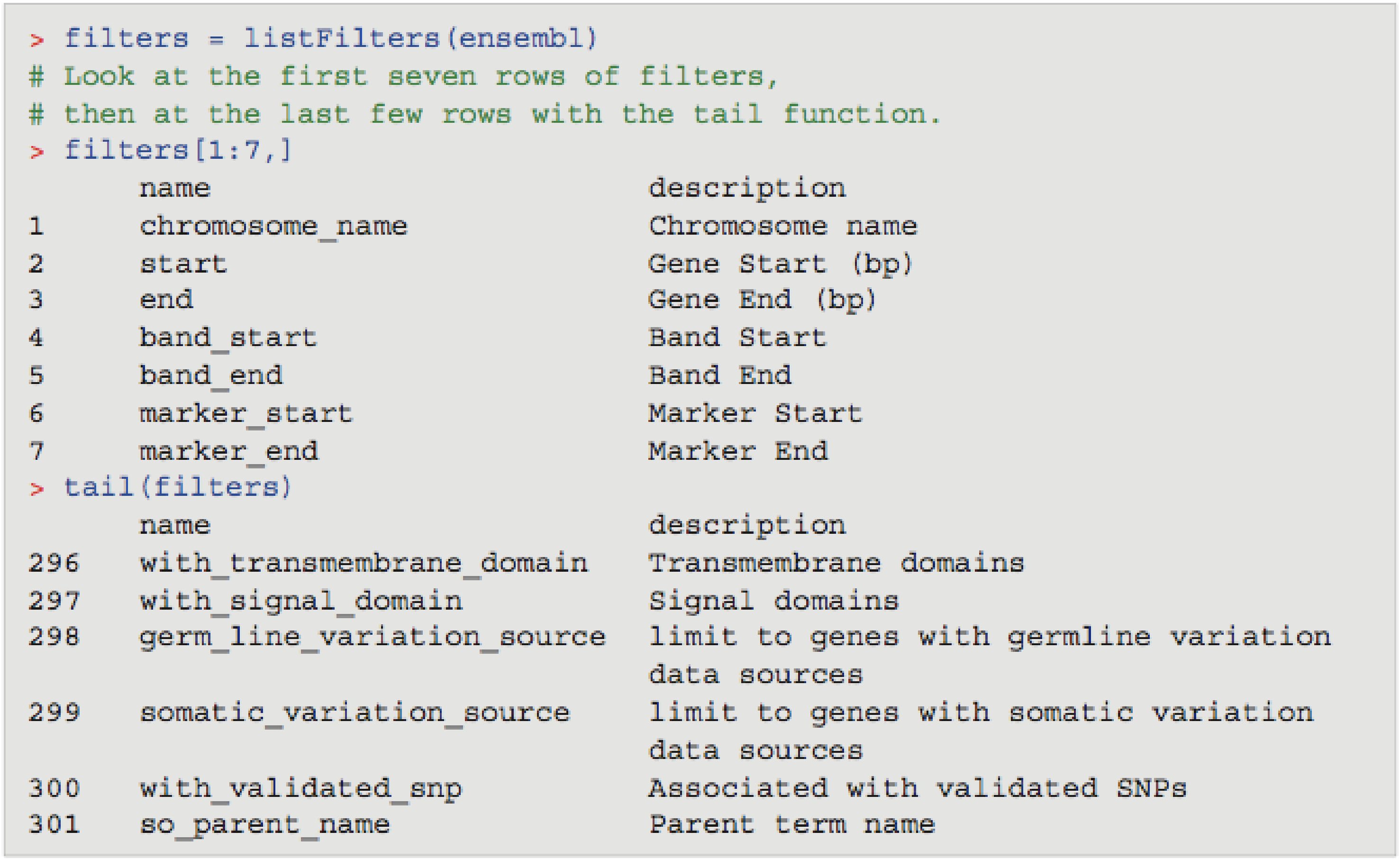
```
> source("http://bioconductor.org/biocLite.R")
> biocLite("biomaRt")
> library("biomaRt")
# We need to choose a BioMart database.
> listMarts()
# Choices include ensembl, vega, unimart, or many others.
> ensembl <- useMart("ensembl")
> listDatasets(ensembl)
# We can browse the datasets and select human
> ensembl = useDataset("hsapiens_gene_ensembl", mart=ensembl)
```

First obtain R and RStudio (both are freely available for PC or Mac).

Type these commands (in blue) to install biomaRt, load it, list the available "marts" (databases) and data sets. Comments are given in green.

biomaRt R package example 1:

Given NCBI gene identifiers for five globins, what are the official (HGNC) gene symbols and the GC content?

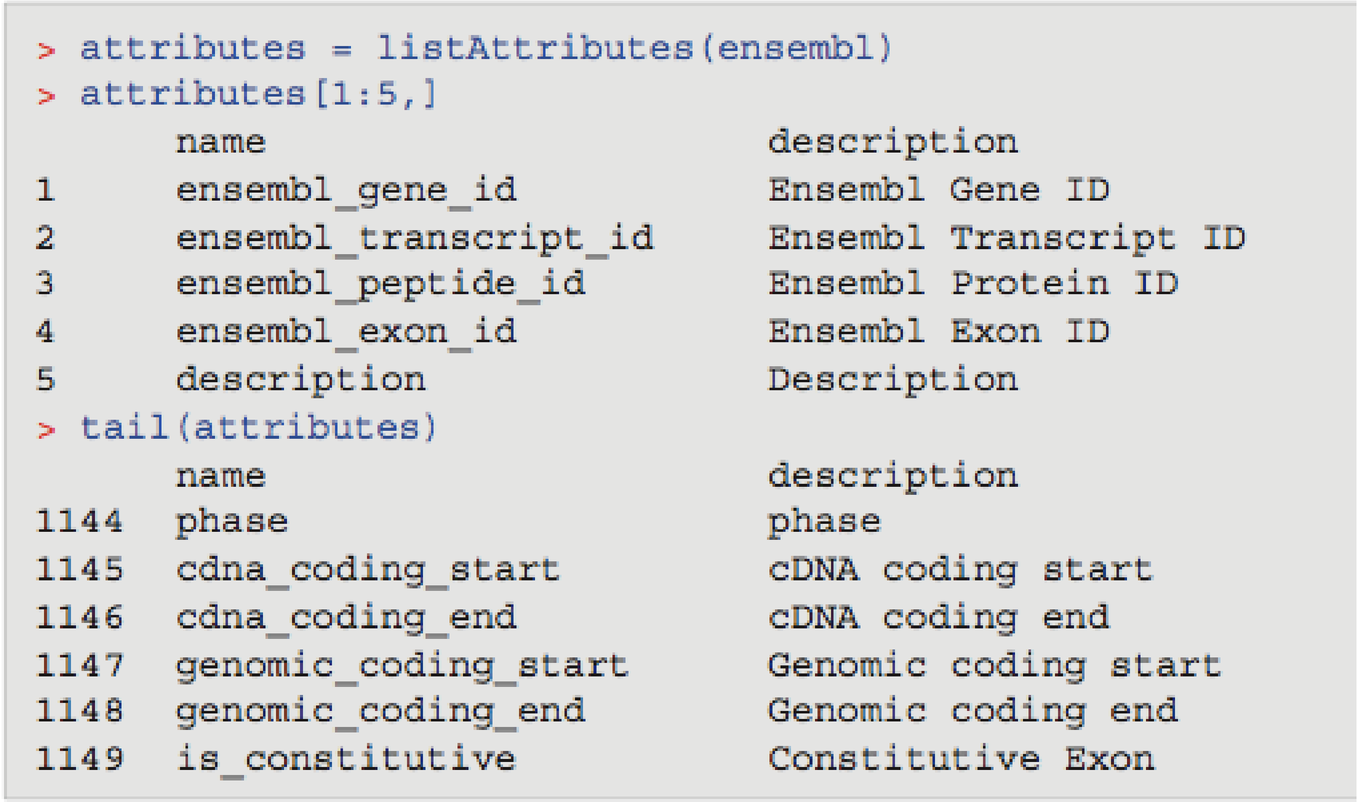


```
> filters = listFilters(ensembl)
# Look at the first seven rows of filters,
# then at the last few rows with the tail function.
> filters[1:7,]
      name                        description
1     chromosome_name             Chromosome name
2     start                       Gene Start (bp)
3     end                         Gene End (bp)
4     band_start                  Band Start
5     band_end                    Band End
6     marker_start                Marker Start
7     marker_end                  Marker End
> tail(filters)
      name                        description
296   with_transmembrane_domain   Transmembrane domains
297   with_signal_domain          Signal domains
298   germ_line_variation_source  limit to genes with germline variation
                                  data sources
299   somatic_variation_source    limit to genes with somatic variation
                                  data sources
300   with_validated_snp          Associated with validated SNPs
301   so_parent_name              Parent term name
```

Choose filters (vectors that restrict your query to features of interest).

biomaRt R package example 1:

Given NCBI gene identifiers for five globins, what are the official (HGNC) gene symbols and the GC content?

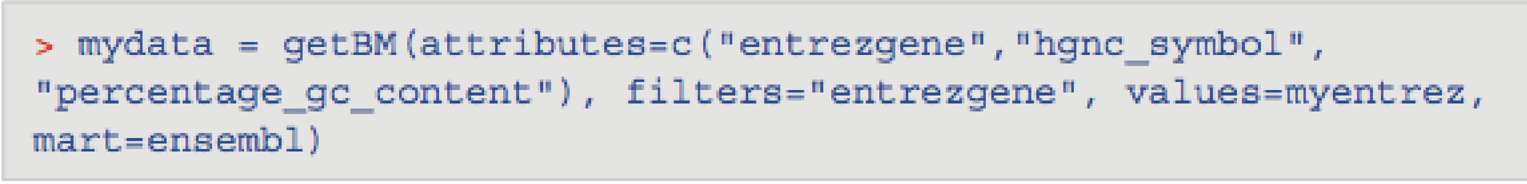


```
> attributes = listAttributes(ensembl)
> attributes[1:5,]
      name                      description
1     ensembl_gene_id           Ensembl Gene ID
2     ensembl_transcript_id     Ensembl Transcript ID
3     ensembl_peptide_id        Ensembl Protein ID
4     ensembl_exon_id           Ensembl Exon ID
5     description               Description
> tail(attributes)
      name                      description
1144  phase                     phase
1145  cdna_coding_start         cDNA coding start
1146  cdna_coding_end           cDNA coding end
1147  genomic_coding_start      Genomic coding start
1148  genomic_coding_end        Genomic coding end
1149  is_constitutive           Constitutive Exon
```

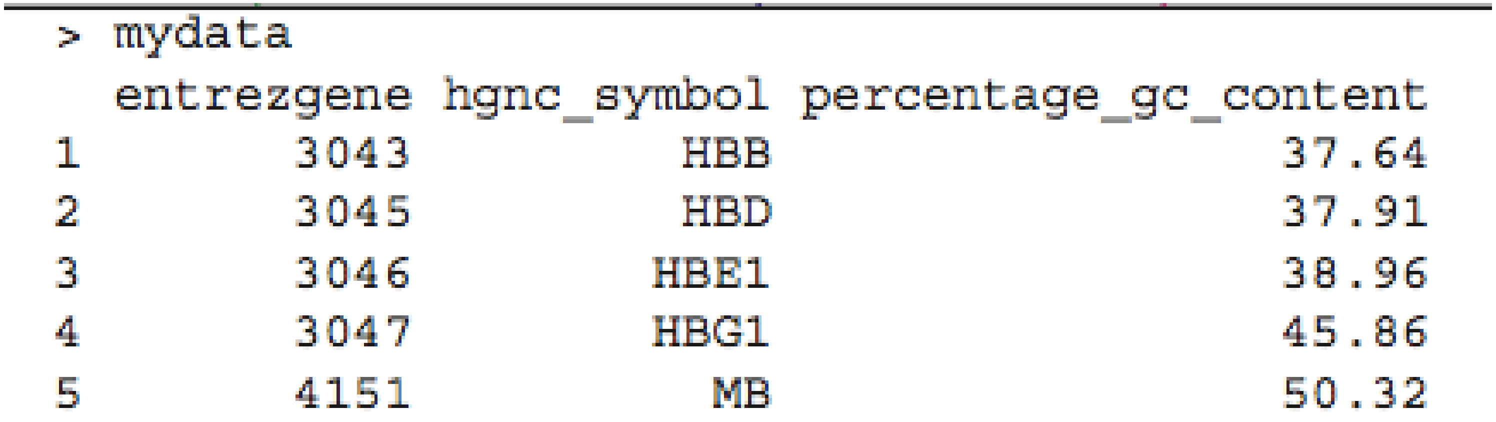
List attributes: specify the output you would like to obtain.

biomaRt R package example 1:

Given NCBI gene identifiers for five globins, what are the official (HGNC) gene symbols and the GC content?



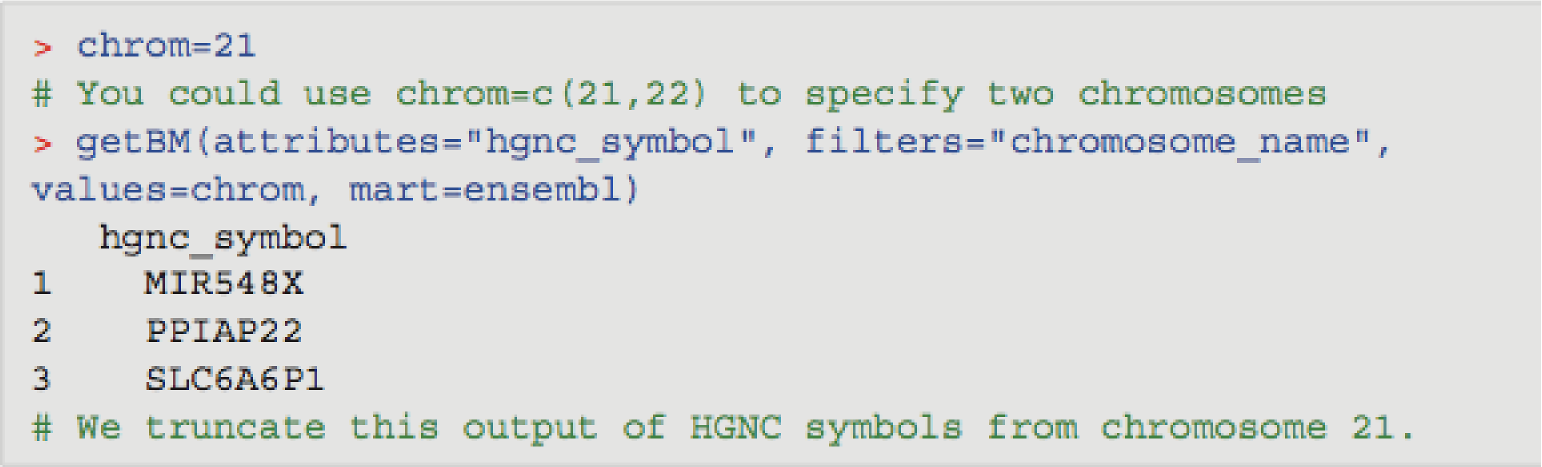
```
> mydata = getBM(attributes=c("entrezgene","hgnc_symbol",
"percentage_gc_content"), filters="entrezgene", values=myentrez,
mart=ensembl)
```



```
> mydata
  entrezgene hgnc_symbol percentage_gc_content
1       3043         HBB                 37.64
2       3045         HBD                 37.91
3       3046        HBE1                 38.96
4       3047        HBG1                 45.86
5       4151          MB                 50.32
```

`biomaRt` R package example 2:

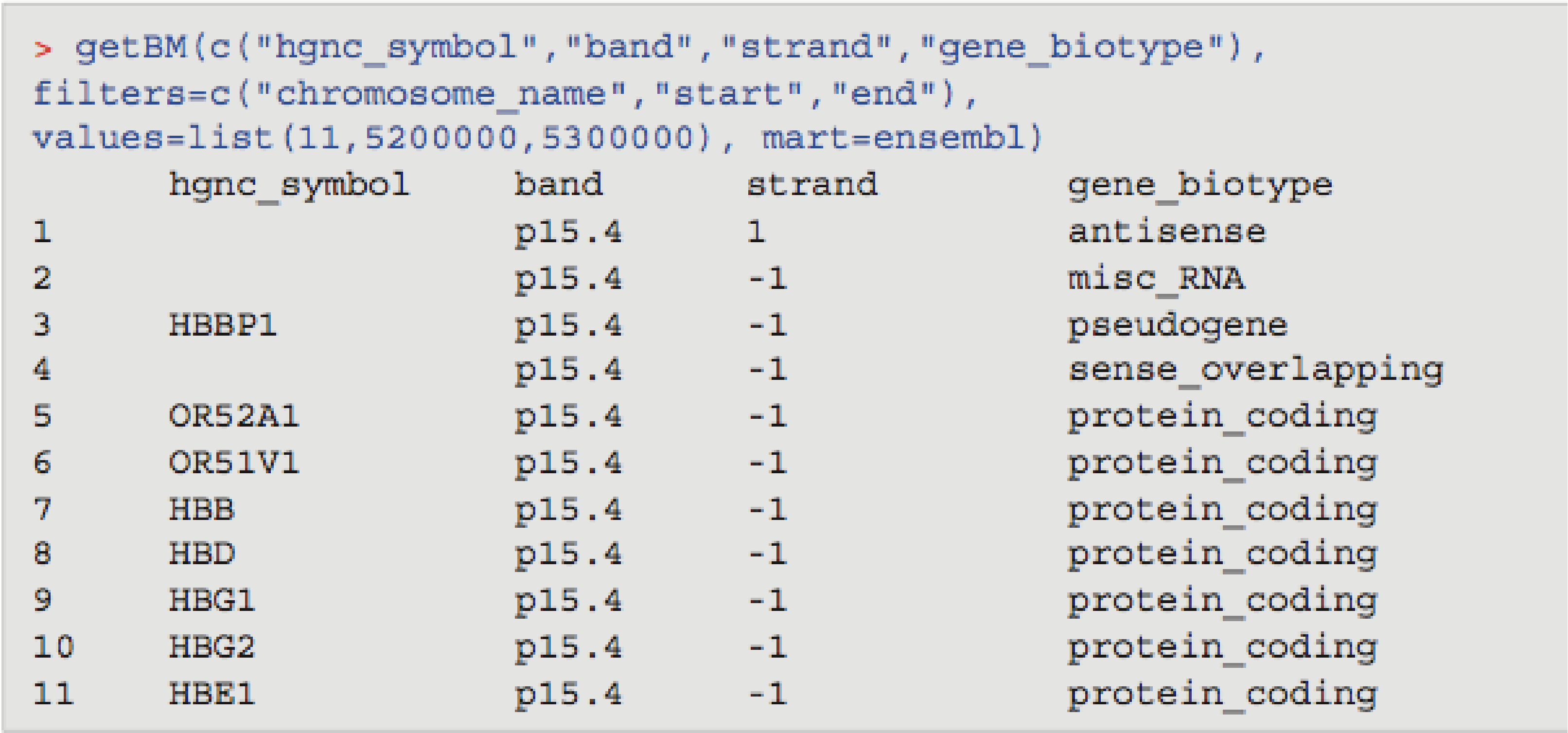
What are the HGNC gene symbols for genes on human chromosome 21?



```
> chrom=21
# You could use chrom=c(21,22) to specify two chromosomes
> getBM(attributes="hgnc_symbol", filters="chromosome_name",
values=chrom, mart=ensembl)
   hgnc_symbol
1     MIR548X
2     PPIAP22
3     SLC6A6P1
# We truncate this output of HGNC symbols from chromosome 21.
```

biomaRt R package example 3:

What Ensembl genes are in a 100,000 base pair region of chromosome 11 surrounding *HBB*? What chromosome band are they on, what strand, and what type of genes are they?

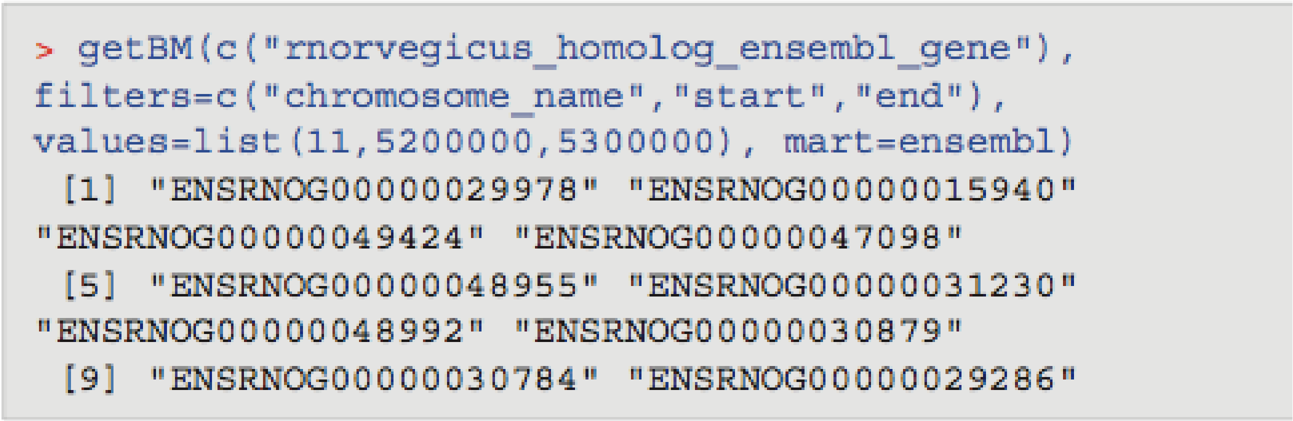


```
> getBM(c("hgnc_symbol","band","strand","gene_biotype"),
filters=c("chromosome_name","start","end"),
values=list(11,5200000,5300000), mart=ensembl)
     hgnc_symbol     band      strand        gene_biotype
1                    p15.4     1             antisense
2                    p15.4     -1            misc_RNA
3    HBBP1           p15.4     -1            pseudogene
4                    p15.4     -1            sense_overlapping
5    OR52A1          p15.4     -1            protein_coding
6    OR51V1          p15.4     -1            protein_coding
7    HBB             p15.4     -1            protein_coding
8    HBD             p15.4     -1            protein_coding
9    HBG1            p15.4     -1            protein_coding
10   HBG2            p15.4     -1            protein_coding
11   HBE1            p15.4     -1            protein_coding
```

Note that we can expand the attributes (e.g., adding "start_position", "end_ position" after "band") for more information.

`biomaRt` R package example 4:

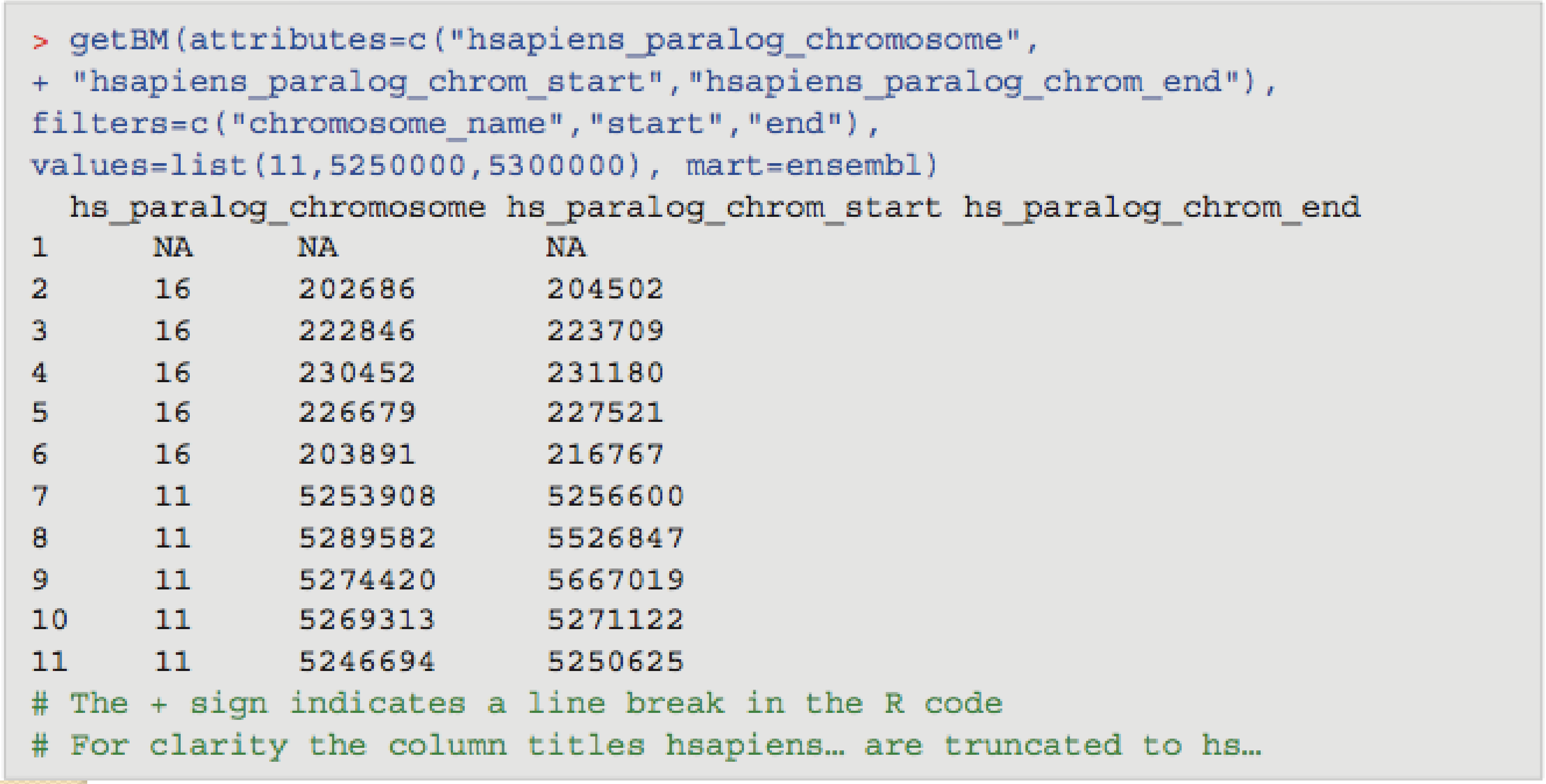
What are the rat homologs of the genes in a 100 kilobase region of human chromosome 11?



```
> getBM(c("rnorvegicus_homolog_ensembl_gene"),
filters=c("chromosome_name","start","end"),
values=list(11,5200000,5300000), mart=ensembl)
 [1] "ENSRNOG00000029978" "ENSRNOG00000015940"
"ENSRNOG00000049424" "ENSRNOG00000047098"
 [5] "ENSRNOG00000048955" "ENSRNOG00000031230"
"ENSRNOG00000048992" "ENSRNOG00000030879"
 [9] "ENSRNOG00000030784" "ENSRNOG00000029286"
```

biomaRt R package example 5:

What are the paralogs of the genes in a 50 kb region of human chromosome 11?



```
> getBM(attributes=c("hsapiens_paralog_chromosome",
+ "hsapiens_paralog_chrom_start","hsapiens_paralog_chrom_end"),
filters=c("chromosome_name","start","end"),
values=list(11,5250000,5300000), mart=ensembl)
  hs_paralog_chromosome hs_paralog_chrom_start hs_paralog_chrom_end
1      NA       NA              NA
2      16       202686          204502
3      16       222846          223709
4      16       230452          231180
5      16       226679          227521
6      16       203891          216767
7      11       5253908         5256600
8      11       5289582         5526847
9      11       5274420         5667019
10     11       5269313         5271122
11     11       5246694         5250625
# The + sign indicates a line break in the R code
# For clarity the column titles hsapiens… are truncated to hs…
```

Since this region includes beta globin genes, we expect the result to include alpha globin gene loci on chromosome 16.

Outline

Introduction

General features of eukaryotic genomes and chromosomes

C value paradox; organization; genome browsers

Analysis of chromosomes using BioMart and biomaRt

ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

Variation in chromosomal DNA

Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

The ENCODE project

- The ENCyclopedia Of DNA Elements (ENCODE) project was launched in 2003

- Pilot phase (completed): devise and test high-throughput approaches to identify functional elements.

- Second phase: technology development.

- Third phase: production. Expand the ENCODE project to analyze the remaining 99 percent of the human genome.

The ENCODE project

Scope of ENCODE: build a list of all sequence-based functional elements in human DNA. This includes:

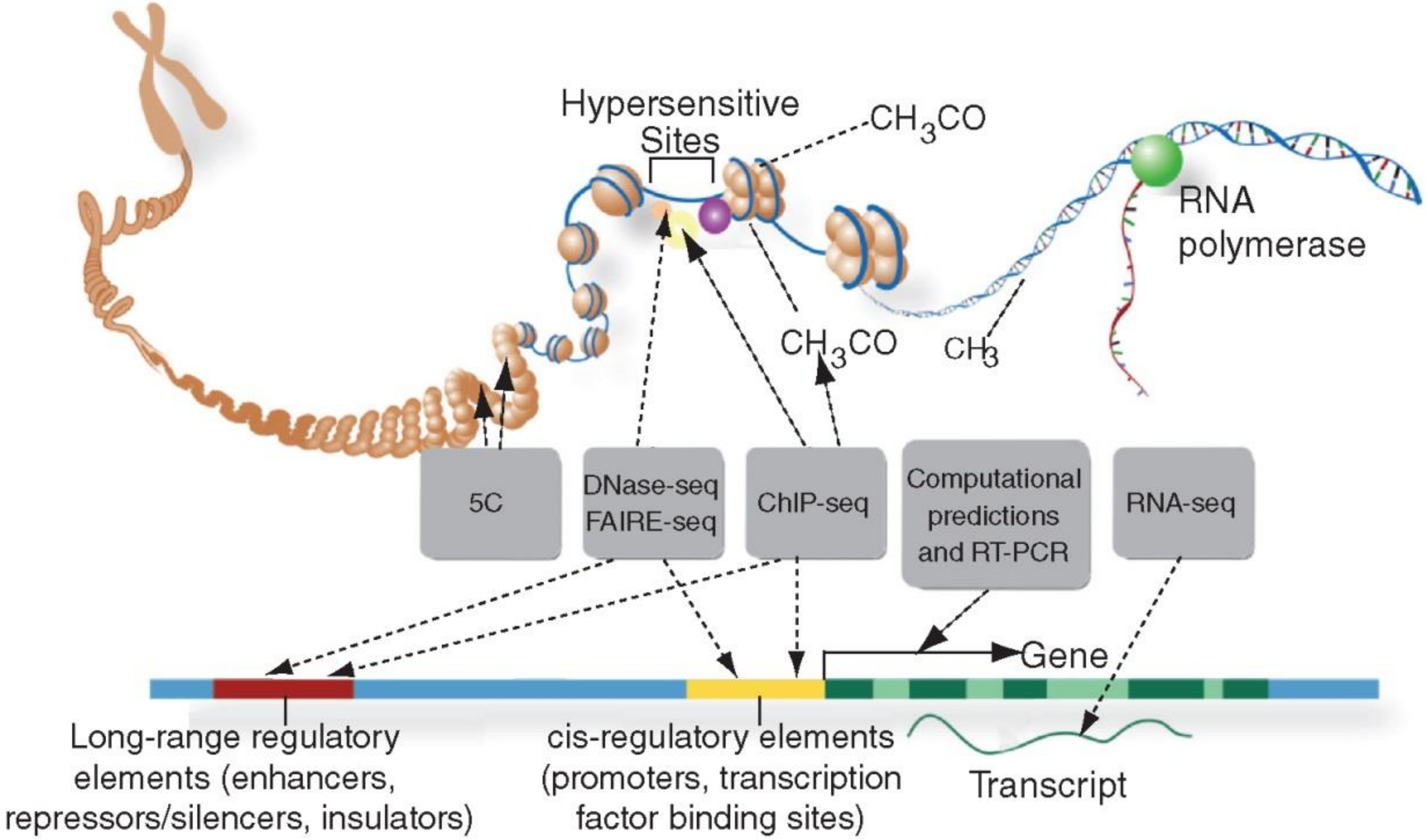
‣ protein-coding genes

‣ non-protein-coding genes

‣ regulatory elements involved in the control of gene transcription

‣ DNA sequences that mediate chromosomal structure and dynamics.

The ENCODE Project catalog of functional elements


Hypersensitive Sites
CH_3CO
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq
FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

ENCODE has catalogued functional elements in human, mouse, *Drosophila*, and a nematode.

Conclusions of the ENCODE project

- The human genome is pervasively transcribed.
- 80.4% of the human genome is functionally active.
- Many noncoding transcripts were identified.
- Novel transcriptional start sites were identified and characterized in detail.
- Histone modification and chromatin accessibility predict the presence and activity of transcription start sites.
- Of the 80.4% of the genome spanned by elements defined by ENCODE as functional, if we exclude RNA elements and histone elements, 44.2% of the genome is covered.

Critiques of the ENCODE project

(1) DNA may have biochemical activity (as described by the ENCODE project) without having function in an evolutionary sense.

(2) Suppose the ENCODE project were extended to a set of compact genomes (e.g., *Takifugu rubipres*; 400 Mb) and large genomes (e.g., a lungfish). There are two possible outcomes. First, functional elements could be constant in number, regardless of *C* value. The density of functional elements per kilobase would be dramatically smaller in such large genomes. A second outcome is that functional elements as defined by ENCODE increase in proportion to *C* value (independent of organismal complexity). Would lungfish having 300-fold larger genome size and 300-fold more functional elements then be expected to display more organismal complexity than related *Takifugu* having compact genomes?

Outline

Introduction

General features of eukaryotic genomes and chromosomes

- *C* value paradox; organization; genome browsers
- Analysis of chromosomes using BioMart and biomaRt
- ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

- Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

- Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

- Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

Variation in chromosomal DNA

- Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

Repetitive DNA in eukaryotes

Bacterial genomes are usually compact, with ~1 gene per kilobase and relatively small intergenic regions.

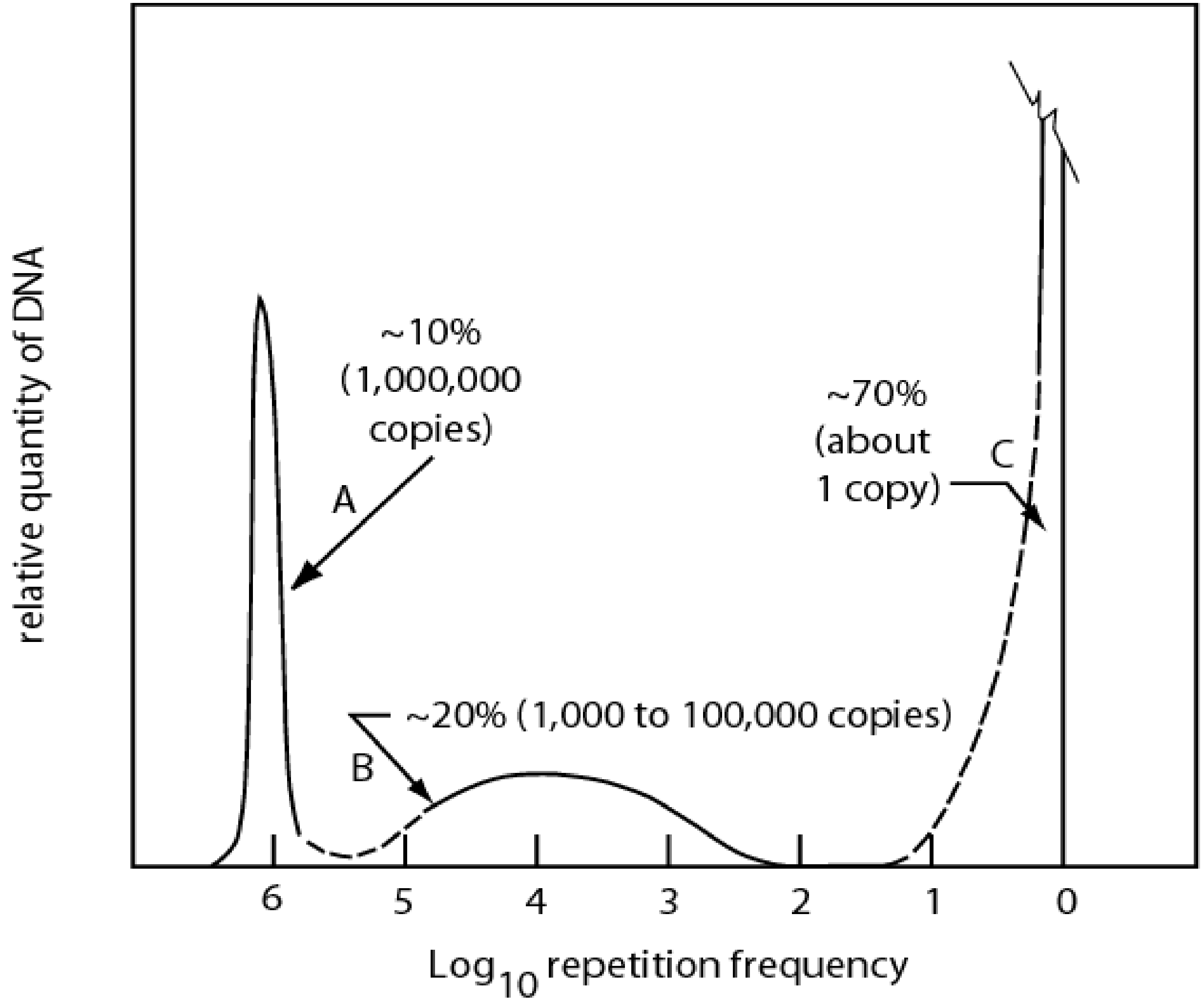
Eukaryotic genes have large intergenic and intronic regions.

Britten & Kohne's analysis of repetitive DNA

In the 1960s, Britten and Kohne defined the repetitive nature of genomic DNA in a variety of organisms. They isolated genomic DNA, sheared it, dissociated the DNA strands, and measured the rates of DNA reassociation.

For dozens of eukaryotes—but not bacteria or viruses—large amount of DNA reassociates extremely rapidly. This represents repetitive DNA.

Britten and Kohne (1968) identified repetitive DNA classes

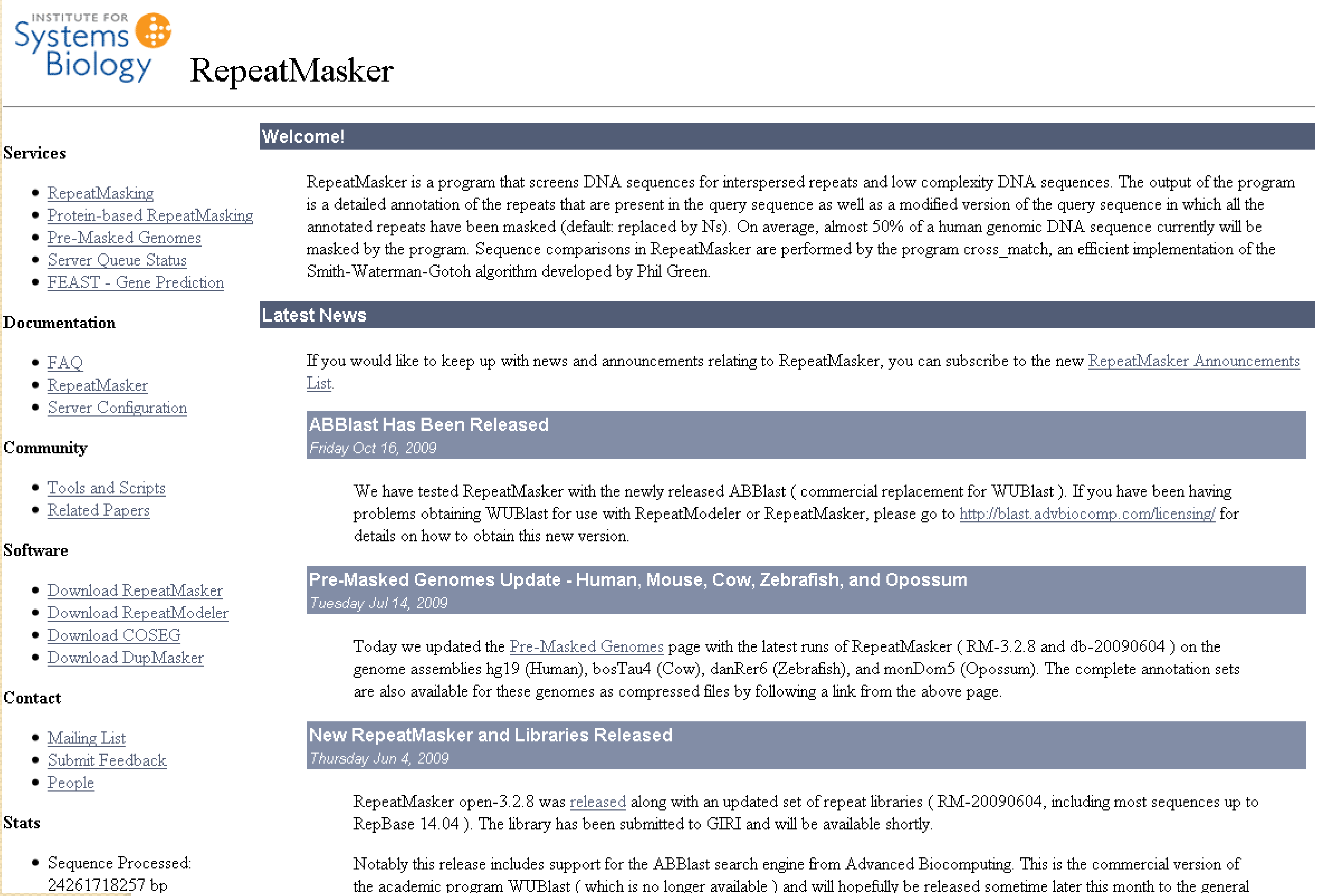


B&FG 3e
Fig. 8.6
Page 324

Software to detect repetitive DNA

It is essential to identify repetitive DNA in eukaryotic genomes. RepBase Update is a database of known repeats and low-complexity regions.

RepeatMasker is a program that searches DNA queries against RepBase. There are many RepeatMasker sites available on-line.



INSTITUTE FOR Systems Biology

RepeatMasker

Services

- RepeatMasking
- Protein-based RepeatMasking
- Pre-Masked Genomes
- Server Queue Status
- FEAST - Gene Prediction

Documentation

- FAQ
- RepeatMasker
- Server Configuration

Community

- Tools and Scripts
- Related Papers

Software

- Download RepeatMasker
- Download RepeatModeler
- Download COSEG
- Download DupMasker

Contact

- Mailing List
- Submit Feedback
- People

Stats

- Sequence Processed: 24261718257 bp

Welcome!

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). On average, almost 50% of a human genomic DNA sequence currently will be masked by the program. Sequence comparisons in RepeatMasker are performed by the program cross_match, an efficient implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green.

Latest News

If you would like to keep up with news and announcements relating to RepeatMasker, you can subscribe to the new RepeatMasker Announcements List.

ABBlast Has Been Released

Friday Oct 16, 2009

We have tested RepeatMasker with the newly released ABBlast (commercial replacement for WUBlast). If you have been having problems obtaining WUBlast for use with RepeatModeler or RepeatMasker, please go to http://blast.advbiocomp.com/licensing/ for details on how to obtain this new version.

Pre-Masked Genomes Update - Human, Mouse, Cow, Zebrafish, and Opossum

Tuesday Jul 14, 2009

Today we updated the Pre-Masked Genomes page with the latest runs of RepeatMasker (RM-3.2.8 and db-20090604) on the genome assemblies hg19 (Human), bosTau4 (Cow), danRer6 (Zebrafish), and monDom5 (Opossum). The complete annotation sets are also available for these genomes as compressed files by following a link from the above page.

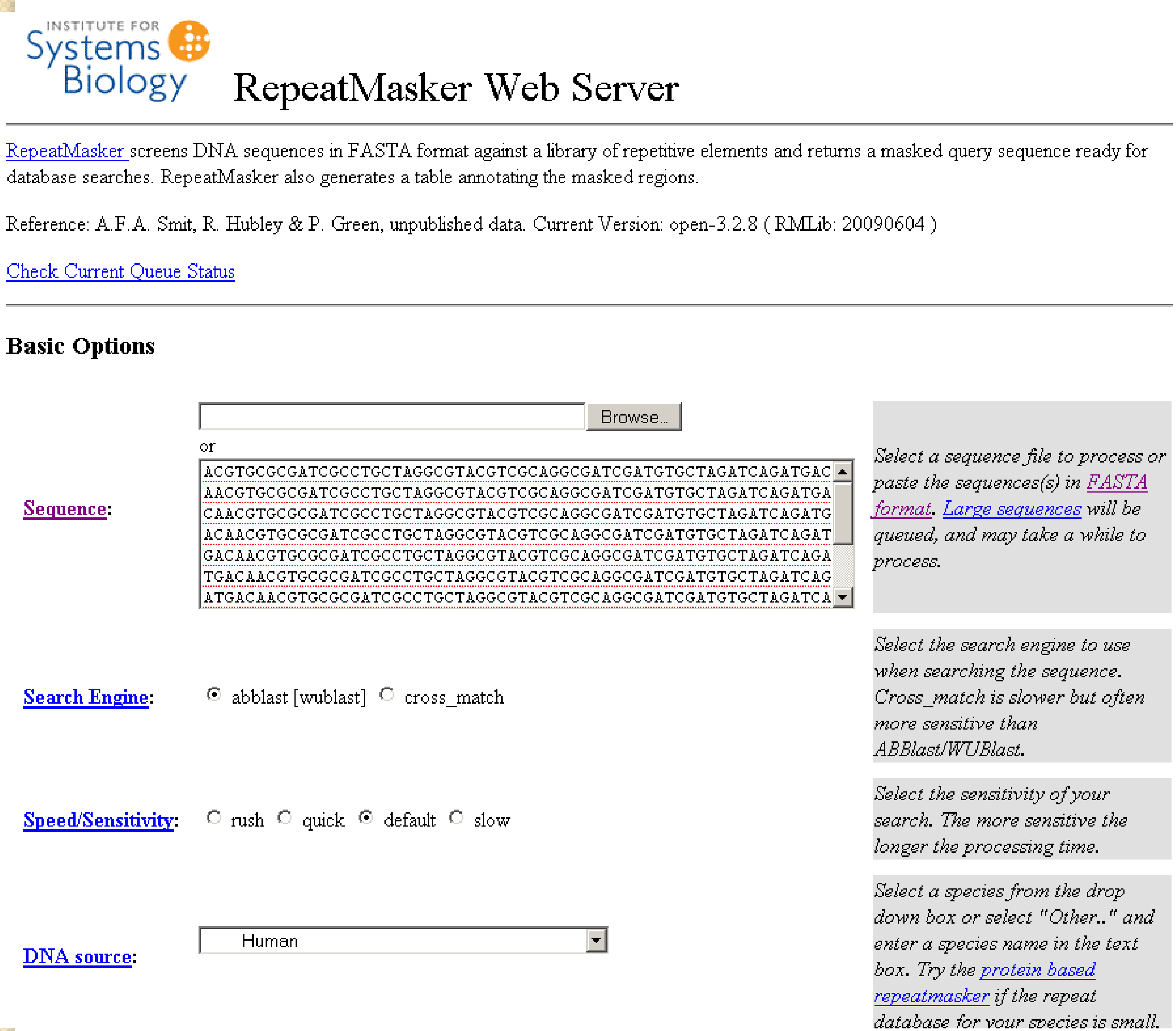
New RepeatMasker and Libraries Released

Thursday Jun 4, 2009

RepeatMasker open-3.2.8 was released along with an updated set of repeat libraries (RM-20090604, including most sequences up to RepBase 14.04). The library has been submitted to GIRI and will be available shortly.

Notably this release includes support for the ABBlast search engine from Advanced Biocomputing. This is the commercial version of the academic program WUBlast (which is no longer available) and will hopefully be released sometime later this month to the general

http://www.repeatmasker.org/



RepeatMasker Web Server

RepeatMasker screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for database searches. RepeatMasker also generates a table annotating the masked regions.

Reference: A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-3.2.8 (RMLib: 20090604)

Check Current Queue Status

Basic Options

Sequence: [ ] Browse...

or

```
ACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCAGATGAC
AACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCAGATGA
CAACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCAGATG
ACAACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCAGAT
GACAACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCAGA
TGACAACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCAG
ATGACAACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAGGCGATCGATGTGCTAGATCA
```

Select a sequence file to process or paste the sequences(s) in FASTA format. Large sequences will be queued, and may take a while to process.

Search Engine: ● abblast [wublast] ○ cross_match

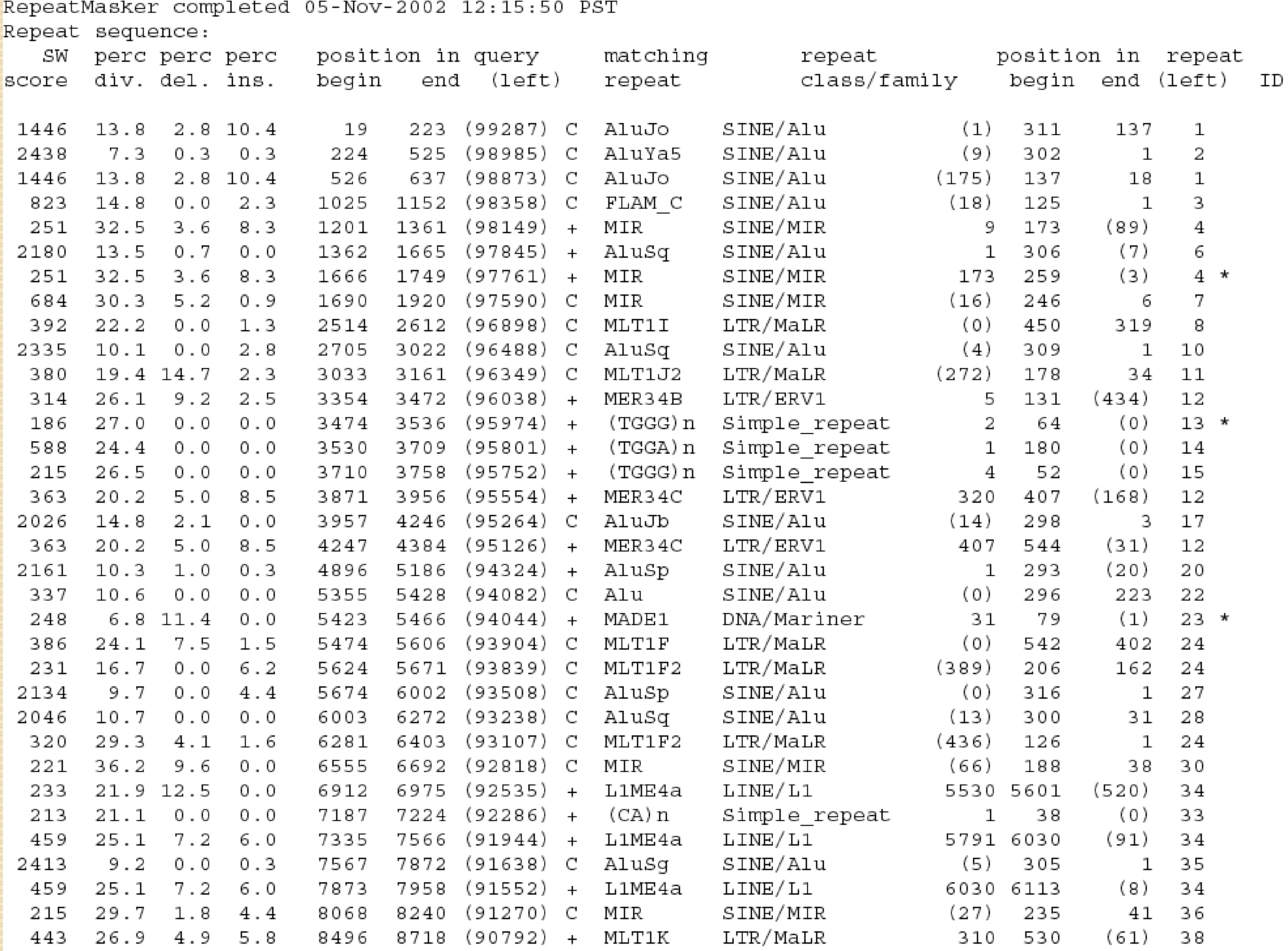
Select the search engine to use when searching the sequence. Cross_match is slower but often more sensitive than ABBlast/WUBlast.

Speed/Sensitivity: ○ rush ○ quick ● default ○ slow

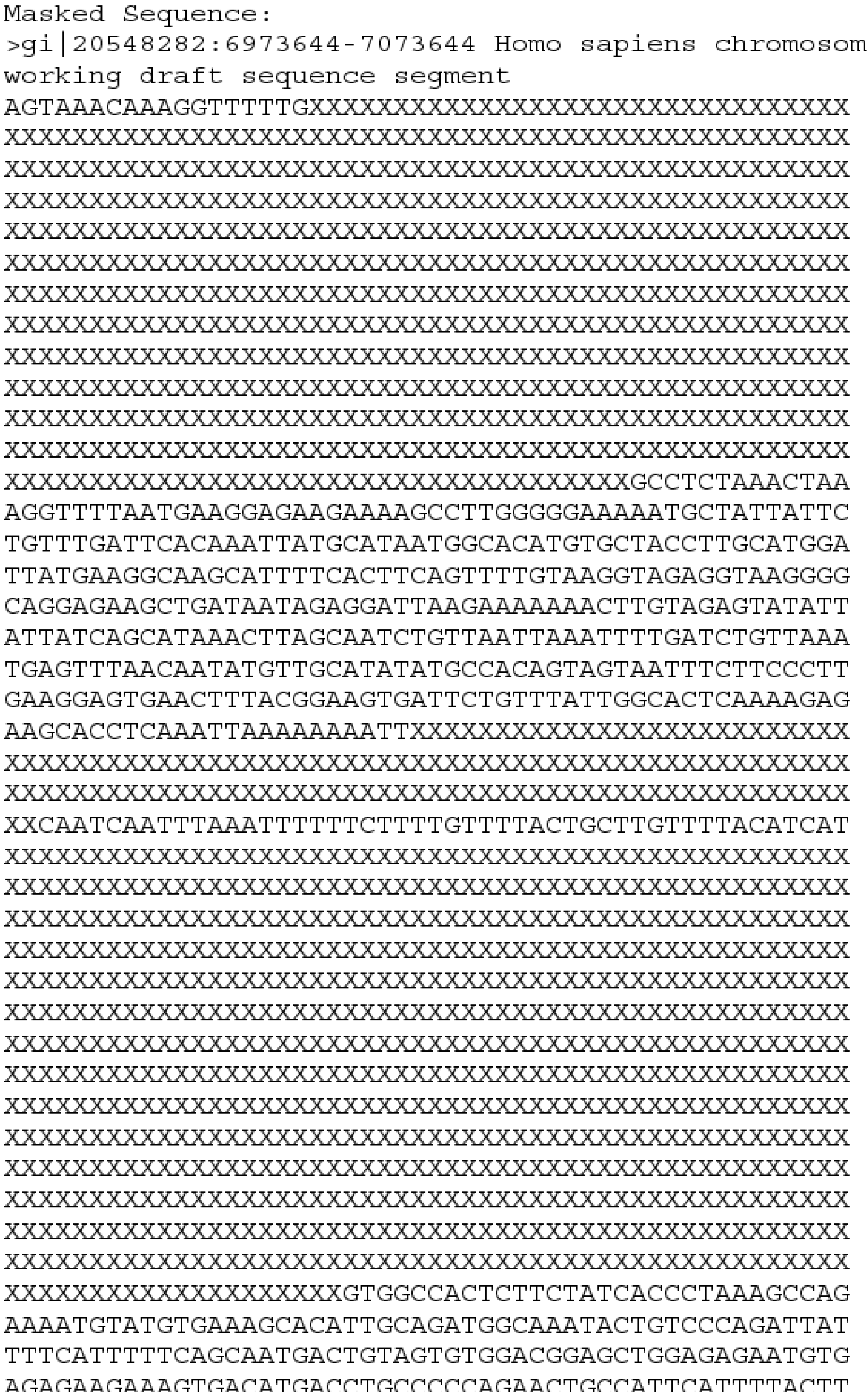
Select the sensitivity of your search. The more sensitive the longer the processing time.

DNA source: Human

Select a species from the drop down box or select "Other.." and enter a species name in the text box. Try the protein based repeatmasker if the repeat database for your species is small.



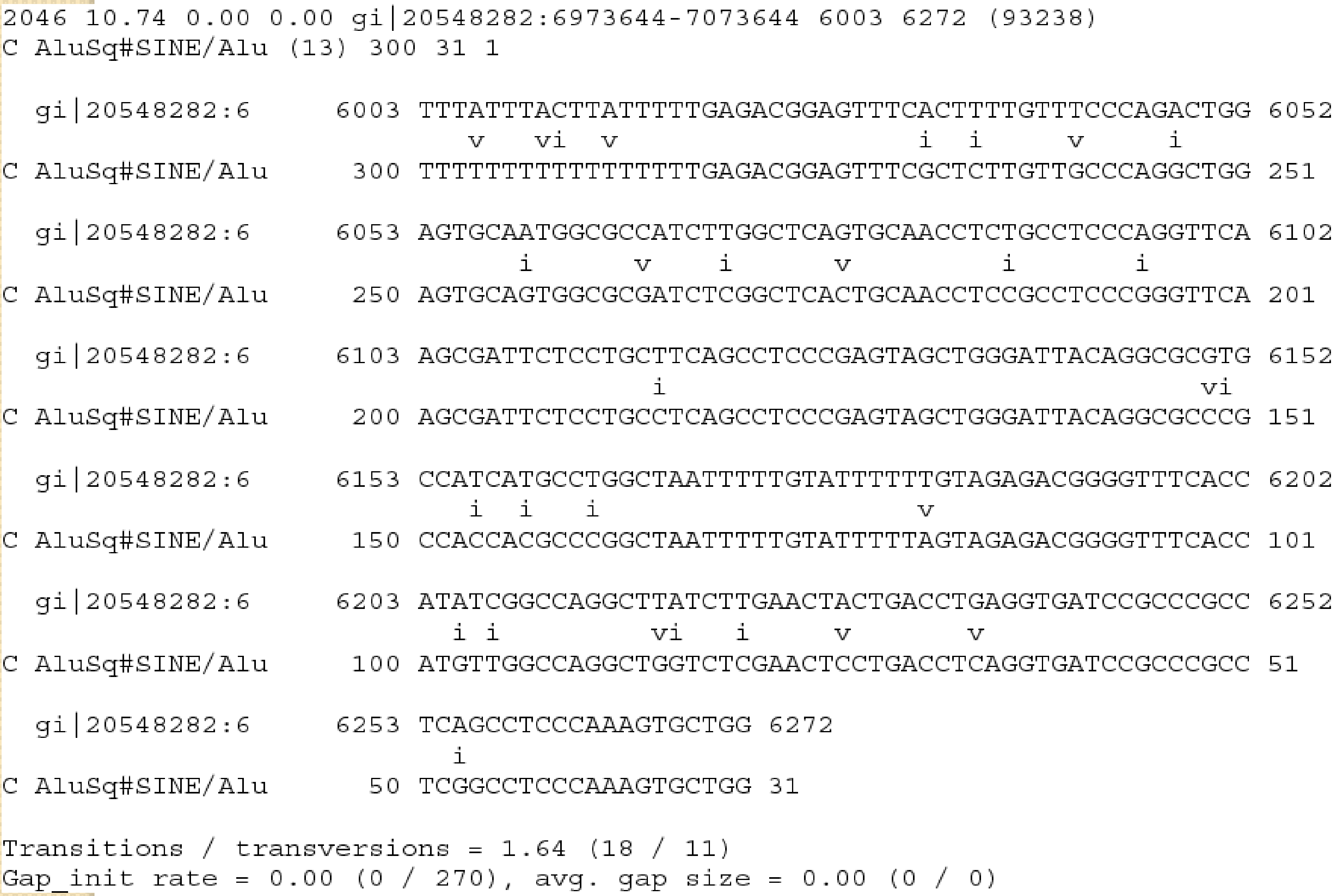
```
RepeatMasker completed 05-Nov-2002 12:15:50 PST
Repeat sequence:
   SW  perc perc perc  position in query     matching      repeat          position in  repeat
score  div. del. ins.  begin   end  (left)   repeat        class/family     begin  end (left)   ID

 1446  13.8  2.8 10.4     19    223 (99287) C  AluJo    SINE/Alu           (1)  311    137   1
 2438   7.3  0.3  0.3    224    525 (98985) C  AluYa5   SINE/Alu           (9)  302      1   2
 1446  13.8  2.8 10.4    526    637 (98873) C  AluJo    SINE/Alu         (175)  137     18   1
  823  14.8  0.0  2.3   1025   1152 (98358) C  FLAM_C   SINE/Alu          (18)  125      1   3
  251  32.5  3.6  8.3   1201   1361 (98149) +  MIR      SINE/MIR             9  173   (89)   4
 2180  13.5  0.7  0.0   1362   1665 (97845) +  AluSq    SINE/Alu             1  306    (7)   6
  251  32.5  3.6  8.3   1666   1749 (97761) +  MIR      SINE/MIR           173  259    (3)   4 *
  684  30.3  5.2  0.9   1690   1920 (97590) C  MIR      SINE/MIR          (16)  246      6   7
  392  22.2  0.0  1.3   2514   2612 (96898) C  MLT1I    LTR/MaLR           (0)  450    319   8
 2335  10.1  0.0  2.8   2705   3022 (96488) C  AluSq    SINE/Alu           (4)  309      1  10
  380  19.4 14.7  2.3   3033   3161 (96349) C  MLT1J2   LTR/MaLR         (272)  178     34  11
  314  26.1  9.2  2.5   3354   3472 (96038) +  MER34B   LTR/ERV1             5  131  (434)  12
  186  27.0  0.0  0.0   3474   3536 (95974) +  (TGGG)n  Simple_repeat        2   64    (0)  13 *
  588  24.4  0.0  0.0   3530   3709 (95801) +  (TGGA)n  Simple_repeat        1  180    (0)  14
  215  26.5  0.0  0.0   3710   3758 (95752) +  (TGGG)n  Simple_repeat        4   52    (0)  15
  363  20.2  5.0  8.5   3871   3956 (95554) +  MER34C   LTR/ERV1           320  407  (168)  12
 2026  14.8  2.1  0.0   3957   4246 (95264) C  AluJb    SINE/Alu          (14)  298      3  17
  363  20.2  5.0  8.5   4247   4384 (95126) +  MER34C   LTR/ERV1           407  544   (31)  12
 2161  10.3  1.0  0.3   4896   5186 (94324) +  AluSp    SINE/Alu             1  293   (20)  20
  337  10.6  0.0  0.0   5355   5428 (94082) C  Alu      SINE/Alu           (0)  296    223  22
  248   6.8 11.4  0.0   5423   5466 (94044) +  MADE1    DNA/Mariner         31   79    (1)  23 *
  386  24.1  7.5  1.5   5474   5606 (93904) C  MLT1F    LTR/MaLR           (0)  542    402  24
  231  16.7  0.0  6.2   5624   5671 (93839) C  MLT1F2   LTR/MaLR         (389)  206    162  24
 2134   9.7  0.0  4.4   5674   6002 (93508) C  AluSp    SINE/Alu           (0)  316      1  27
 2046  10.7  0.0  0.0   6003   6272 (93238) C  AluSq    SINE/Alu          (13)  300     31  28
  320  29.3  4.1  1.6   6281   6403 (93107) C  MLT1F2   LTR/MaLR         (436)  126      1  24
  221  36.2  9.6  0.0   6555   6692 (92818) C  MIR      SINE/MIR          (66)  188     38  30
  233  21.9 12.5  0.0   6912   6975 (92535) +  L1ME4a   LINE/L1           5530 5601  (520)  34
  213  21.1  0.0  0.0   7187   7224 (92286) +  (CA)n    Simple_repeat        1   38    (0)  33
  459  25.1  7.2  6.0   7335   7566 (91944) +  L1ME4a   LINE/L1           5791 6030   (91)  34
 2413   9.2  0.0  0.3   7567   7872 (91638) C  AluSg    SINE/Alu           (5)  305      1  35
  459  25.1  7.2  6.0   7873   7958 (91552) +  L1ME4a   LINE/L1           6030 6113    (8)  34
  215  29.7  1.8  4.4   8068   8240 (91270) C  MIR      SINE/MIR          (27)  235     41  36
  443  26.9  4.9  5.8   8496   8718 (90792) +  MLT1K    LTR/MaLR           310  530   (61)  38
```



```
Masked Sequence:
>gi|20548282:6973644-7073644 Homo sapiens chromosome 10
working draft sequence segment
AGTAAACAAAGGTTTTTGXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGCCTCTAAACTAA
AGGTTTTAATGAAGGAGAAGAAAAGCCTTGGGGGAAAAATGCTATTATTC
TGTTTGATTCACAAATTATGCATAATGGCACATGTGCTACCTTGCATGGA
TTATGAAGGCAAGCATTTTCACTTCAGTTTTGTAAGGTAGAGGTAAGGGG
CAGGAGAAGCTGATAATAGAGGATTAAGAAAAAAACTTGTAGAGTATATT
ATTATCAGCATAAACTTAGCAATCTGTTAATTAAATTTTGATCTGTTAAA
TGAGTTTAACAATATGTTGCATATATGCCACAGTAGTAATTTCTTCCCTT
GAAGGAGTGAACTTTACGGAAGTGATTCTGTTTATTGGCACTCAAAAGAG
AAGCACCTCAAATTAAAAAAAATTXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXCAATCAATTTAAATTTTTTCTTTTGTTTTACTGCTTGTTTTACATCAT
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXGTGGCCACTCTTCTATCACCCTAAAGCCAG
AAAATGTATGTGAAAGCACATTGCAGATGGCAAATACTGTCCCAGATTAT
TTTCATTTTTCAGCAATGACTGTAGTGTGGACGGAGCTGGAGAGAATGTG
AGAGAAGAAAGTGACATGACCTGCCCCCAGAACTGCCATTCATTTTACTT
```

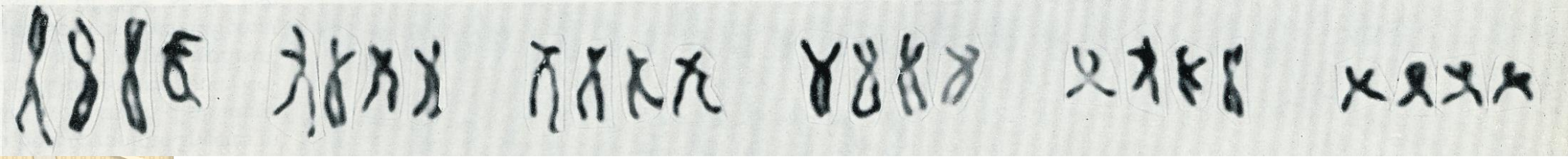
RepeatMasker masks repetitive DNA (FASTA format)

RepeatMasker identifies *Alu* repeats



```
2046 10.74 0.00 0.00 gi|20548282:6973644-7073644 6003 6272 (93238)
C AluSq#SINE/Alu (13) 300 31 1

  gi|20548282:6     6003 TTTATTTACTTATTTTTGAGACGGAGTTTCACTTTTGTTTCCCAGACTGG 6052
                            v   vi  v                  i  i     v     i
C AluSq#SINE/Alu     300 TTTTTTTTTTTTTTTTTGAGACGGAGTTTCGCTCTTGTTGCCCAGGCTGG 251

  gi|20548282:6     6053 AGTGCAATGGCGCCATCTTGGCTCAGTGCAACCTCTGCCTCCCAGGTTCA 6102
                               i      v    i      v         i       i
C AluSq#SINE/Alu     250 AGTGCAGTGGCGCGATCTCGGCTCACTGCAACCTCCGCCTCCCGGGTTCA 201

  gi|20548282:6     6103 AGCGATTCTCCTGCTTCAGCCTCCCGAGTAGCTGGGATTACAGGCGCGTG 6152
                                       i                                vi
C AluSq#SINE/Alu     200 AGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCGCCCG 151

  gi|20548282:6     6153 CCATCATGCCTGGCTAATTTTTGTATTTTTTGTAGAGACGGGGTTTCACC 6202
                            i  i   i                   v
C AluSq#SINE/Alu     150 CCACCACGCCCGGCTAATTTTTGTATTTTTAGTAGAGACGGGGTTTCACC 101

  gi|20548282:6     6203 ATATCGGCCAGGCTTATCTTGAACTACTGACCTGAGGTGATCCGCCCGCC 6252
                           i i         vi   i     v       v
C AluSq#SINE/Alu     100 ATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCCGCCCGCC 51

  gi|20548282:6     6253 TCAGCCTCCCAAAGTGCTGG 6272
                           i
C AluSq#SINE/Alu      50 TCGGCCTCCCAAAGTGCTGG 31

Transitions / transversions = 1.64 (18 / 11)
Gap_init rate = 0.00 (0 / 270), avg. gap size = 0.00 (0 / 0)
```

Repetitive DNA content of eukaryotic genomes



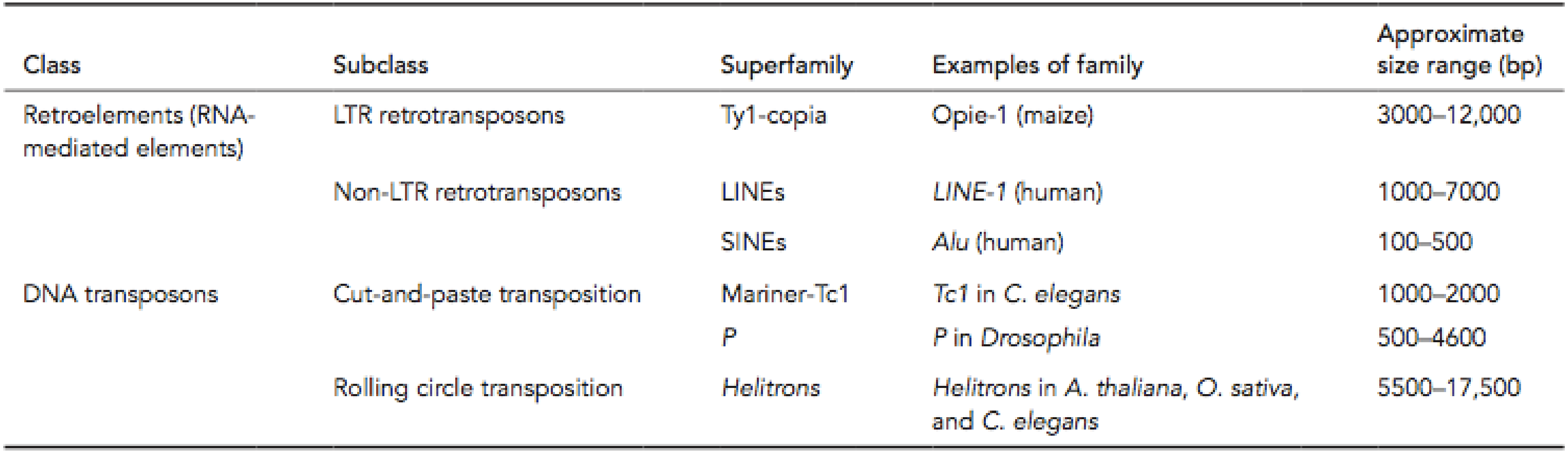
1. Interspersed repeats
2. Processed pseudogenes
3. Simple sequence repeats
4. Segmental duplications
5. Blocks of tandem repeats

Five main classes of repetitive DNA

1. Interspersed repeats (transposon-derived repeats) constitute ~45% of the human genome. They involve RNA intermediates (retroelements) or DNA intermediates (DNA transposons).

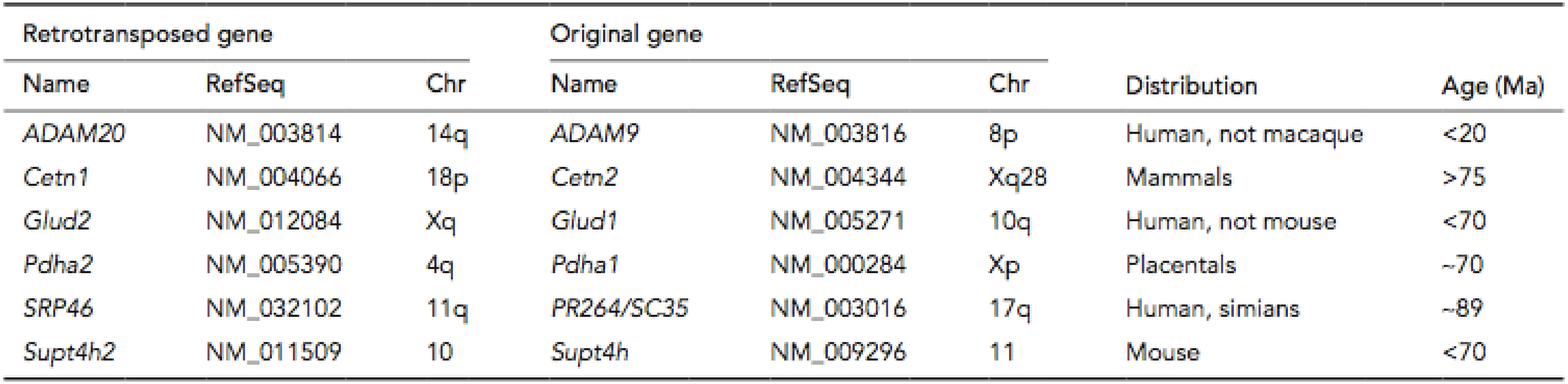
- Long-terminal repeat transposons (RNA-mediated)
- Long interspersed elements (LINEs); these encode a reverse transcriptase
- Short interspersed elements (SINEs)(RNA-mediated); these include *Alu* repeats
- DNA transposons (3% of human genome)

Examples of repeat classes and transposable elements



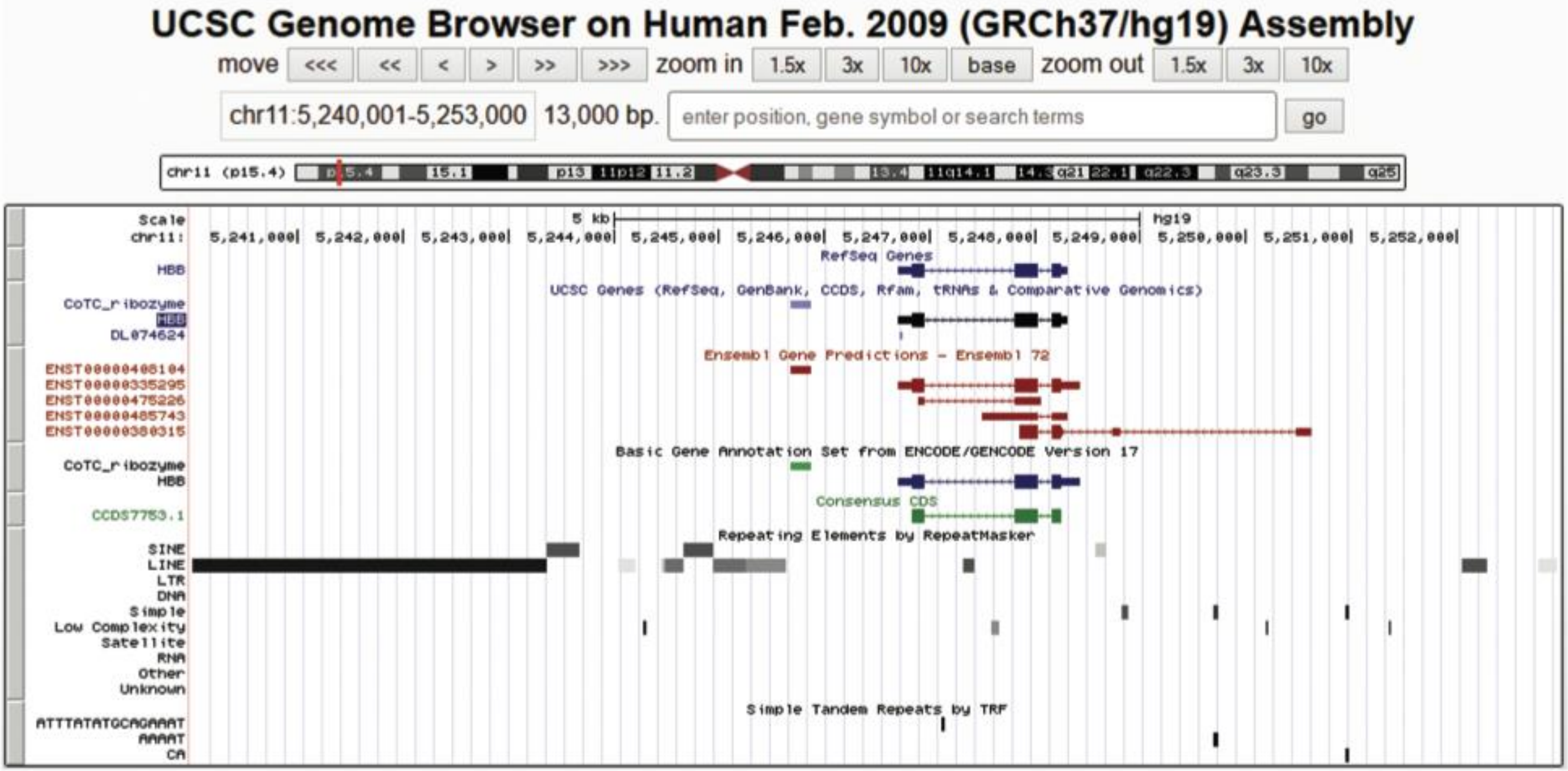
Class	Subclass	Superfamily	Examples of family	Approximate size range (bp)
Retroelements (RNA-mediated elements)	LTR retrotransposons	Ty1-copia	Opie-1 (maize)	3000–12,000
	Non-LTR retrotransposons	LINEs	*LINE-1* (human)	1000–7000
		SINEs	*Alu* (human)	100–500
DNA transposons	Cut-and-paste transposition	Mariner-Tc1	*Tc1* in *C. elegans*	1000–2000
		P	*P* in *Drosophila*	500–4600
	Rolling circle transposition	*Helitrons*	*Helitrons* in *A. thaliana, O. sativa,* and *C. elegans*	5500–17,500

Examples of mammalian genes generated by retrotransposition



Retrotransposed gene			Original gene				
Name	RefSeq	Chr	Name	RefSeq	Chr	Distribution	Age (Ma)
ADAM20	NM_003814	14q	*ADAM9*	NM_003816	8p	Human, not macaque	<20
Cetn1	NM_004066	18p	*Cetn2*	NM_004344	Xq28	Mammals	>75
Glud2	NM_012084	Xq	*Glud1*	NM_005271	10q	Human, not mouse	<70
Pdha2	NM_005390	4q	*Pdha1*	NM_000284	Xp	Placentals	~70
SRP46	NM_032102	11q	*PR264/SC35*	NM_003016	17q	Human, simians	~89
Supt4h2	NM_011509	10	*Supt4h*	NM_009296	11	Mouse	<70

Interspersed repeats via the UCSC Genome and Table Browser



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr11:5,240,001-5,253,000 13,000 bp.

Go to chr11:5,240,001-5,253,000 (13,000 bases) in the beta globin region. Set the RepeatMasker track to "full" to see repetitive DNA elements such as SINE, LINE, LTR, and DNA transposon.

Interspersed repeats via the UCSC Genome and Table Browser

(b) Access to tabular data on repeat elements using the UCSC Table Browser

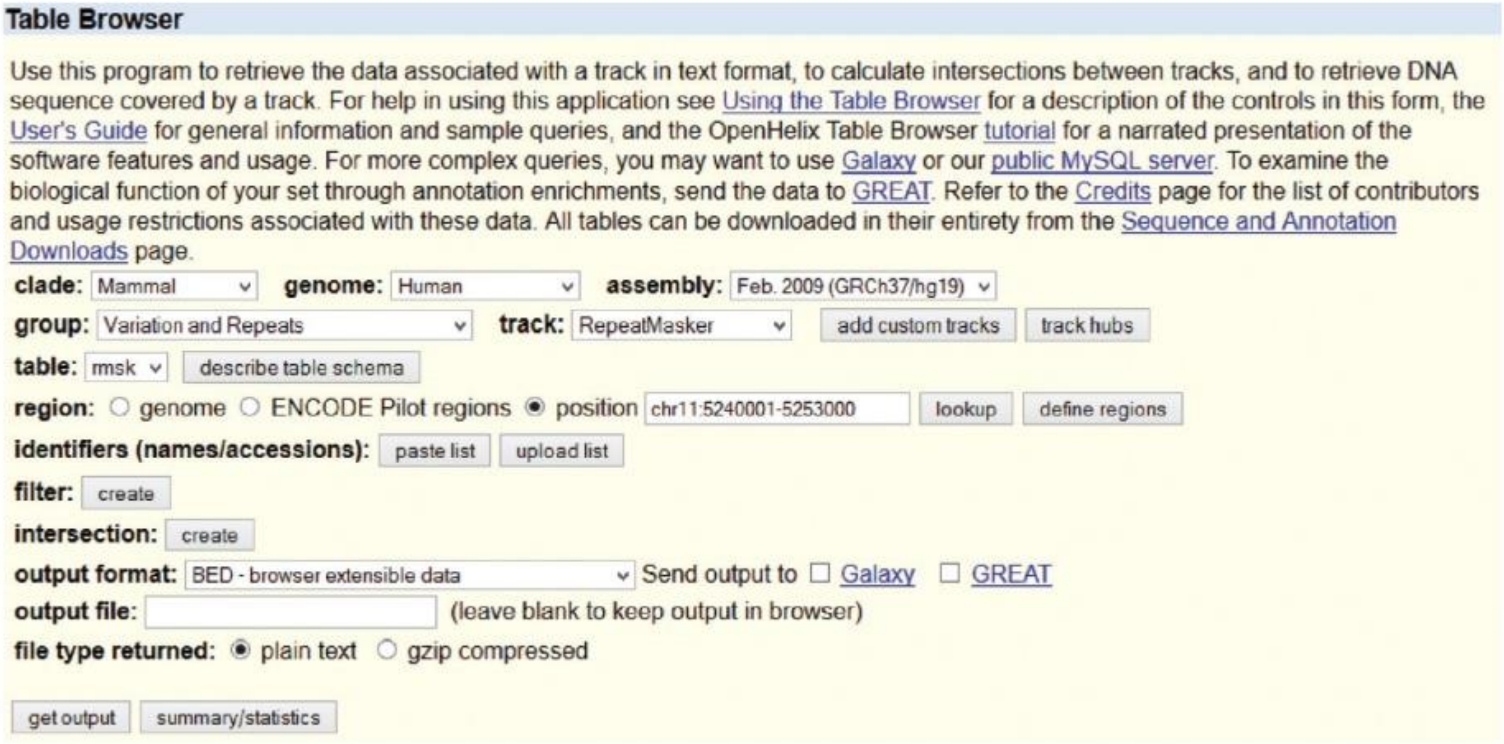


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Variation and Repeats **track:** RepeatMasker

table: rmsk

region: ○ genome ○ ENCODE Pilot regions ● position chr11:5240001-5253000

identifiers (names/accessions):

filter:

intersection:

output format: BED - browser extensible data Send output to □ Galaxy □ GREAT

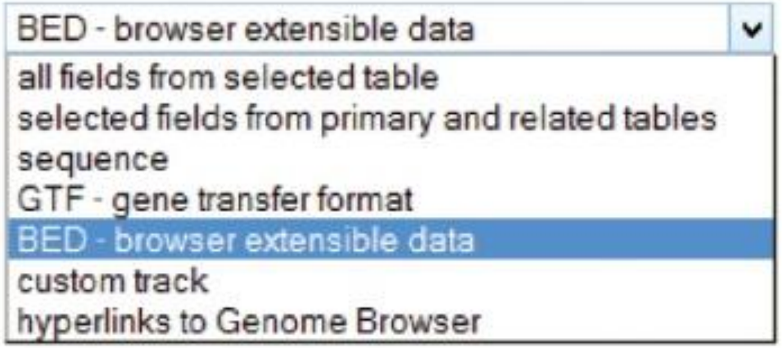
output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

The UCSC Table Browser is complementary to the Genome Browser. Information is presented in a tabular output.

Interspersed repeats via the UCSC Genome and Table Browser

(c) Options for Table Browser output formats

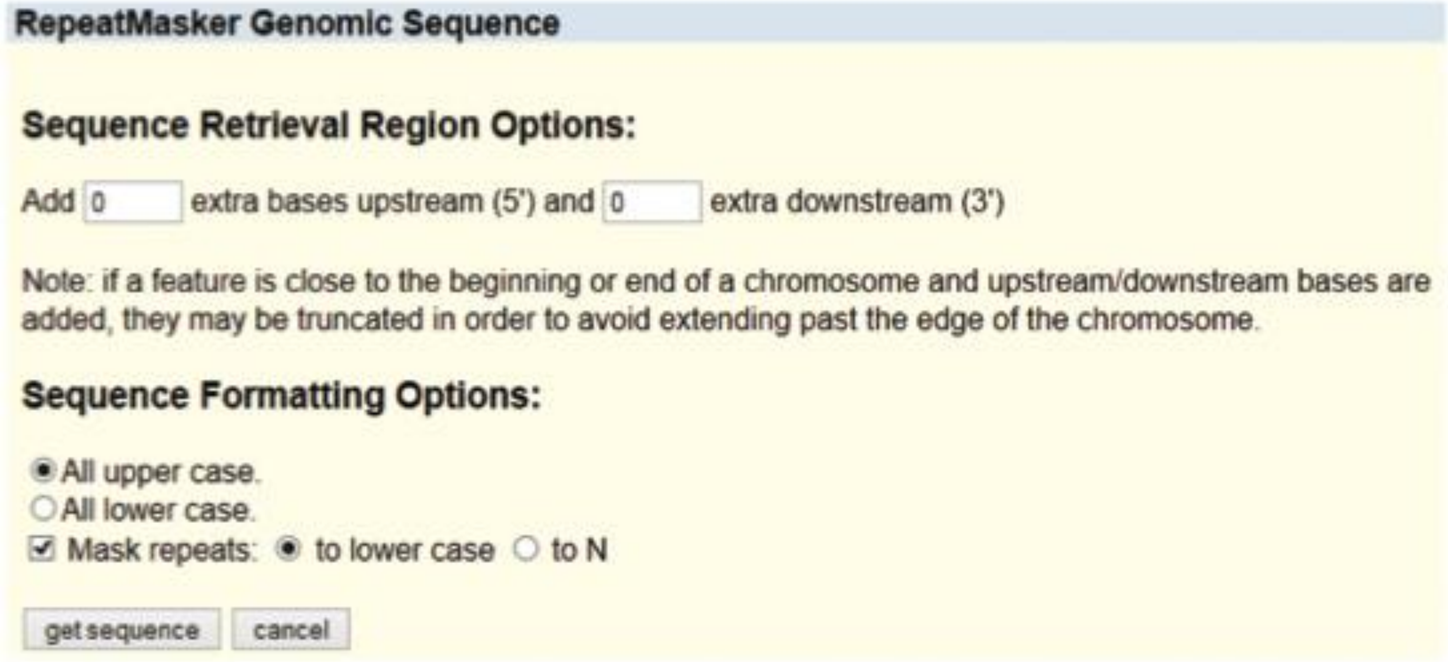


BED - browser extensible data

- all fields from selected table
- selected fields from primary and related tables
- sequence
- GTF - gene transfer format
- BED - browser extensible data
- custom track
- hyperlinks to Genome Browser

UCSC Table Browser output formats include browser extensible data (BED) files (see UCSC site for details).

Repeatmasker output (at UCSC Table Browser)



RepeatMasker Genomic Sequence

Sequence Retrieval Region Options:

Add 0 extra bases upstream (5') and 0 extra downstream (3')

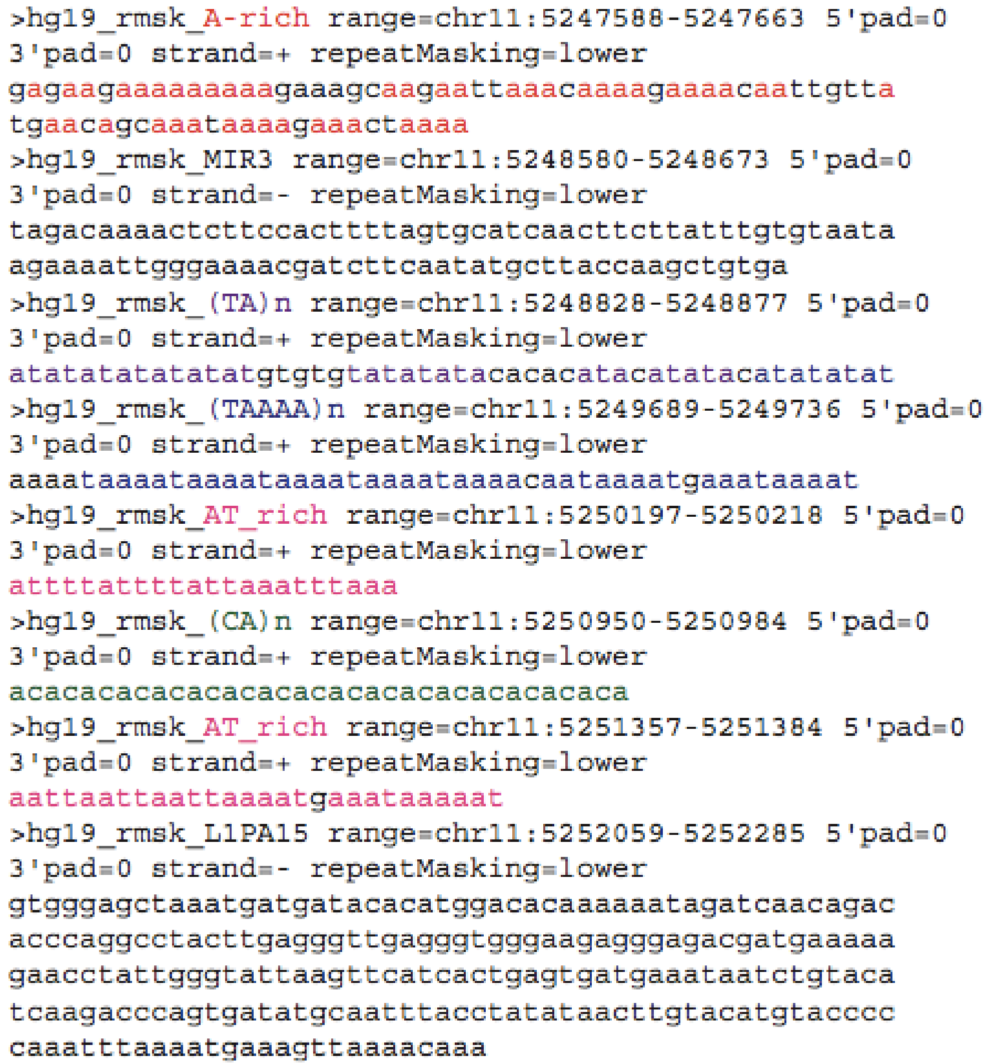
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

● All upper case.
○ All lower case.
☑ Mask repeats: ● to lower case ○ to N

You can select different output options.

Repeatmasker output (at UCSC Table Browser)



```
>hg19_rmsk_A-rich range=chr11:5247588-5247663 5'pad=0
3'pad=0 strand=+ repeatMasking=lower
gagaagaaaaaaaaagaaagcaagaattaaacaaaagaaaacaattgtta
tgaacagcaaataaaagaaactaaaa
>hg19_rmsk_MIR3 range=chr11:5248580-5248673 5'pad=0
3'pad=0 strand=- repeatMasking=lower
tagacaaaactcttccacttttagtgcatcaacttcttatttgtgtaata
agaaaattgggaaaacgatcttcaatatgcttaccaagctgtga
>hg19_rmsk_(TA)n range=chr11:5248828-5248877 5'pad=0
3'pad=0 strand=+ repeatMasking=lower
atatatatatatatgtgtgtatatatacacacatacatatacatatatat
>hg19_rmsk_(TAAAA)n range=chr11:5249689-5249736 5'pad=0
3'pad=0 strand=+ repeatMasking=lower
aaaataaaataaaataaaataaaataaaacaataaaatgaaataaaat
>hg19_rmsk_AT_rich range=chr11:5250197-5250218 5'pad=0
3'pad=0 strand=+ repeatMasking=lower
attttattttattaaatttaaa
>hg19_rmsk_(CA)n range=chr11:5250950-5250984 5'pad=0
3'pad=0 strand=+ repeatMasking=lower
acacacacacacacacacacacacacacacacaca
>hg19_rmsk_AT_rich range=chr11:5251357-5251384 5'pad=0
3'pad=0 strand=+ repeatMasking=lower
aattaattaattaaaatgaaataaaaat
>hg19_rmsk_L1PA15 range=chr11:5252059-5252285 5'pad=0
3'pad=0 strand=- repeatMasking=lower
gtgggagctaaatgatgatacacatggacacaaaaaatagatcaacagac
acccaggcctacttgagggttgagggtgggaagagggagacgatgaaaaa
gaacctattgggtattaagttcatcactgagtgatgaaataatctgtaca
tcaagacccagtgatatgcaatttacctatataacttgtacatgtacccc
caaatttaaaatgaaagttaaaacaaa
```

Repeats (such as polyA) are indicated in color.

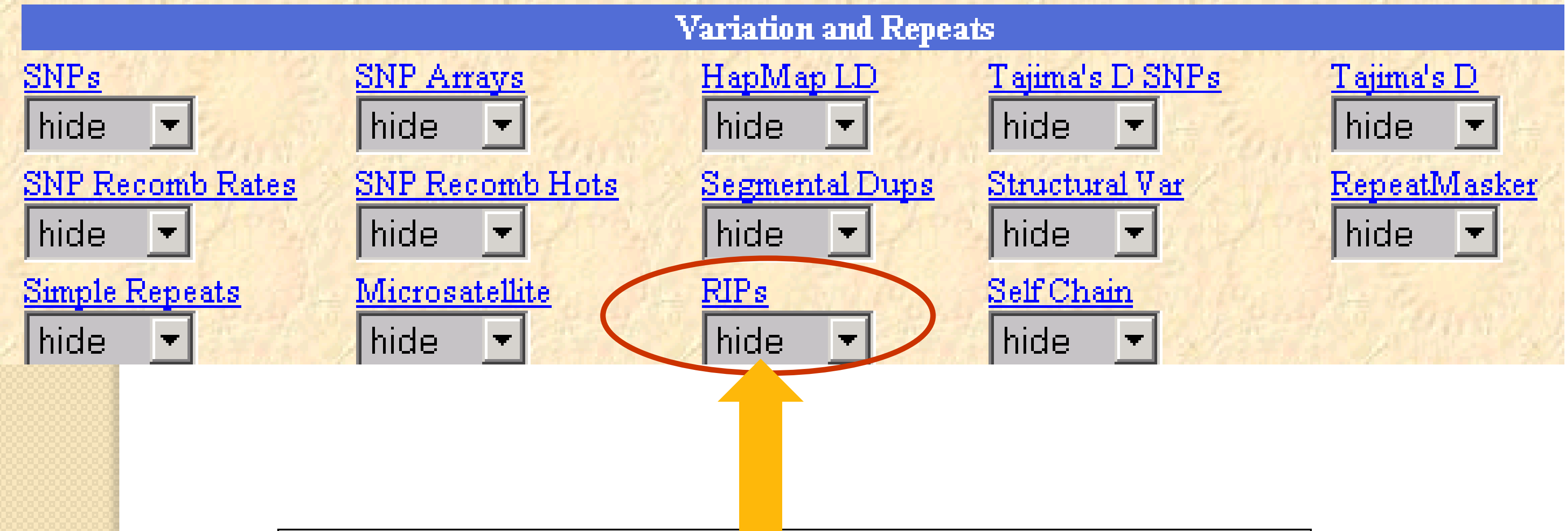
Five main classes of repetitive DNA

1. Interspersed repeats (transposon-derived repeats)

Examples include retrotransposed genes that lack introns, such as:

ADAM20	NM_003814	14q (original gene on 8p)
Cetn1	NM_004066	18p (original gene on Xq)
Glud2	NM_012084	Xq (original gene on 10q)
Pdha2	NM_005390	4q (original gene on Xp)

Interspersed repeats in the UCSC genome browser



Variation and Repeats

SNPs hide	SNP Arrays hide	HapMap LD hide	Tajima's D SNPs hide	Tajima's D hide
SNP Recomb Rates hide	SNP Recomb Hots hide	Segmental Dups hide	Structural Var hide	RepeatMasker hide
Simple Repeats hide	Microsatellite hide	RIPs hide	Self Chain hide	

Retrotransposon Insertion Polymorphisms

http://genome.ucsc.edu

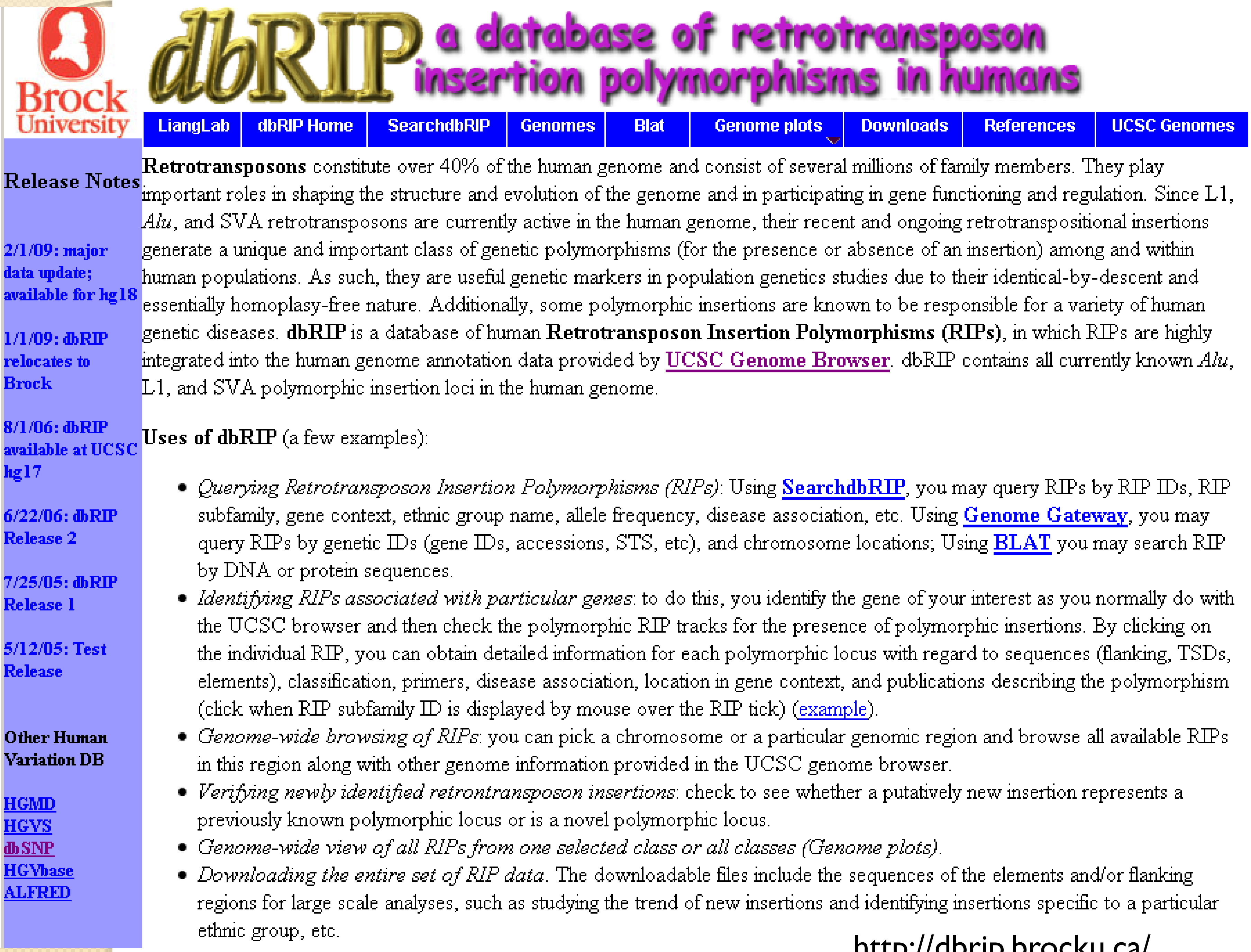
"Retrotransposons constitute over 40% of the human genome and consist of several millions of family members. They play important roles in shaping the structure and evolution of the genome and in participating in gene functioning and regulation. Since L1, Alu, and SVA retrotransposons are currently active in the human genome, their recent and ongoing retrotranspositional insertions generate a unique and important class of genetic polymorphisms (for the presence or absence of an insertion) among and within human populations. As such, they are useful genetic markers in population genetics studies due to their identical-by-descent and essentially homoplasy-free nature. Additionally, some polymorphic insertions are known to be responsible for a variety of human genetic diseases. dbRIP is a database of human Retrotransposon Insertion Polymorphisms (RIPs). dbRIP contains all currently known Alu, L1, and SVA polymorphic insertion loci in the human genome."

--dbRIP

Homoplasy: having some states arise more than once on a tree.

Wang J et al. (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat. 27:323-329.

Retrotransposons constitute over 40% of the human genome and play important roles in the evolution of the genome. Since certain types of retrotransposons, particularly members of the Alu, L1, and SVA families, are still active, their recent and ongoing propagation generates a unique and important class of human genomic diversity/polymorphism (for the presence and absence of an insertion) with some elements known to cause genetic diseases. So far, over 2,300, 500, and 80 Alu, L1, and SVA insertions, respectively, have been reported to be polymorphic and many more are yet to be discovered. We present here the Database of Retrotransposon Insertion Polymorphisms (dbRIP; http://falcon.roswellpark.org:9090), a highly integrated and interactive database of human retrotransposon insertion polymorphisms (RIPs). dbRIP currently contains a nonredundant list of 1,625, 407, and 63 polymorphic Alu, L1, and SVA elements, respectively, or a total of 2,095 RIPs. In dbRIP, we deploy the utilities and annotated data of the genome browser developed at the University of California at Santa Cruz (UCSC) for user-friendly queries and integrative browsing of RIPs along with all other genome annotation information. Users can query the database by a variety of means and have access to the detailed information related to a RIP, including detailed insertion sequences and genotype data. dbRIP represents the first database providing comprehensive, integrative, and interactive compilation of RIP data, and it will be a useful resource for researchers working in the area of human genetics.



dbRIP a database of retrotransposon insertion polymorphisms in humans

Release Notes

2/1/09: major data update; available for hg18

1/1/09: dbRIP relocates to Brock

8/1/06: dbRIP available at UCSC hg17

6/22/06: dbRIP Release 2

7/25/05: dbRIP Release 1

5/12/05: Test Release

Other Human Variation DB

HGMD
HGVS
dbSNP
HGVbase
ALFRED

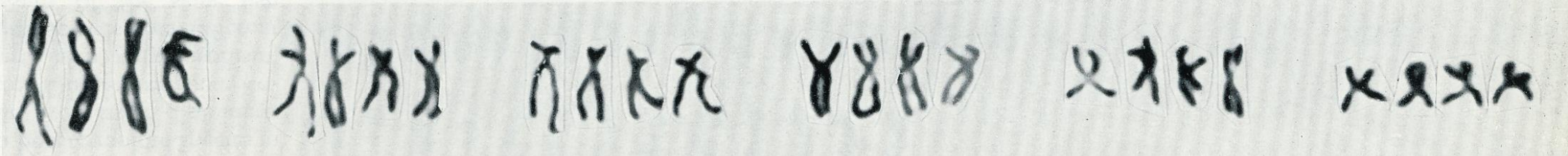
Retrotransposons constitute over 40% of the human genome and consist of several millions of family members. They play important roles in shaping the structure and evolution of the genome and in participating in gene functioning and regulation. Since L1, *Alu*, and SVA retrotransposons are currently active in the human genome, their recent and ongoing retrotranspositional insertions generate a unique and important class of genetic polymorphisms (for the presence or absence of an insertion) among and within human populations. As such, they are useful genetic markers in population genetics studies due to their identical-by-descent and essentially homoplasy-free nature. Additionally, some polymorphic insertions are known to be responsible for a variety of human genetic diseases. **dbRIP** is a database of human **Retrotransposon Insertion Polymorphisms (RIPs)**, in which RIPs are highly integrated into the human genome annotation data provided by **UCSC Genome Browser**. dbRIP contains all currently known *Alu*, L1, and SVA polymorphic insertion loci in the human genome.

Uses of dbRIP (a few examples):

- *Querying Retrotransposon Insertion Polymorphisms (RIPs)*: Using **SearchdbRIP**, you may query RIPs by RIP IDs, RIP subfamily, gene context, ethnic group name, allele frequency, disease association, etc. Using **Genome Gateway**, you may query RIPs by genetic IDs (gene IDs, accessions, STS, etc), and chromosome locations; Using **BLAT** you may search RIP by DNA or protein sequences.
- *Identifying RIPs associated with particular genes*: to do this, you identify the gene of your interest as you normally do with the UCSC browser and then check the polymorphic RIP tracks for the presence of polymorphic insertions. By clicking on the individual RIP, you can obtain detailed information for each polymorphic locus with regard to sequences (flanking, TSDs, elements), classification, primers, disease association, location in gene context, and publications describing the polymorphism (click when RIP subfamily ID is displayed by mouse over the RIP tick) (example).
- *Genome-wide browsing of RIPs*: you can pick a chromosome or a particular genomic region and browse all available RIPs in this region along with other genome information provided in the UCSC genome browser.
- *Verifying newly identified retrontransposon insertions*: check to see whether a putatively new insertion represents a previously known polymorphic locus or is a novel polymorphic locus.
- *Genome-wide view of all RIPs from one selected class or all classes (Genome plots)*.
- *Downloading the entire set of RIP data*. The downloadable files include the sequences of the elements and/or flanking regions for large scale analyses, such as studying the trend of new insertions and identifying insertions specific to a particular ethnic group, etc.

http://dbrip.brocku.ca/

Repetitive DNA content of eukaryotic genomes



1. Interspersed repeats
2. Processed pseudogenes
3. Simple sequence repeats
4. Segmental duplications
5. Blocks of tandem repeats

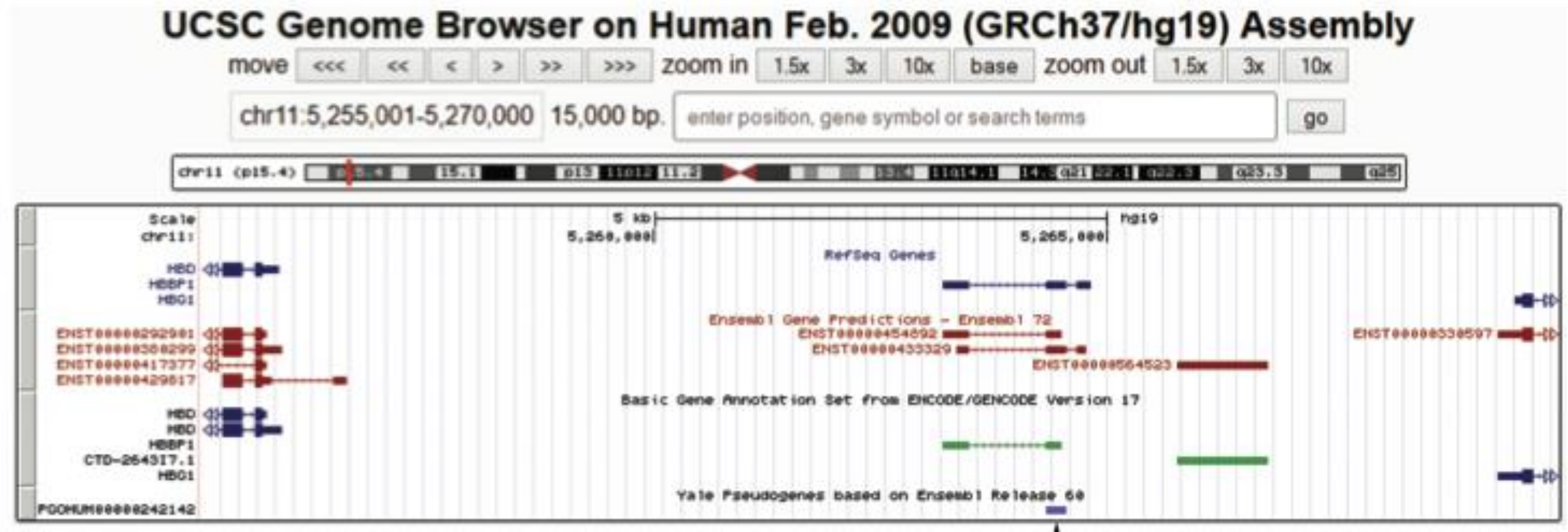
Five main classes of repetitive DNA

2. Processed pseudogenes

These genes have a stop codon or frameshift mutation and do not encode a functional protein. They commonly arise from retrotransposition, or following gene duplication and subsequent gene loss.

For a superb on-line resource, visit http://www.pseudogene.org. Gerstein and colleagues (2006) suggest that there are ~19,000 pseudogenes in the human genome, slightly fewer than the number of functional protein-coding genes. (11,000 non-processed, 8,000 processed [lack introns].)

Pseudogenes: view of *HBBP1* pseudogene (15,000 bp view)

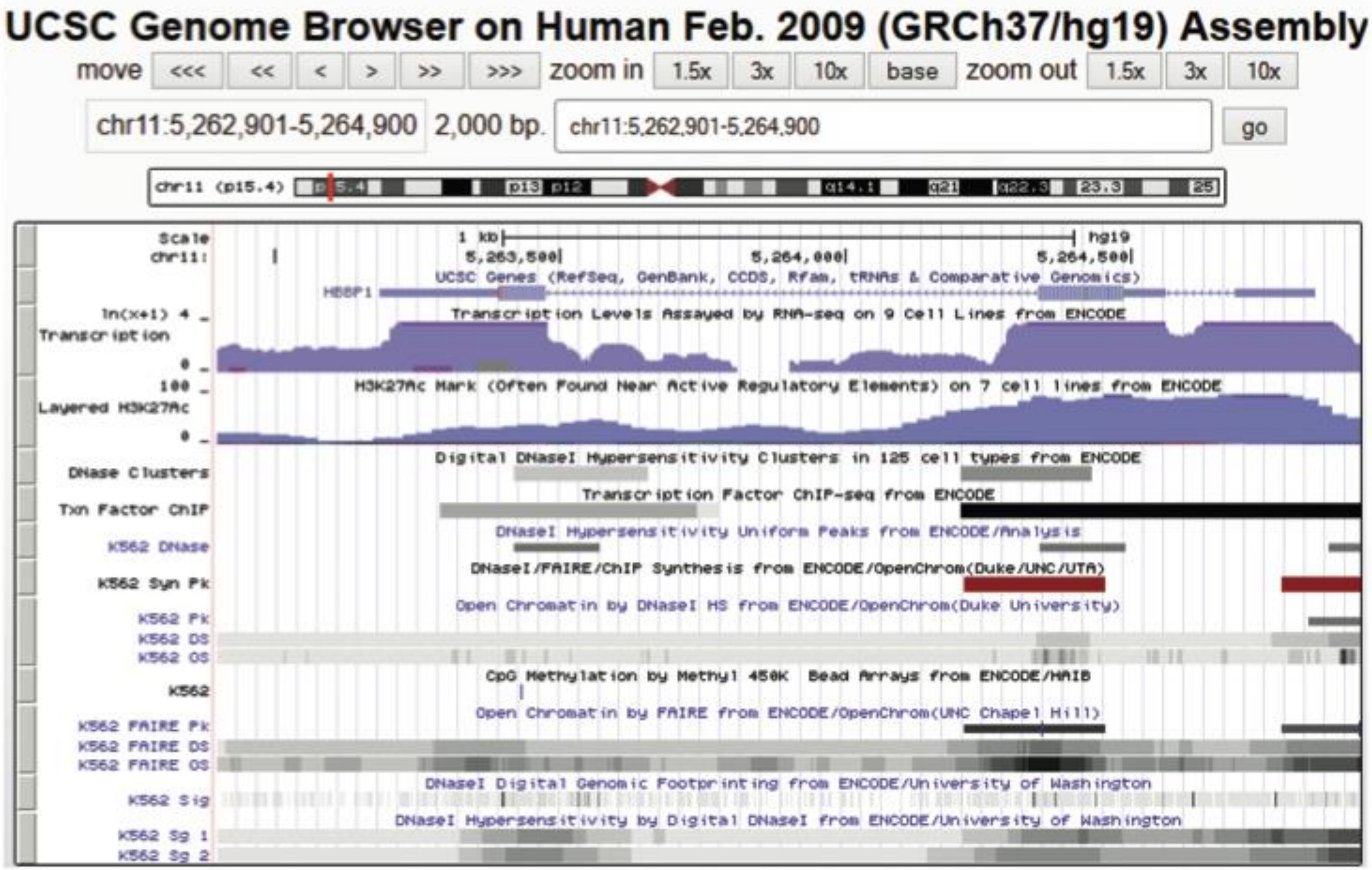


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr11:5,255,001-5,270,000 15,000 bp.

View of a globin pseudogene and its neighboring genes.

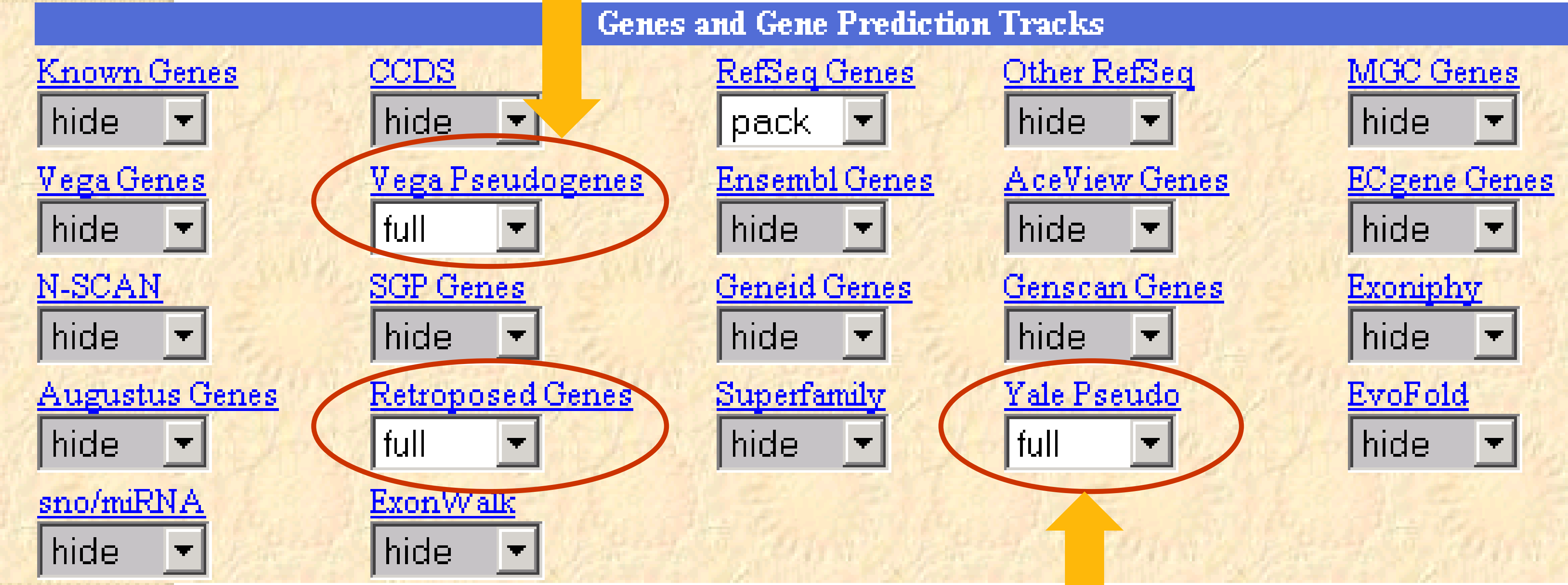
Pseudogenes: view of *HBBP1* pseudogene (2,000 bp view)


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
chr11:5,262,901-5,264,900 2,000 bp. chr11:5,262,901-5,264,900 go
chr11 (p15.4) p15.4 p13 p12 q14.1 q21 q22.3 23.3 25
Scale
chr11:
1 kb
hg19
5,263,500
5,264,000
5,264,500
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
HBBP1
ln(x+1) 4
Transcription
0
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
K562 DNase
DNaseI Hypersensitivity Uniform Peaks from ENCODE/Analysis
K562 Syn Pk
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
K562 Pk
K562 DS
K562 OS
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
K562
CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB
K562 FAIRE Pk
K562 FAIRE DS
K562 FAIRE OS
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
K562 Sig
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
K562 Sg 1
K562 Sg 2
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington

ENCODE annotation tracks are included, suggesting transcription of RNA of the pseudogene.

Pseudogenes in the UCSC genome browser

VEGA pseudogenes

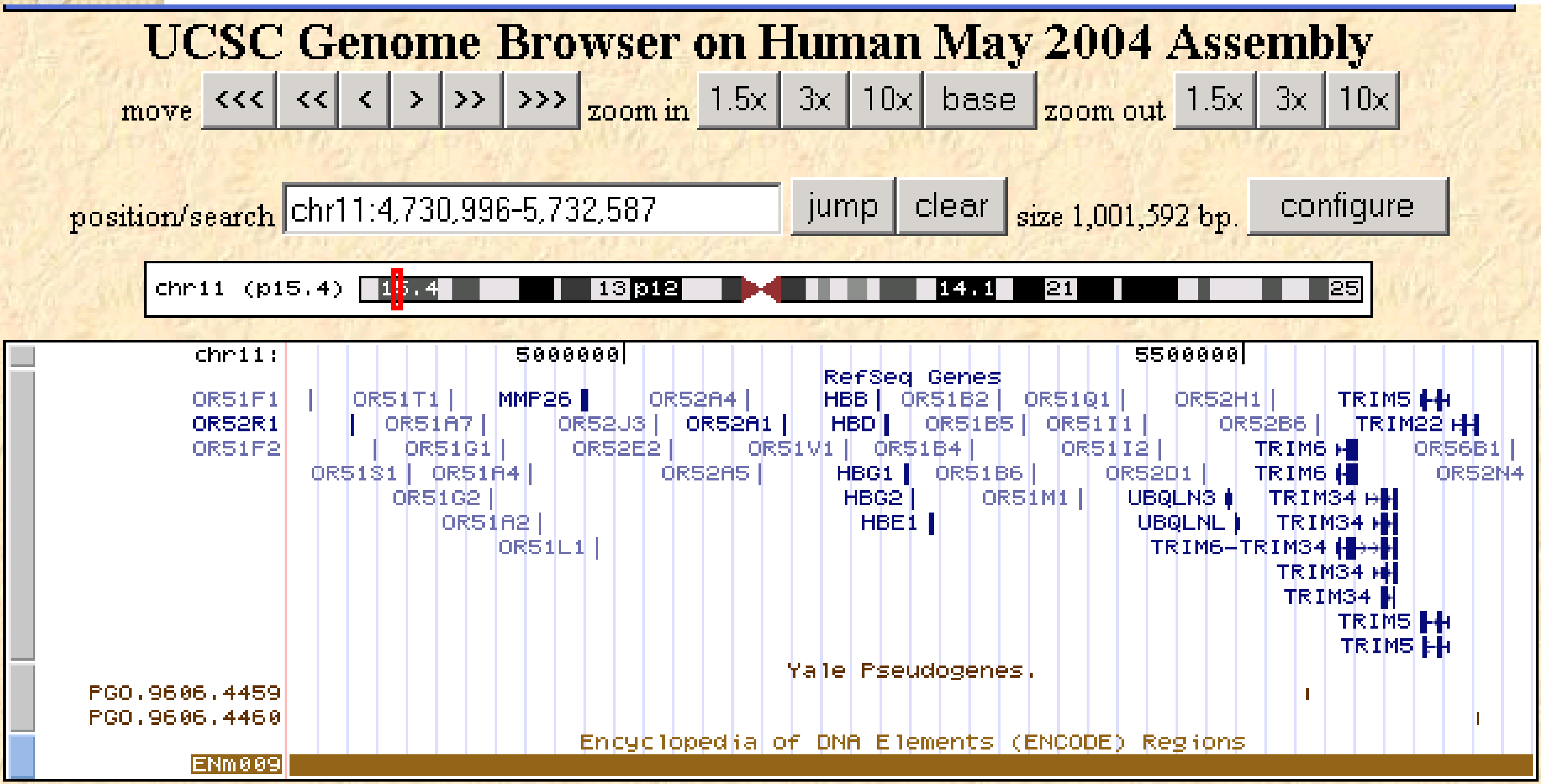


Genes and Gene Prediction Tracks

Known Genes	CCDS	RefSeq Genes	Other RefSeq	MGC Genes
hide	hide	pack	hide	hide
Vega Genes	Vega Pseudogenes	Ensembl Genes	AceView Genes	ECgene Genes
hide	full	hide	hide	hide
N-SCAN	SGP Genes	Geneid Genes	Genscan Genes	Exoniphy
hide	hide	hide	hide	hide
Augustus Genes	Retroposed Genes	Superfamily	Yale Pseudo	EvoFold
hide	full	hide	full	hide
sno/miRNA	ExonWalk			
hide	hide			

Yale pseudogenes

Pseudogenes in the beta globin region



UCSC Genome Browser on Human May 2004 Assembly

position/search: chr11:4,730,996-5,732,587

size 1,001,592 bp.

Vertebrate Genome Annotation (VEGA) database

- From the VEGA home page (http://vega.sanger.ac.uk):

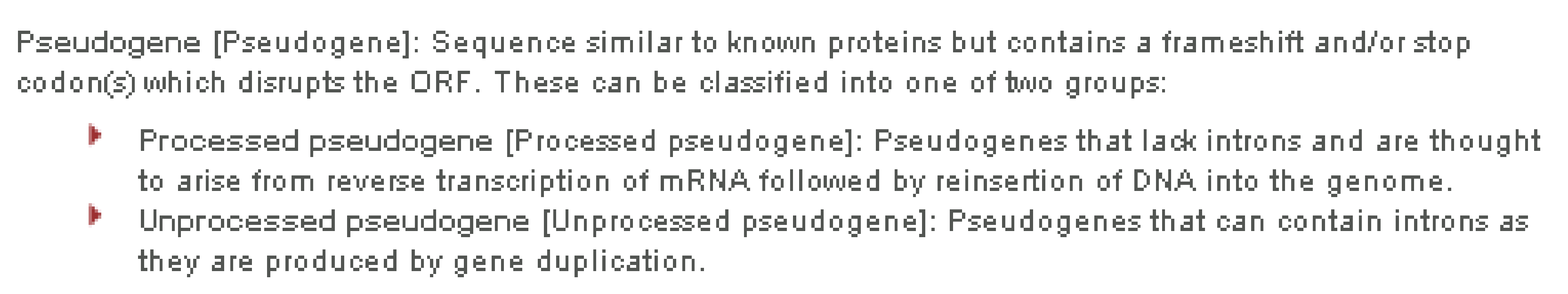
  "The Vertebrate Genome Annotation (VEGA) database build 30 is designed to be a central repository for manual annotation of different vertebrate finished genome sequence. In collaboration with the genome sequencing centres Vega attempts to present consistent high-quality curation of the published chromosome sequences."

  "Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases as well as a series of ab initio gene predictions (GENSCAN, Fgenes)."

  "In addition, comparative analysis using vertebrate datasets such as the Riken mouse cDNAs and Genoscope *Tetraodon nigroviridis* Ecores (Evolutionary Conserved Regions) are used for novel gene discovery."

Vertebrate Genome Annotation (VEGA) database

- VEGA definition of pseudogenes (http://vega.sanger.ac.uk):

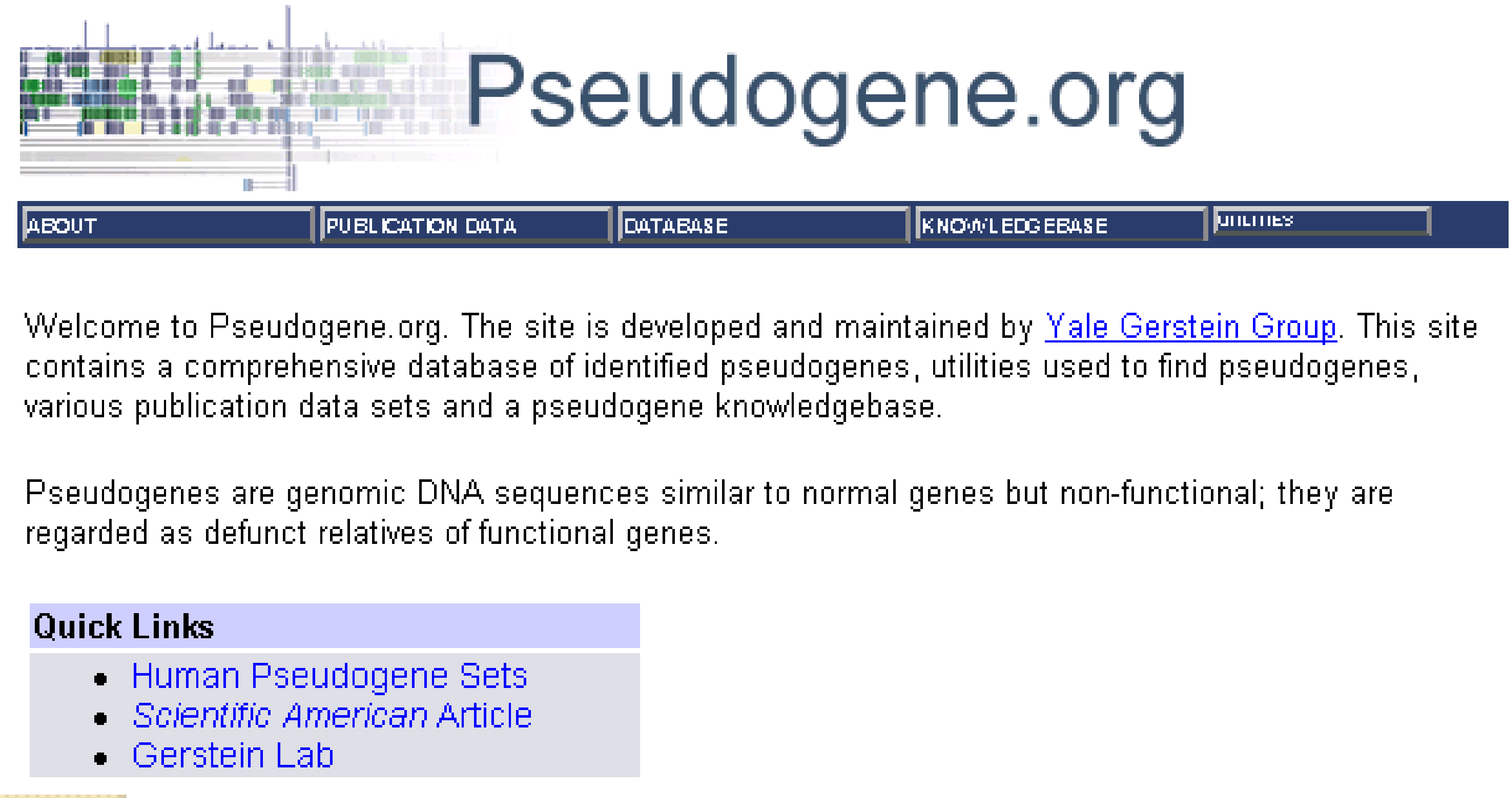


Pseudogene [Pseudogene]: Sequence similar to known proteins but contains a frameshift and/or stop codon(s) which disrupts the ORF. These can be classified into one of two groups:

- Processed pseudogene [Processed pseudogene]: Pseudogenes that lack introns and are thought to arise from reverse transcription of mRNA followed by reinsertion of DNA into the genome.
- Unprocessed pseudogene [Unprocessed pseudogene]: Pseudogenes that can contain introns as they are produced by gene duplication.

Yale pseudogene database

http://www.pseudogene.org



Welcome to Pseudogene.org. The site is developed and maintained by Yale Gerstein Group. This site contains a comprehensive database of identified pseudogenes, utilities used to find pseudogenes, various publication data sets and a pseudogene knowledgebase.

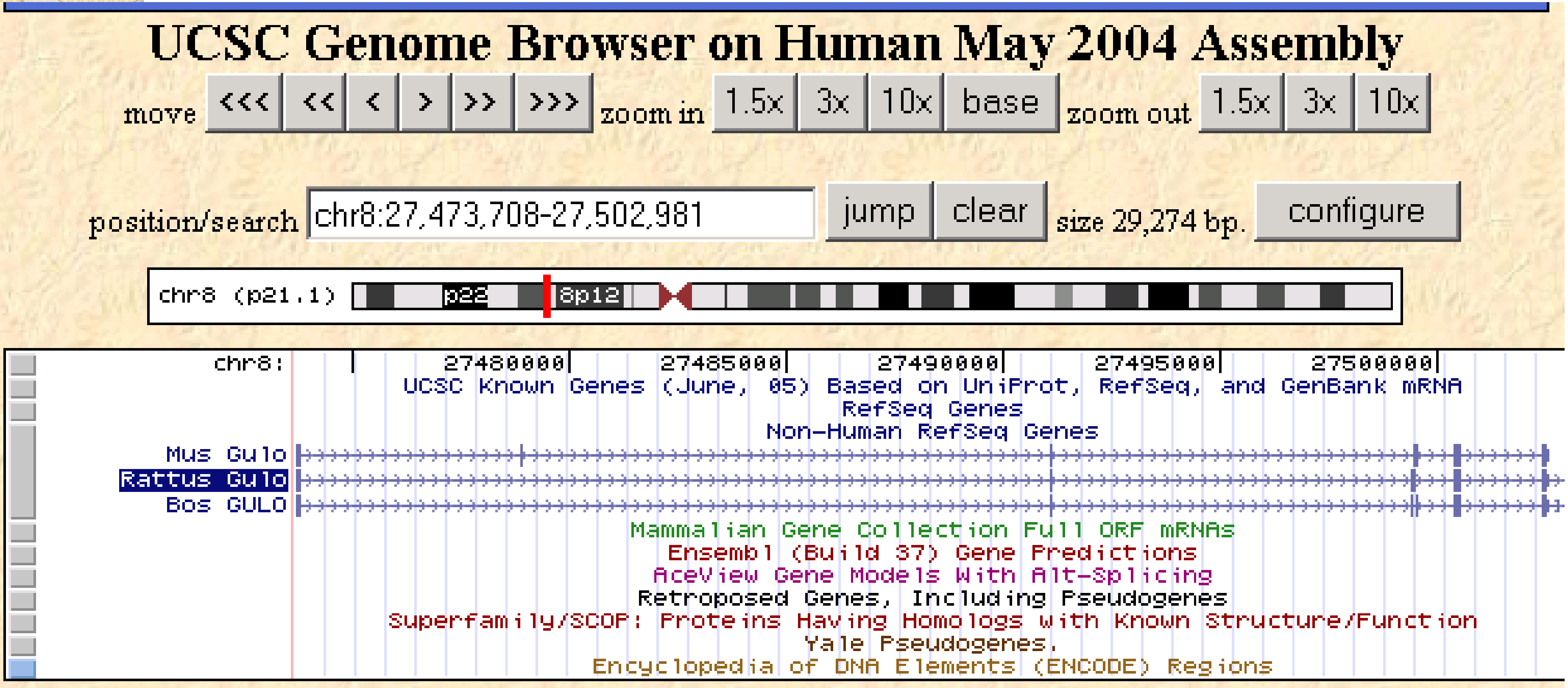
Pseudogenes are genomic DNA sequences similar to normal genes but non-functional; they are regarded as defunct relatives of functional genes.

Quick Links

- Human Pseudogene Sets
- *Scientific American* Article
- Gerstein Lab

Pseudogenes: example

Mouse GULO, required for vitamin C biosynthesis, has become a pseudogene in the primate lineage (γGULO). Here is an output for GULO on the human genome:



UCSC Genome Browser on Human May 2004 Assembly

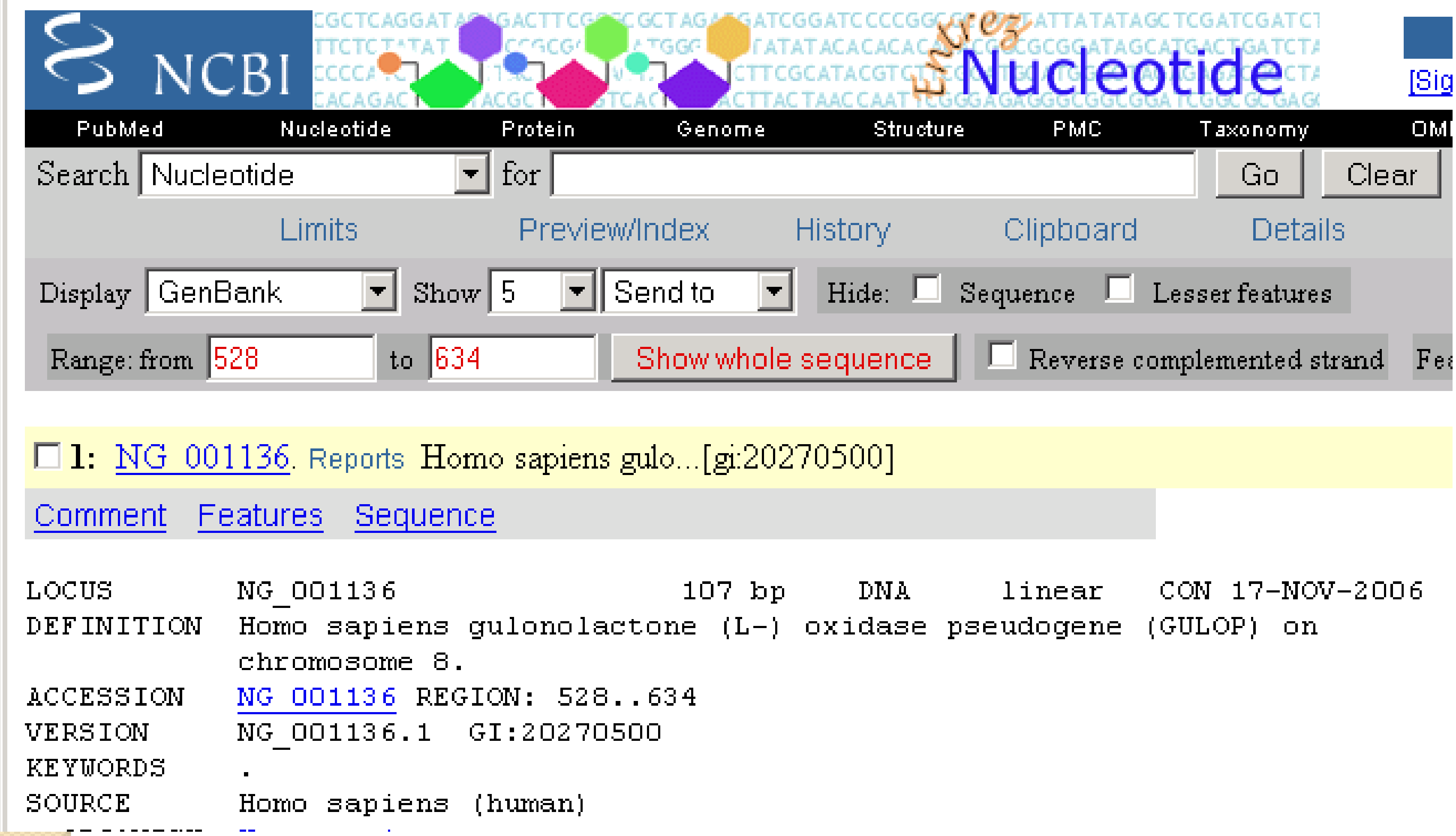
position/search chr8:27,473,708-27,502,981

size 29,274 bp.

move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
jump
clear
configure
chr8 (p21.1)
p22
8p12
chr8:
27480000
27485000
27490000
27495000
27500000
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
RefSeq Genes
Non-Human RefSeq Genes
Mus Gulo
Rattus Gulo
Bos GULO
Mammalian Gene Collection Full ORF mRNAs
Ensembl (Build 37) Gene Predictions
AceView Gene Models With Alt-Splicing
Retroposed Genes, Including Pseudogenes
Superfamily/SCOP: Proteins Having Homologs with Known Structure/Function
Yale Pseudogenes.
Encyclopedia of DNA Elements (ENCODE) Regions

Pseudogenes: example

GULO pseudogene in NCBI nucleotide:



Search: Nucleotide for

Display: GenBank Show: 5

Hide: □ Sequence □ Lesser features

Range: from 528 to 634

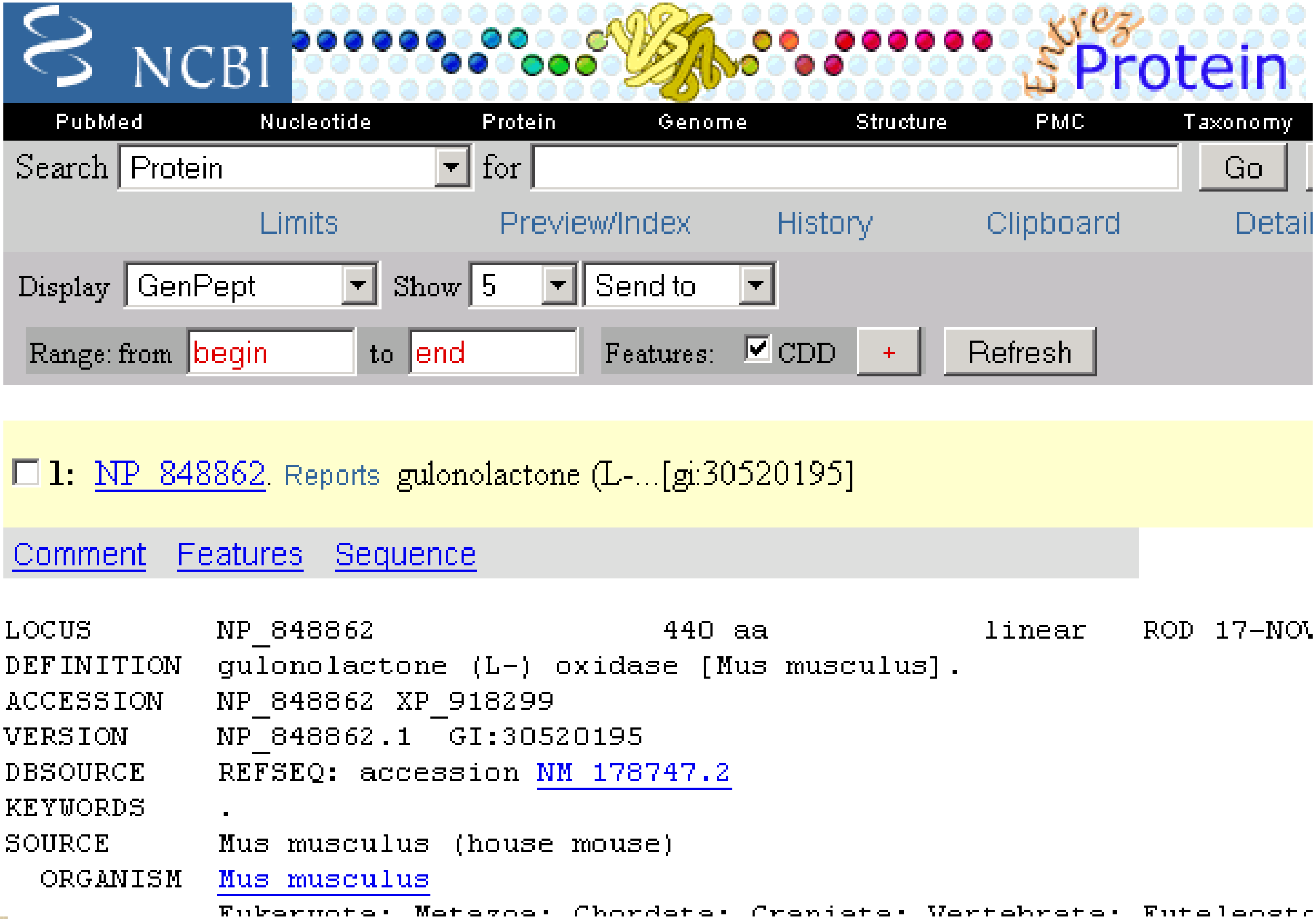
□ Reverse complemented strand

□ 1: NG_001136. Reports Homo sapiens gulo...[gi:20270500]

```
LOCUS       NG_001136                107 bp    DNA     linear   CON 17-NOV-2006
DEFINITION  Homo sapiens gulonolactone (L-) oxidase pseudogene (GULOP) on
            chromosome 8.
ACCESSION   NG_001136 REGION: 528..634
VERSION     NG_001136.1  GI:20270500
KEYWORDS    .
SOURCE      Homo sapiens (human)
```

Pseudogenes: example

Mouse GULO in NCBI Protein:



Search Protein

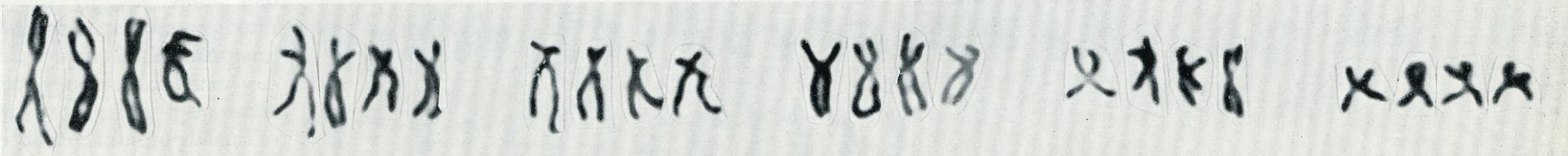
Display GenPept Show 5

Range: from begin to end Features: ☑ CDD

□ 1: NP_848862. Reports gulonolactone (L-...[gi:30520195]

```
LOCUS       NP_848862                 440 aa            linear   ROD 17-NOV
DEFINITION  gulonolactone (L-) oxidase [Mus musculus].
ACCESSION   NP_848862 XP_918299
VERSION     NP_848862.1  GI:30520195
DBSOURCE    REFSEQ: accession NM_178747.2
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostc
```

Repetitive DNA content of eukaryotic genomes



1. Interspersed repeats
2. Processed pseudogenes
3. Simple sequence repeats
4. Segmental duplications
5. Blocks of tandem repeats

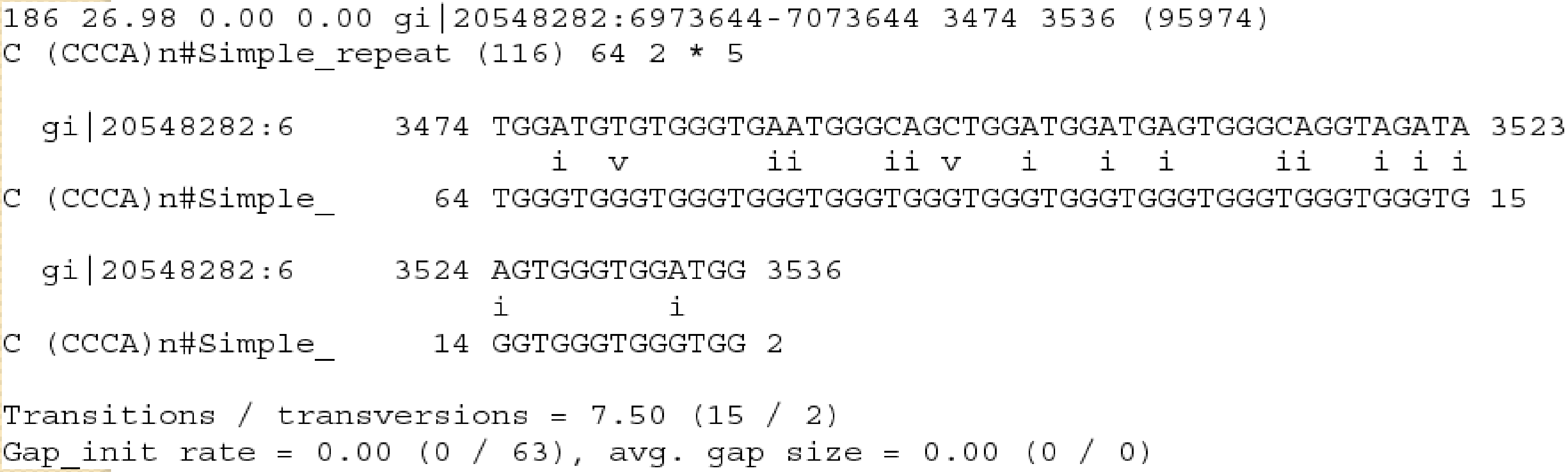
Five main classes of repetitive DNA

3. Simple sequence repeats

Microsatellites: from one to a dozen base pairs
Examples: $(A)_n$, $(CA)_n$, $(CGG)_n$
These may be formed by replication slippage.
Minisatellites: a dozen to 500 base pairs

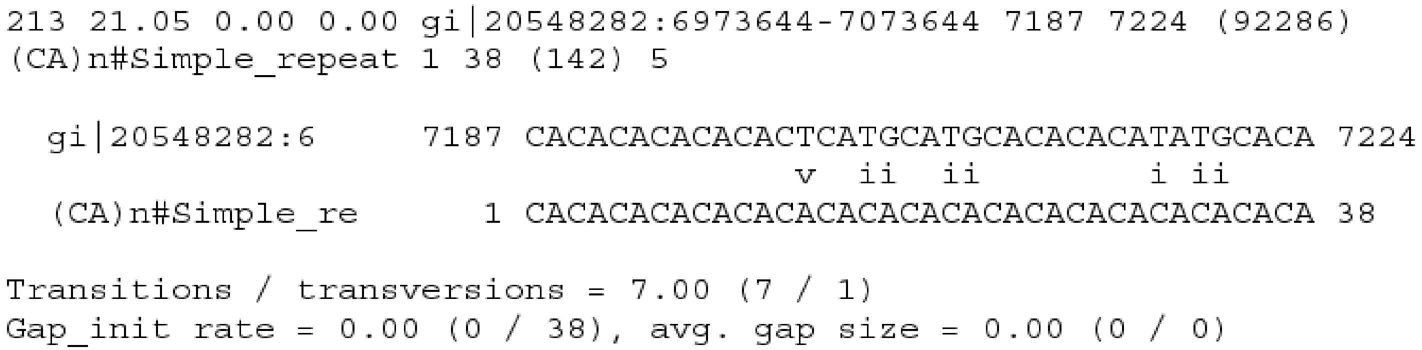
Simple sequence repeats of a particular length and composition occur preferentially in different species.
In humans, an expansion of triplet repeats such as CAG is associated with at least 14 disorders (including Huntington's disease).

Example of a simple sequence repeat (CCCA or GGGT) in human genomic DNA

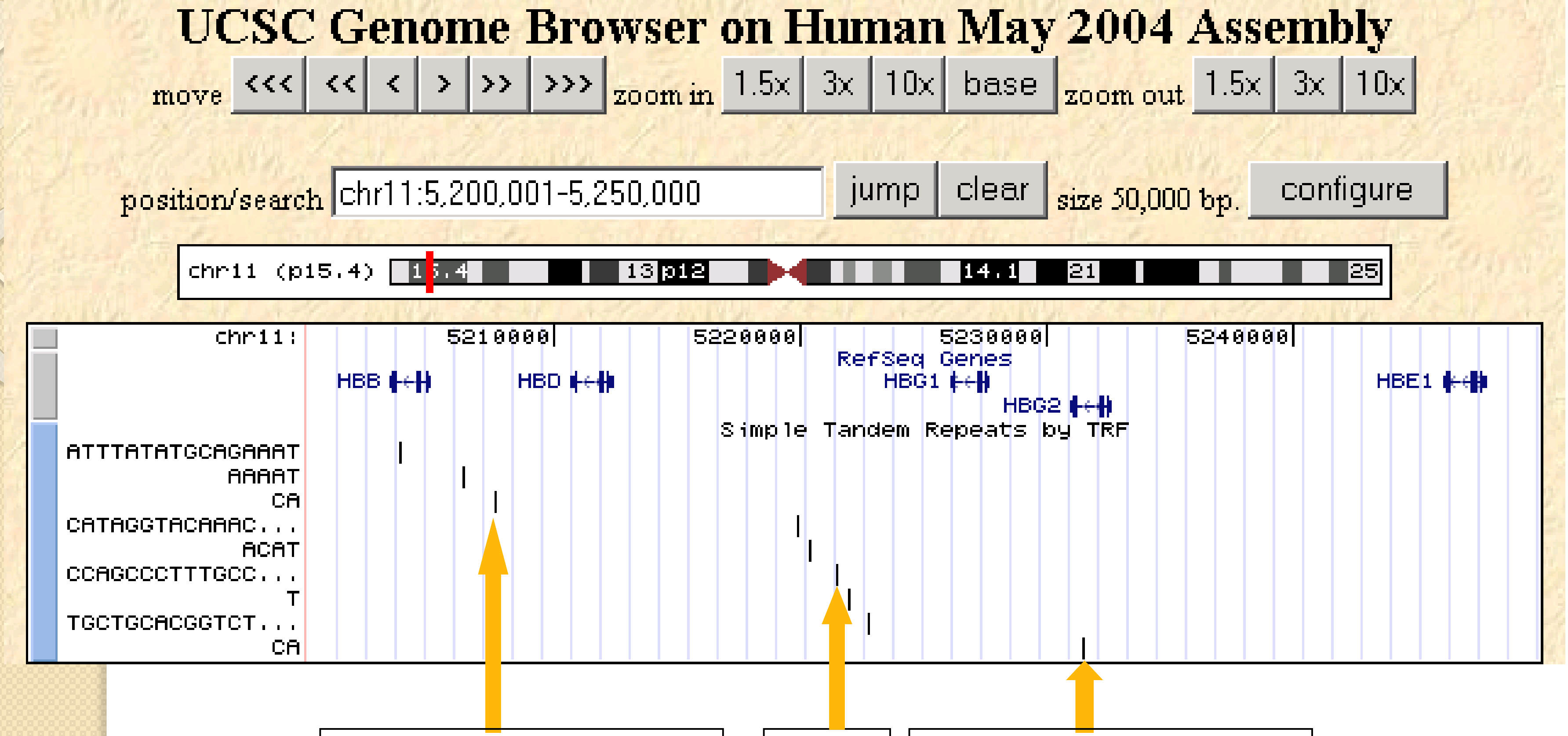


```
186 26.98 0.00 0.00 gi|20548282:6973644-7073644 3474 3536 (95974)
C (CCCA)n#Simple_repeat (116) 64 2 * 5

  gi|20548282:6     3474 TGGATGTGTGGGTGAATGGGCAGCTGGATGGATGAGTGGGCAGGTAGATA 3523
                            i  v       ii    ii v   i   i  i     ii   i i i
C (CCCA)n#Simple_     64 TGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTG 15

  gi|20548282:6     3524 AGTGGGTGGATGG 3536
                         i        i
C (CCCA)n#Simple_     14 GGTGGGTGGGTGG 2

Transitions / transversions = 7.50 (15 / 2)
Gap_init rate = 0.00 (0 / 63), avg. gap size = 0.00 (0 / 0)
```

RepeatMasker identifies simple sequence repeats



```
213 21.05 0.00 0.00 gi|20548282:6973644-7073644 7187 7224 (92286)
(CA)n#Simple_repeat 1 38 (142) 5

  gi|20548282:6     7187 CACACACACACACTCATGCATGCACACACATATGCACA 7224
                                      v  ii  ii        i ii
  (CA)n#Simple_re      1 CACACACACACACACACACACACACACACACACACACA 38

Transitions / transversions = 7.00 (7 / 1)
Gap_init rate = 0.00 (0 / 38), avg. gap size = 0.00 (0 / 0)
```



UCSC Genome Browser on Human May 2004 Assembly

position/search chr11:5,200,001-5,250,000

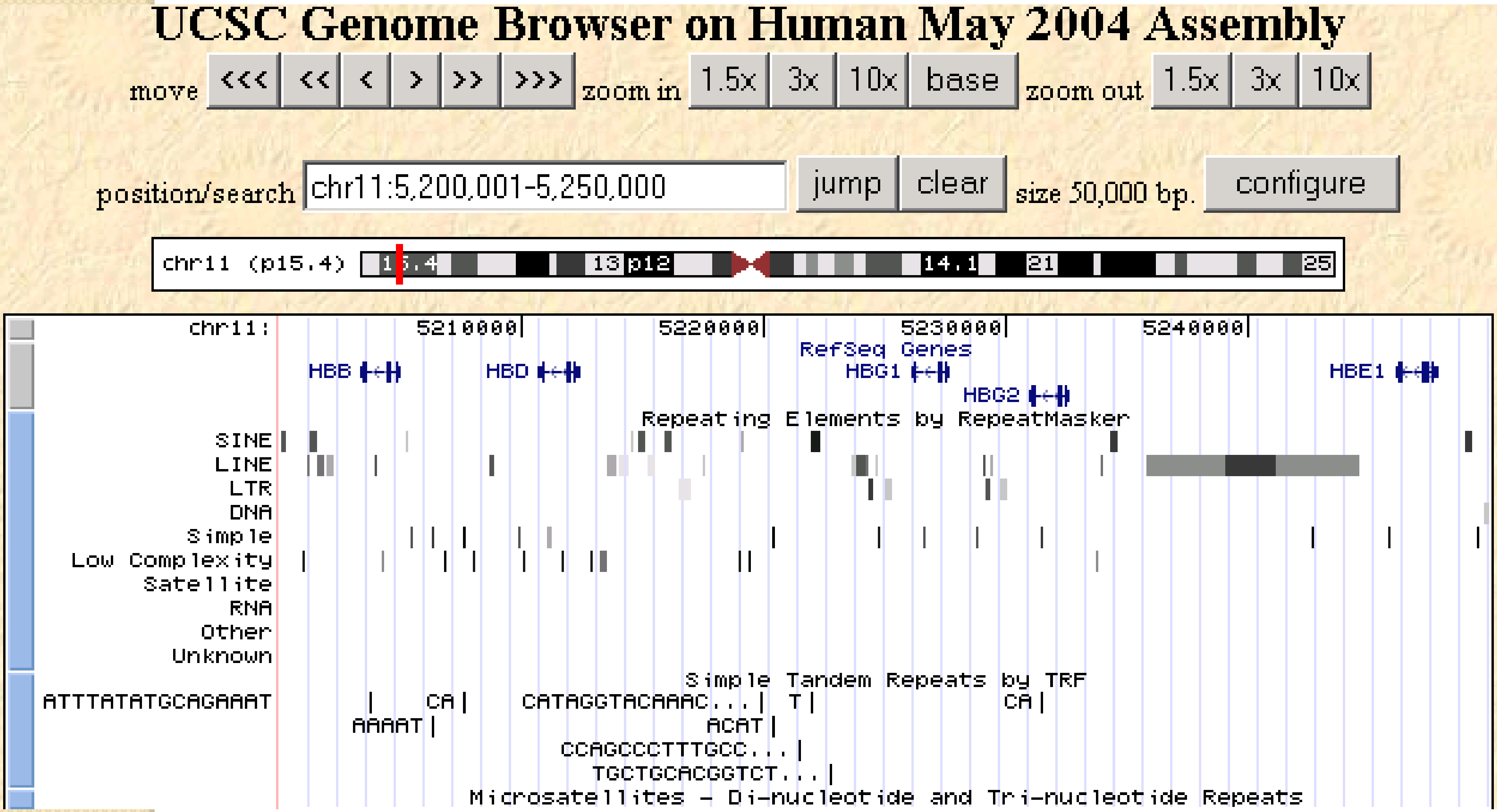
size 50,000 bp.

20 copies of CA

T48

23 copies of CA

Beta globin locus: tandem repeats, microsatellites, and RepeatMasker

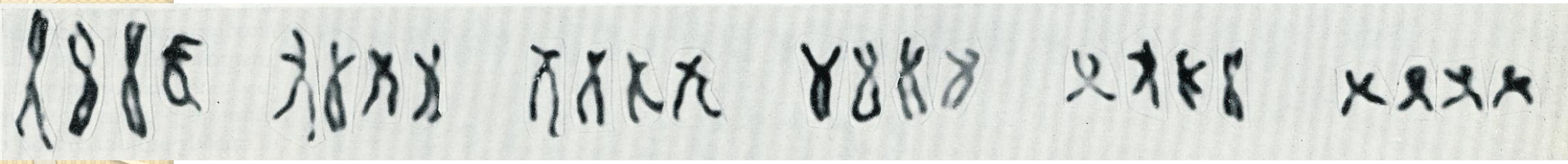


UCSC Genome Browser on Human May 2004 Assembly

position/search chr11:5,200,001-5,250,000

size 50,000 bp.

Repetitive DNA content of eukaryotic genomes



1. Interspersed repeats
2. Processed pseudogenes
3. Simple sequence repeats
4. Segmental duplications
5. Blocks of tandem repeats

Five main classes of repetitive DNA

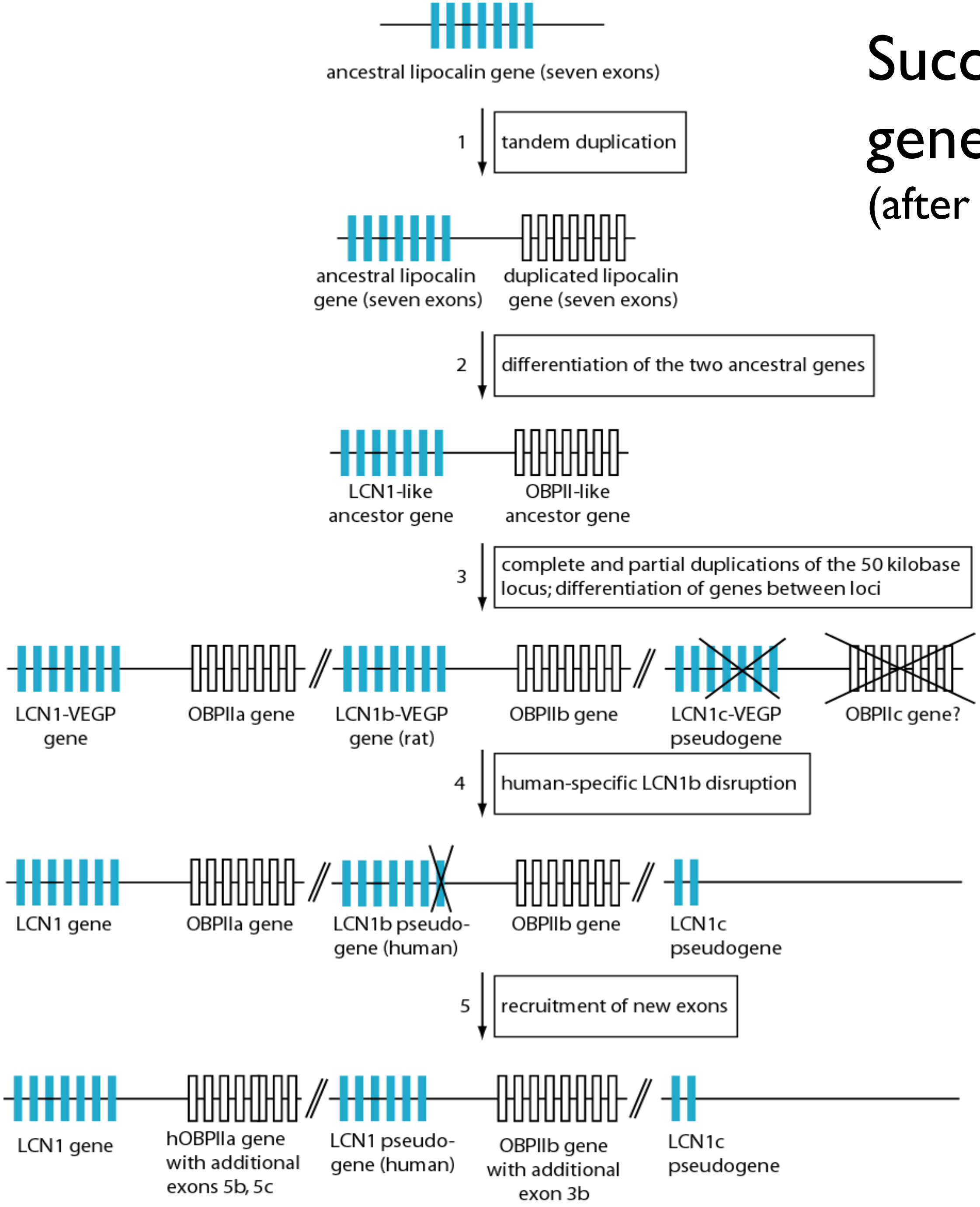
4. Segmental duplications

These are blocks of about 1 kilobase to 300 kb that are copied intra- or interchromosomally. Evan Eichler and colleagues estimate that about 5% of the human genome consists of segmental duplications. Duplicated regions often share very high (99%) sequence identity.

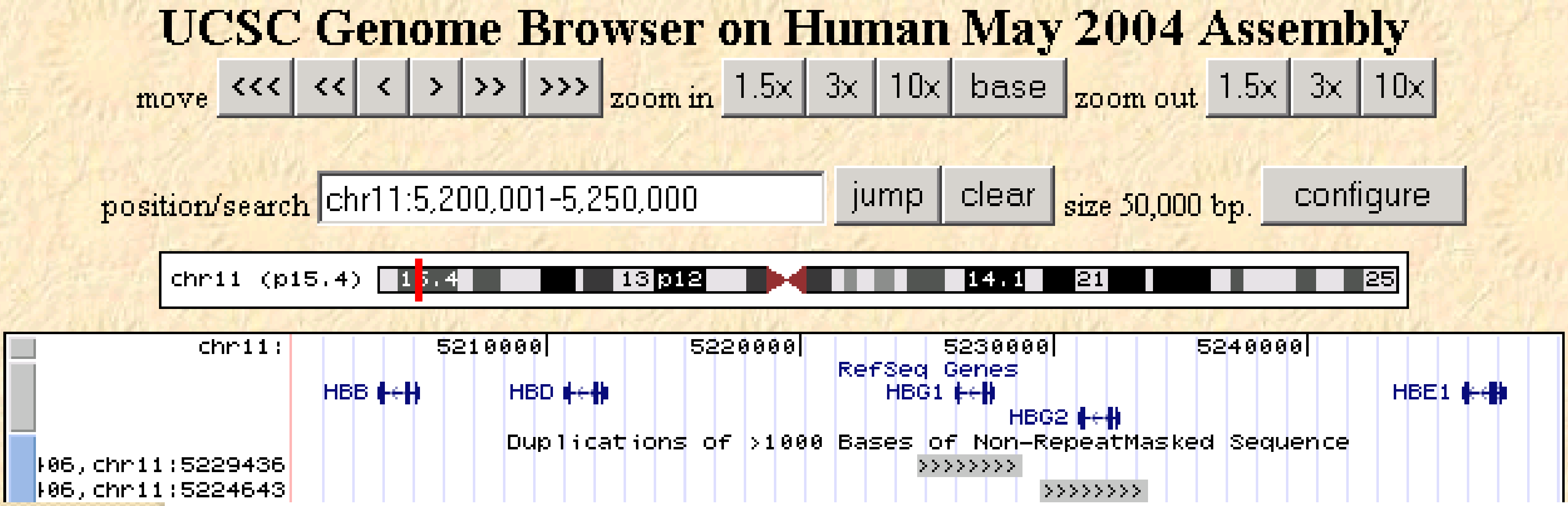
As an example, consider a group of lipocalin genes on human chromosome 9.

Successive tandem gene duplications

(after Lacazette et al., 2000)


ancestral lipocalin gene (seven exons)
1
tandem duplication
ancestral lipocalin gene (seven exons)
duplicated lipocalin gene (seven exons)
2
differentiation of the two ancestral genes
LCN1-like ancestor gene
OBPII-like ancestor gene
3
complete and partial duplications of the 50 kilobase locus; differentiation of genes between loci
LCN1-VEGP gene
OBPIIa gene
LCN1b-VEGP gene (rat)
OBPIIb gene
LCN1c-VEGP pseudogene
OBPIIc gene?
4
human-specific LCN1b disruption
LCN1 gene
OBPIIa gene
LCN1b pseudo-gene (human)
OBPIIb gene
LCN1c pseudogene
5
recruitment of new exons
LCN1 gene
hOBPIIa gene with additional exons 5b, 5c
LCN1 pseudo-gene (human)
OBPIIb gene with additional exon 3b
LCN1c pseudogene

Beta globin locus: segmental duplications



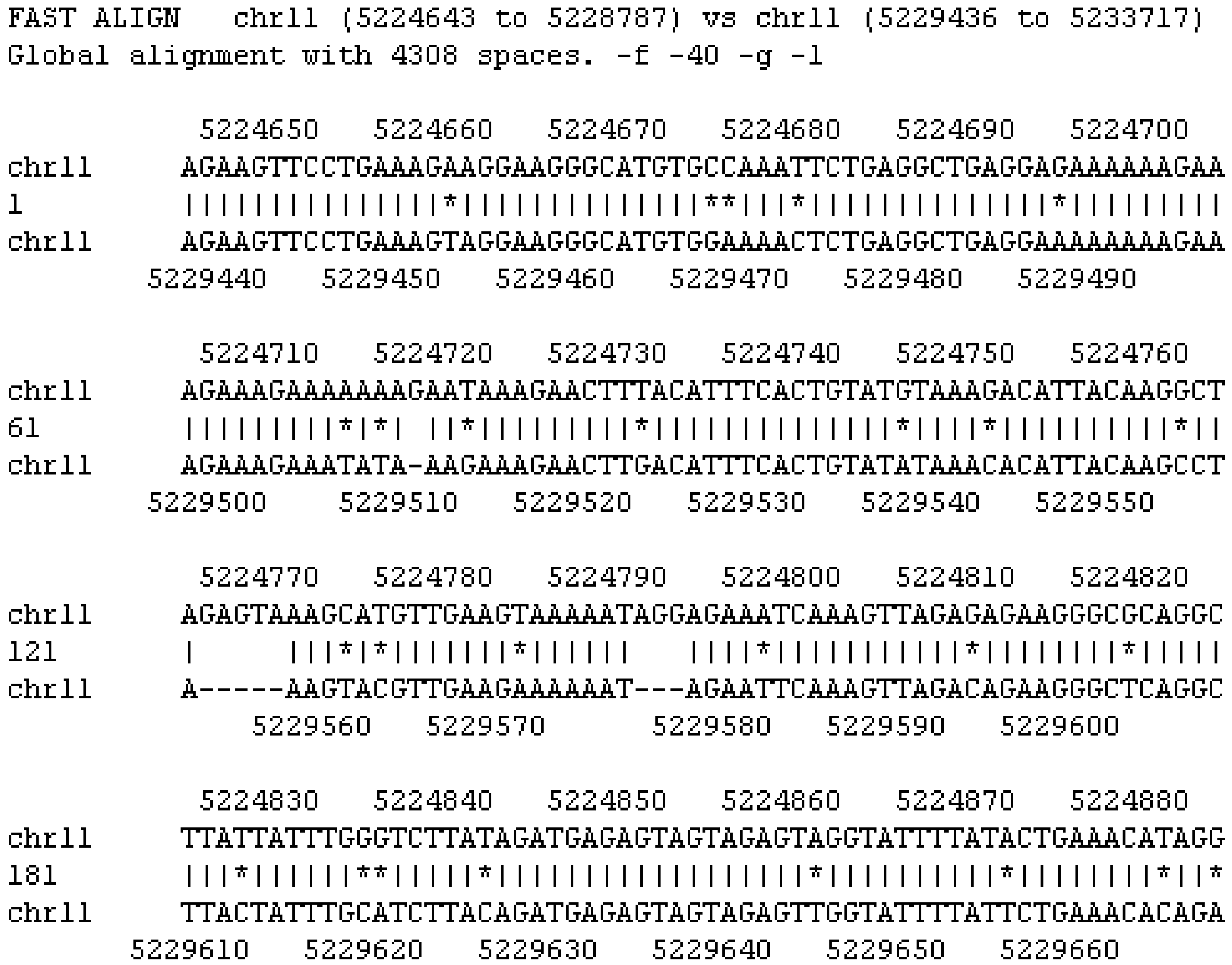
UCSC Genome Browser on Human May 2004 Assembly

position/search chr11:5,200,001-5,250,000

size 50,000 bp.

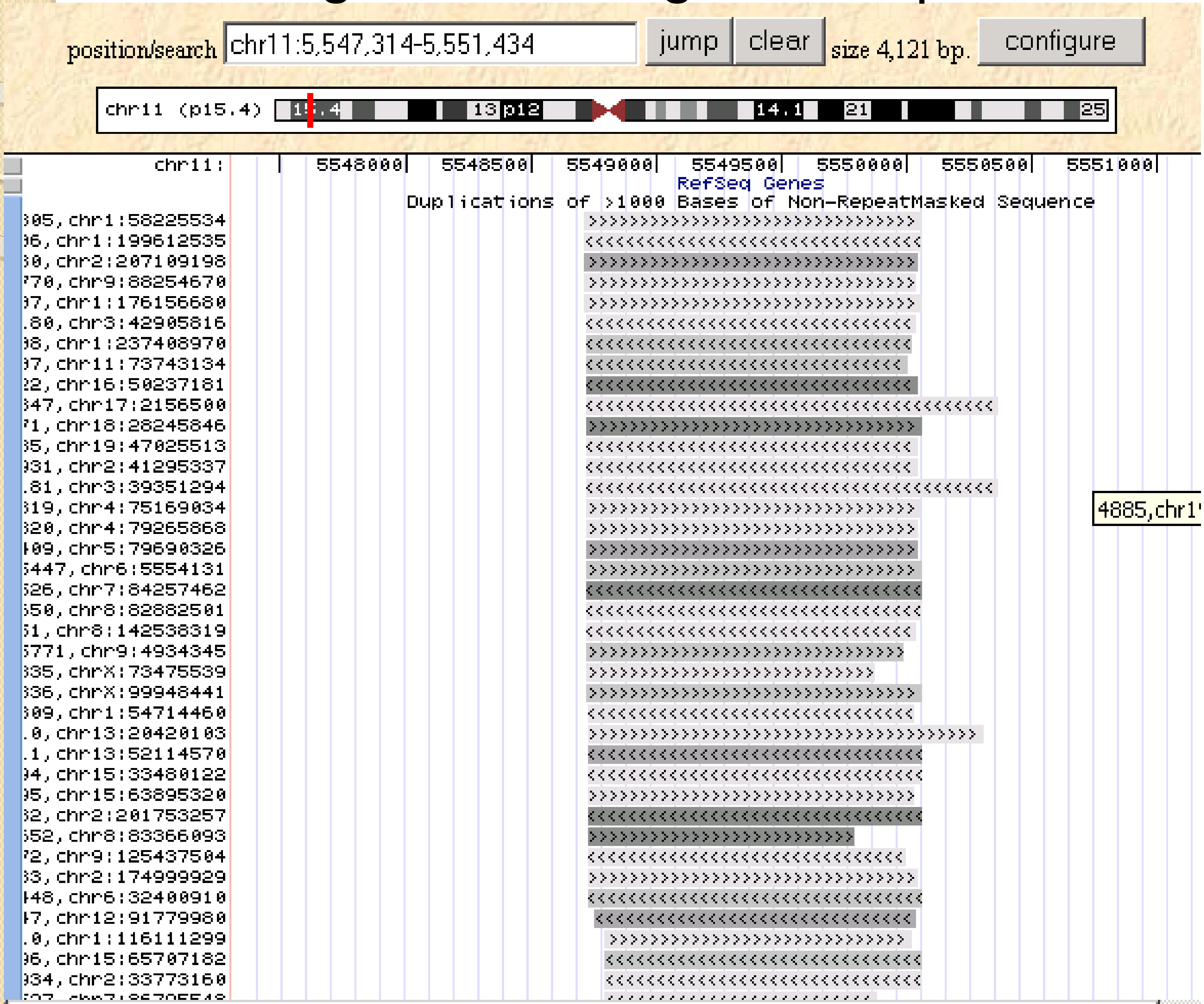
- Light to dark gray: 90 - 98% similarity
- Light to dark yellow: 98 - 99% similarity
- Light to dark orange: greater than 99% similarity
- Red: duplications of greater than 98% similarity that lack sufficient Segmental Duplication Database evidence (most likely missed overlaps)

Beta globin locus: segmental duplications



```
FAST ALIGN   chr11 (5224643 to 5228787) vs chr11 (5229436 to 5233717)
Global alignment with 4308 spaces. -f -40 -g -1

           5224650   5224660   5224670   5224680   5224690   5224700
chr11     AGAAGTTCCTGAAAGAAGGAAGGGCATGTGCCAAATTCTGAGGCTGAGGAGAAAAAAGAA
1         |||||||||||||||*||||||||||||||**|||*||||||||||||||*|||||||||
chr11     AGAAGTTCCTGAAAGTAGGAAGGGCATGTGGAAAACTCTGAGGCTGAGGAAAAAAAAGAA
        5229440   5229450   5229460   5229470   5229480   5229490

           5224710   5224720   5224730   5224740   5224750   5224760
chr11     AGAAAGAAAAAAAGAATAAAGAACTTTACATTTCACTGTATGTAAAGACATTACAAGGCT
61        |||||||||*|*| ||*|||||||||*||||||||||||||*||||*||||||||||*||
chr11     AGAAAGAAATATA-AAGAAAGAACTTGACATTTCACTGTATATAAACACATTACAAGCCT
        5229500    5229510   5229520   5229530   5229540   5229550

           5224770   5224780   5224790   5224800   5224810   5224820
chr11     AGAGTAAAGCATGTTGAAGTAAAAATAGGAGAAATCAAAGTTAGAGAGAAGGGCGCAGGC
121       |     |||*|*|||||||*||||||   ||||*|||||||||||*||||||||*|||||
chr11     A-----AAGTACGTTGAAGAAAAAAT---AGAATTCAAAGTTAGACAGAAGGGCTCAGGC
              5229560   5229570      5229580   5229590   5229600

           5224830   5224840   5224850   5224860   5224870   5224880
chr11     TTATTATTTGGGTCTTATAGATGAGAGTAGTAGAGTAGGTATTTTATACTGAAACATAGG
181       |||*||||||**|||||*||||||||||||||||||*||||||||||*||||||||*||*
chr11     TTACTATTTGCATCTTACAGATGAGAGTAGTAGAGTTGGTATTTTATTCTGAAACACAGA
       5229610   5229620   5229630   5229640   5229650   5229660
```

Beta globin locus: segmental duplications



position/search chr11:5,547,314-5,551,434 jump clear size 4,121 bp. configure

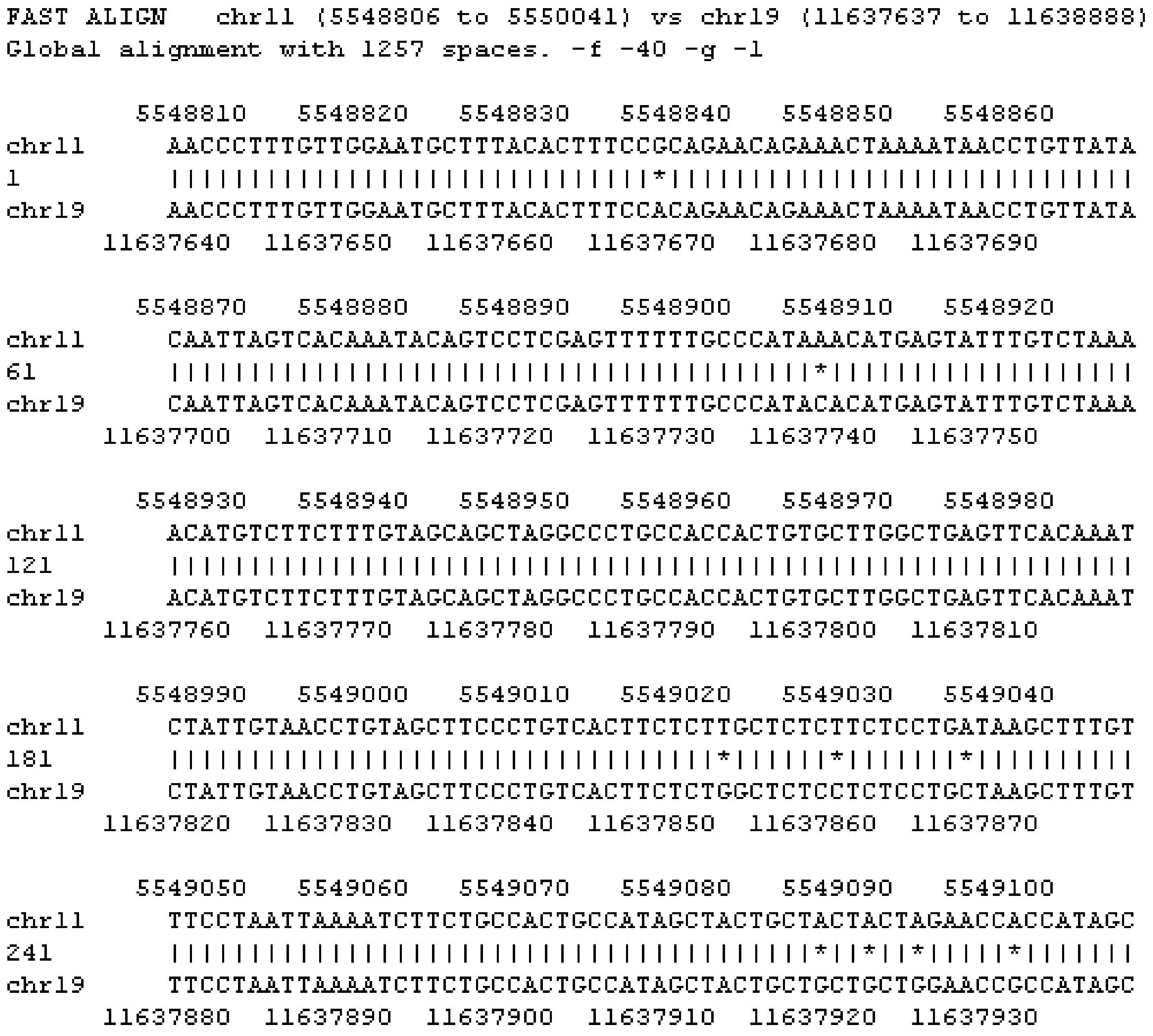
RefSeq Genes

Duplications of >1000 Bases of Non-RepeatMasked Sequence

```
05,chr1:58225534
6,chr1:199612535
0,chr2:207109198
70,chr9:88254670
7,chr1:176156680
.80,chr3:42905816
8,chr1:237408970
7,chr11:73743134
2,chr16:50237181
47,chr17:2156500
1,chr18:28245846
5,chr19:47025513
31,chr2:41295337
.81,chr3:39351294
19,chr4:75169034
20,chr4:79265868
09,chr5:79690326
447,chr6:5554131
26,chr7:84257462
50,chr8:82882501
1,chr8:142538319
771,chr9:4934345
35,chrX:73475539
36,chrX:99948441
09,chr1:54714460
.0,chr13:20420103
.1,chr13:52114570
4,chr15:33480122
5,chr15:63895320
2,chr2:201753257
52,chr8:83366093
2,chr9:125437504
3,chr2:174999929
48,chr6:32400910
7,chr12:91779980
.0,chr1:116111299
6,chr15:65707182
34,chr2:33773160
```

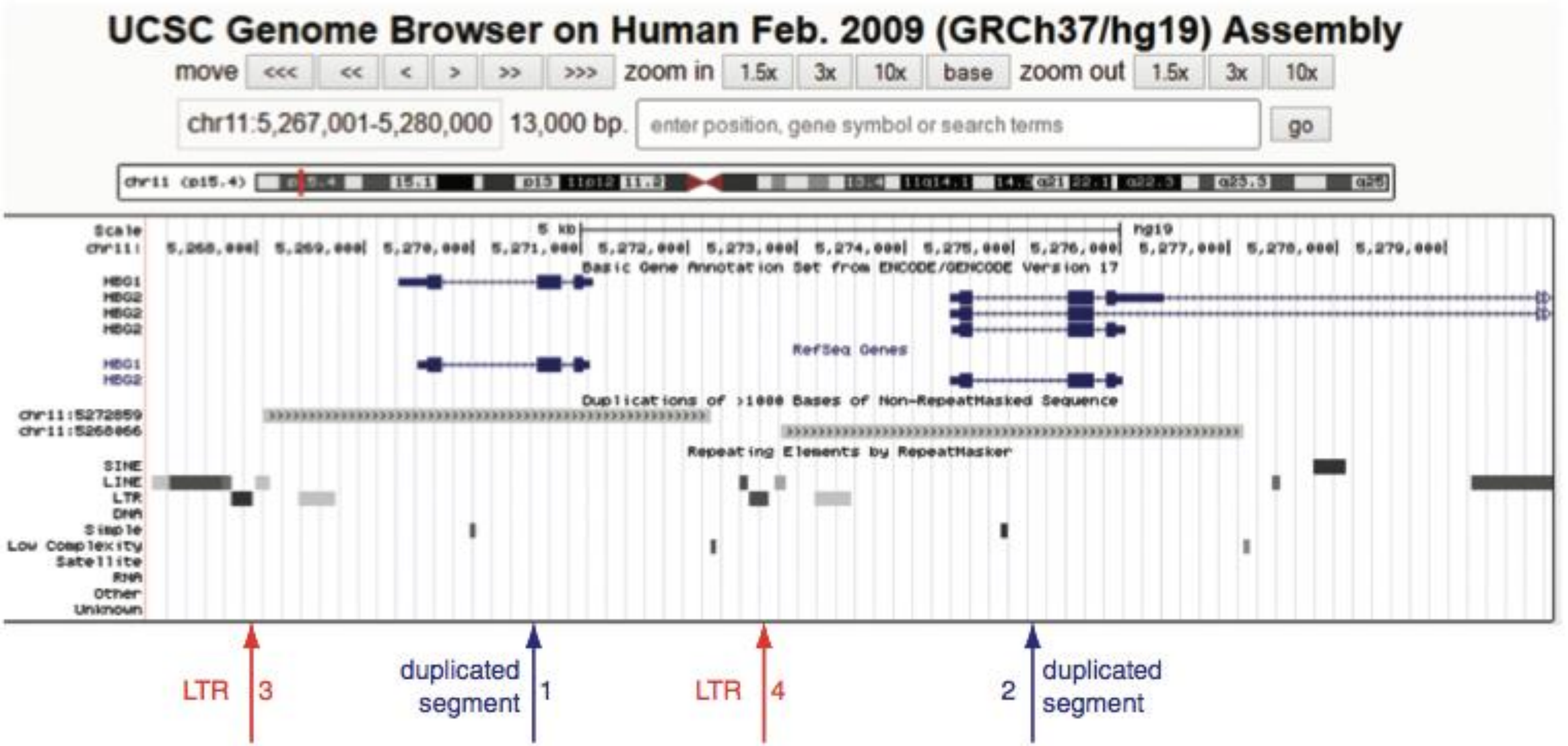
4885,chr1

Beta globin locus: segmental duplications



```
FAST ALIGN   chr11 (5548806 to 5550041) vs chr19 (11637637 to 11638888)
Global alignment with 1257 spaces. -f -40 -g -1

        5548810   5548820   5548830   5548840   5548850   5548860
chr11     AACCCTTTGTTGGAATGCTTTACACTTTCCGCAGAACAGAAACTAAAATAACCTGTTATA
1         ||||||||||||||||||||||||||||||*|||||||||||||||||||||||||||||
chr19     AACCCTTTGTTGGAATGCTTTACACTTTCCACAGAACAGAAACTAAAATAACCTGTTATA
      11637640  11637650  11637660  11637670  11637680  11637690

        5548870   5548880   5548890   5548900   5548910   5548920
chr11     CAATTAGTCACAAATACAGTCCTCGAGTTTTTTGCCCATAAACATGAGTATTTGTCTAAA
61        ||||||||||||||||||||||||||||||||||||||||*|||||||||||||||||||
chr19     CAATTAGTCACAAATACAGTCCTCGAGTTTTTTGCCCATACACATGAGTATTTGTCTAAA
      11637700  11637710  11637720  11637730  11637740  11637750

        5548930   5548940   5548950   5548960   5548970   5548980
chr11     ACATGTCTTCTTTGTAGCAGCTAGGCCCTGCCACCACTGTGCTTGGCTGAGTTCACAAAT
121       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
chr19     ACATGTCTTCTTTGTAGCAGCTAGGCCCTGCCACCACTGTGCTTGGCTGAGTTCACAAAT
      11637760  11637770  11637780  11637790  11637800  11637810

        5548990   5549000   5549010   5549020   5549030   5549040
chr11     CTATTGTAACCTGTAGCTTCCCTGTCACTTCTCTTGCTCTCTTCTCCTGATAAGCTTTGT
181       ||||||||||||||||||||||||||||||||||*||||||*|||||||*||||||||||
chr19     CTATTGTAACCTGTAGCTTCCCTGTCACTTCTCTGGCTCTCCTCTCCTGCTAAGCTTTGT
      11637820  11637830  11637840  11637850  11637860  11637870

        5549050   5549060   5549070   5549080   5549090   5549100
chr11     TTCCTAATTAAAATCTTCTGCCACTGCCATAGCTACTGCTACTACTAGAACCACCATAGC
241       ||||||||||||||||||||||||||||||||||||||||*||*||*|||||*|||||||
chr19     TTCCTAATTAAAATCTTCTGCCACTGCCATAGCTACTGCTGCTGCTGGAACCGCCATAGC
      11637880  11637890  11637900  11637910  11637920  11637930
```

Segmental duplications

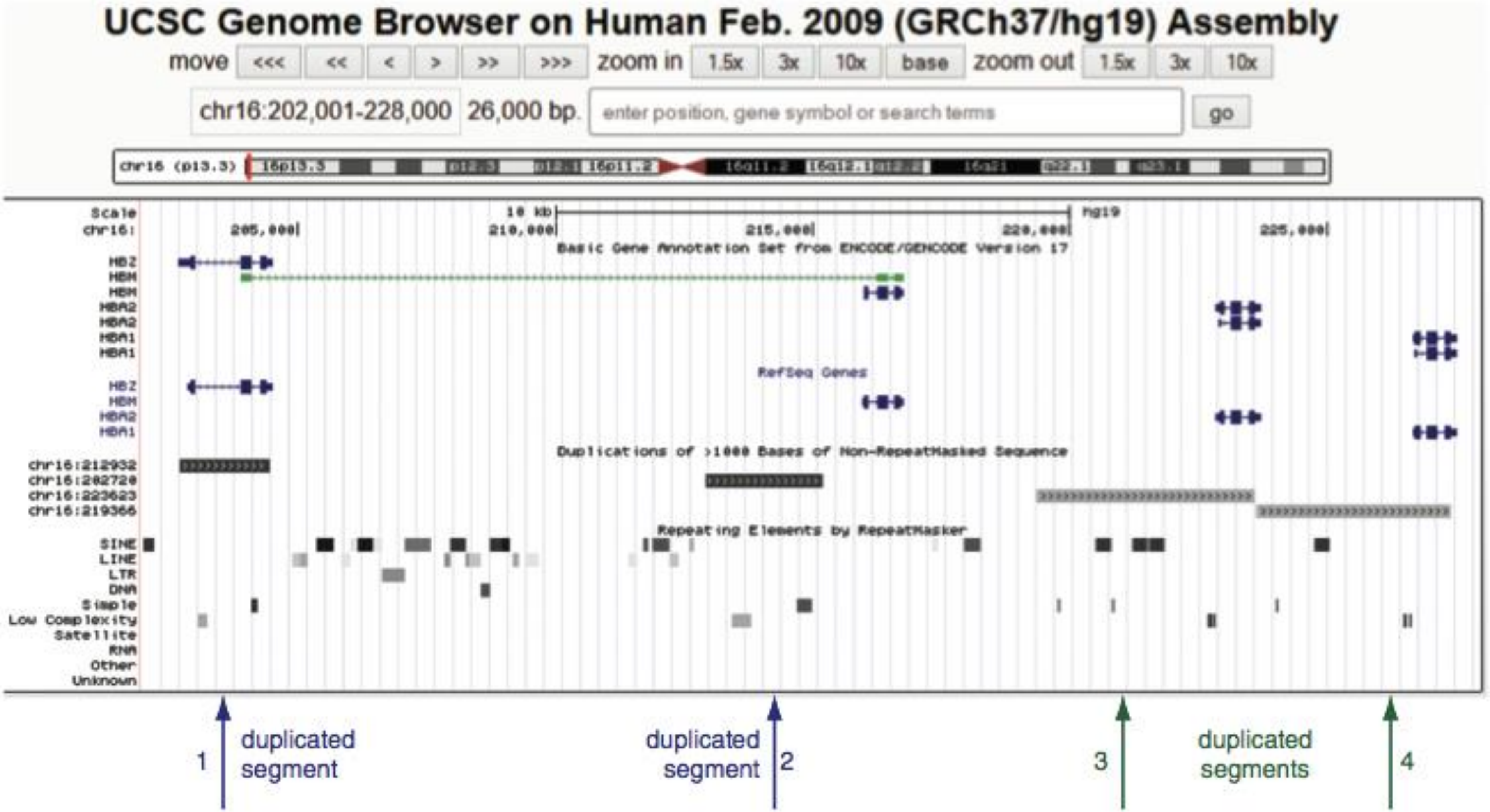


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr11:5,267,001-5,280,000 13,000 bp.

Segmental duplication at the beta globin locus on chromosome 11.

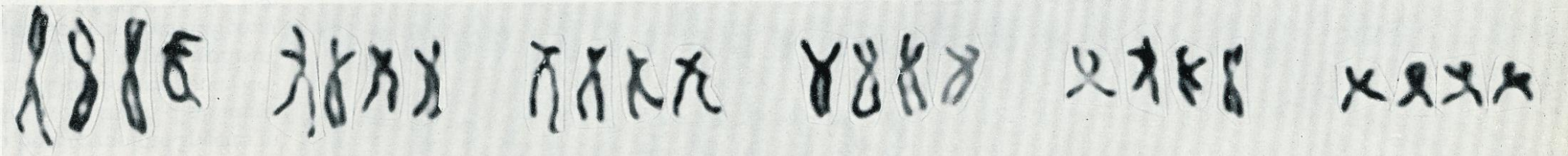
Segmental duplications


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr16:202,001-228,000 26,000 bp.

Segmental duplications at the alpha globin locus on chromosome 16.

Repetitive DNA content of eukaryotic genomes



1. Interspersed repeats
2. Processed pseudogenes
3. Simple sequence repeats
4. Segmental duplications
5. Blocks of tandem repeats

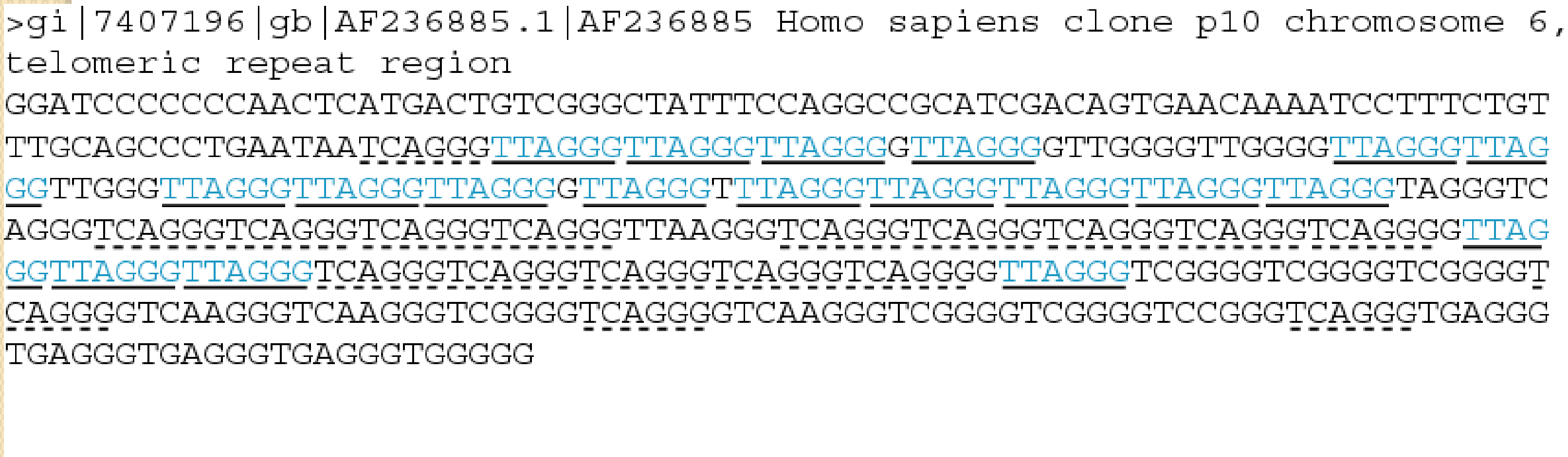
Five main classes of repetitive DNA

5. Blocks of tandem repeats

These include telomeric repeats (e.g. TTAGGG in humans) and centromeric repeats (e.g. a 171 base pair repeat of a satellite DNA in humans).

Such repetitive DNA can span millions of base pairs, and it is often species-specific.

Example of telomeric repeats (obtained by blastn searching $TTAGGG_4$)



```
>gi|7407196|gb|AF236885.1|AF236885 Homo sapiens clone p10 chromosome 6,
telomeric repeat region
GGATCCCCCCCAACTCATGACTGTCGGGCTATTTCCAGGCCGCATCGACAGTGAACAAAATCCTTTCTGT
TTGCAGCCCTGAATAATCAGGGTTAGGGTTAGGGTTAGGGGTTAGGGGTTGGGGTTGGGGTTAGGGTTAG
GGTTGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTAGGGTC
AGGGTCAGGGTCAGGGTCAGGGTCAGGGTTAAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAG
GGTTAGGGTTAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAGGGTCGGGGTCGGGGTCGGGGT
CAGGGGTCAAGGGTCAAGGGTCGGGGTCAGGGGTCAAGGGTCGGGGTCGGGGTCCGGGTCAGGGTGAGGG
TGAGGGTGAGGGTGAGGGTGGGGG
```

Five main classes of repetitive DNA

5. Blocks of tandem repeats

In two exceptional cases, chromosomes lack satellite DNA:

- *Saccharomyces cerevisiae* (very small centromeres)
- Neocentromeres (an ectopic centromere; 60 have been described in human, often associated with disease)

Tandem repeats: telomeric repeats in eukaryotes

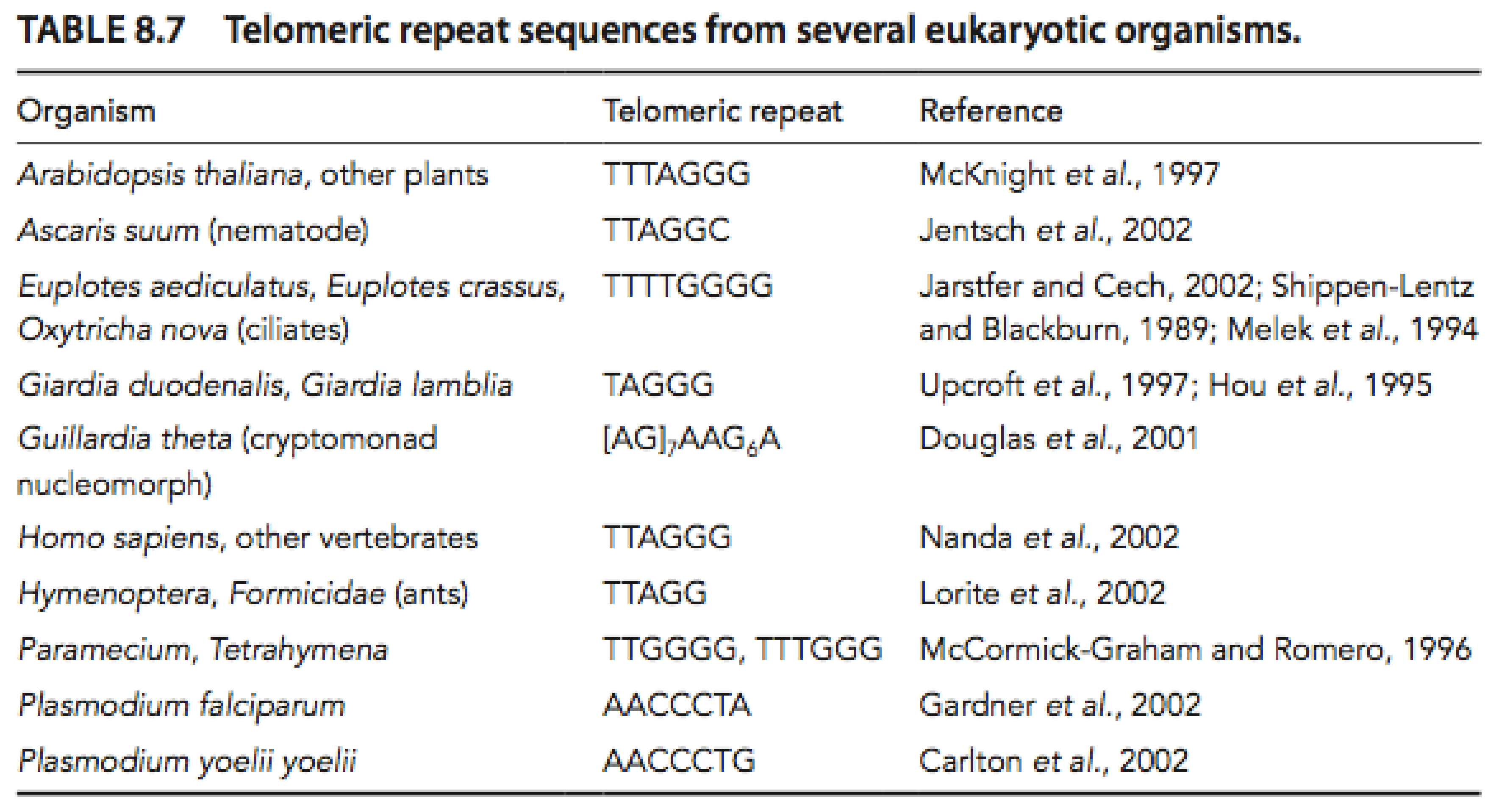
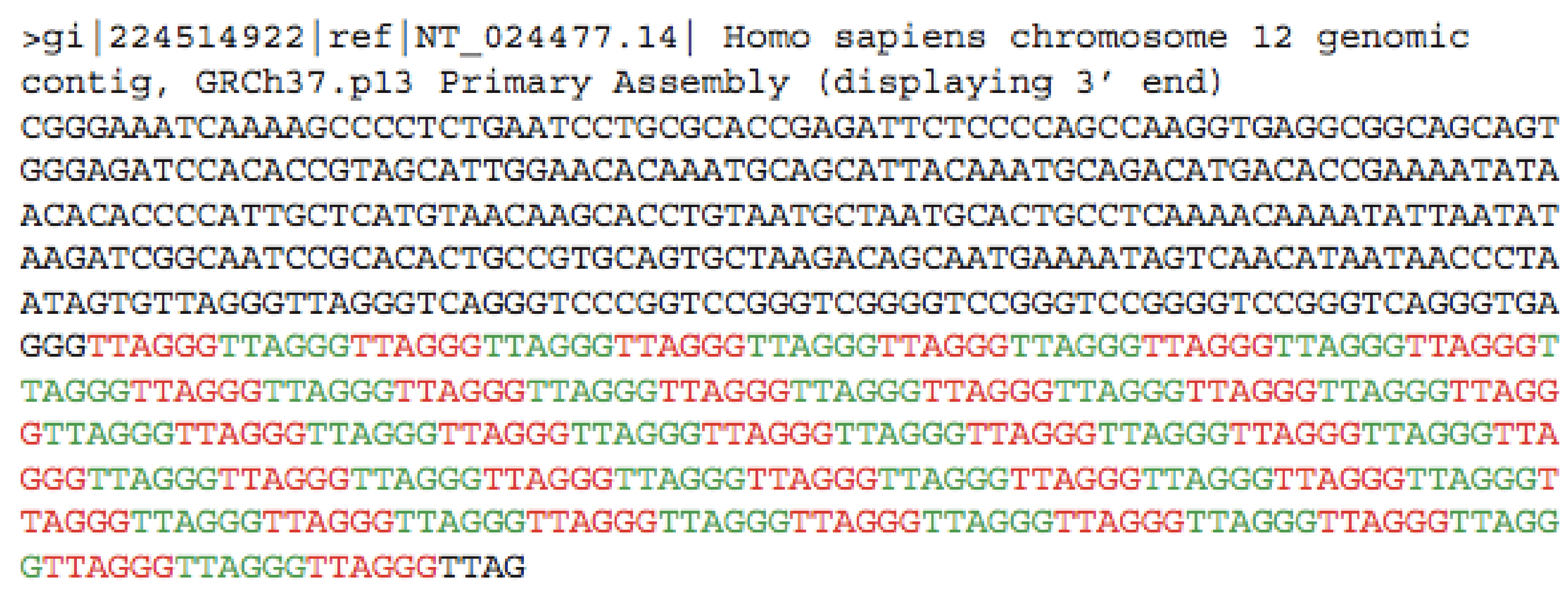


TABLE 8.7 Telomeric repeat sequences from several eukaryotic organisms.

Organism	Telomeric repeat	Reference
Arabidopsis thaliana, other plants	TTTAGGG	McKnight *et al.*, 1997
Ascaris suum (nematode)	TTAGGC	Jentsch *et al.*, 2002
Euplotes aediculatus, *Euplotes crassus*, *Oxytricha nova* (ciliates)	TTTTGGGG	Jarstfer and Cech, 2002; Shippen-Lentz and Blackburn, 1989; Melek *et al.*, 1994
Giardia duodenalis, *Giardia lamblia*	TAGGG	Upcroft *et al.*, 1997; Hou *et al.*, 1995
Guillardia theta (cryptomonad nucleomorph)	$[AG]_7AAG_6A$	Douglas et al., 2001
Homo sapiens, other vertebrates	TTAGGG	Nanda *et al.*, 2002
Hymenoptera, *Formicidae* (ants)	TTAGG	Lorite *et al.*, 2002
Paramecium, *Tetrahymena*	TTGGGG, TTTGGG	McCormick-Graham and Romero, 1996
Plasmodium falciparum	AACCCTA	Gardner *et al.*, 2002
Plasmodium yoelii yoelii	AACCCTG	Carlton *et al.*, 2002

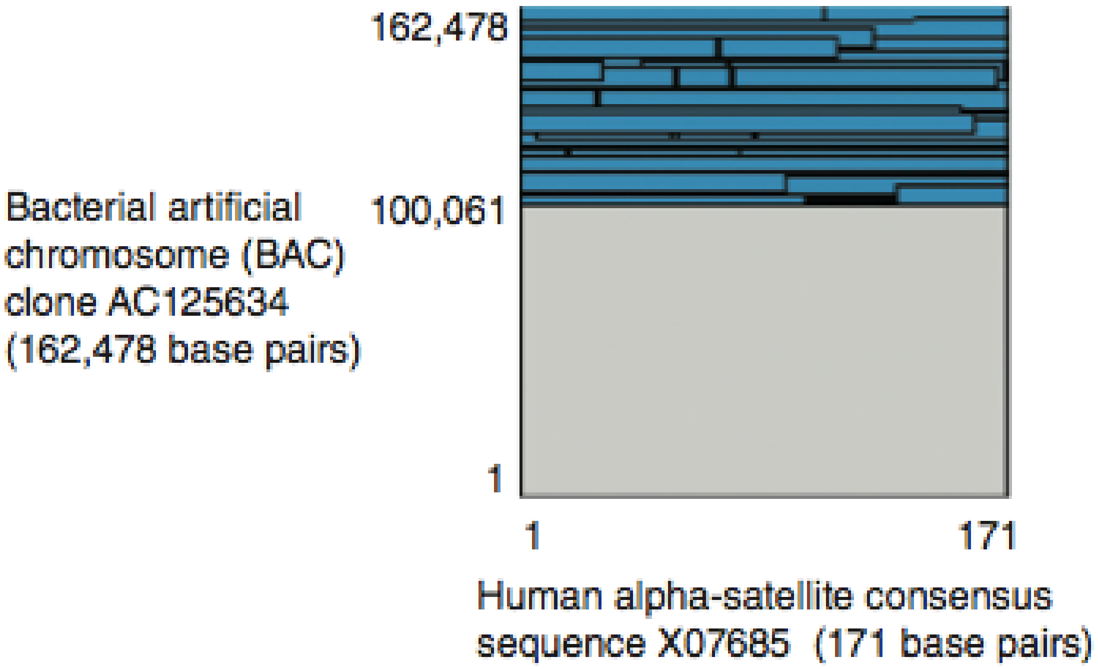
Tandem repeats: TTAGGG in subtelomeric regions



```
>gi|224514922|ref|NT_024477.14| Homo sapiens chromosome 12 genomic
contig, GRCh37.p13 Primary Assembly (displaying 3' end)
CGGGAAATCAAAAGCCCCTCTGAATCCTGCGCACCGAGATTCTCCCCAGCCAAGGTGAGGCGGCAGCAGT
GGGAGATCCACACCGTAGCATTGGAACACAAATGCAGCATTACAAATGCAGACATGACACCGAAAATATA
ACACACCCCATTGCTCATGTAACAAGCACCTGTAATGCTAATGCACTGCCTCAAAACAAAATATTAATAT
AAGATCGGCAATCCGCACACTGCCGTGCAGTGCTAAGACAGCAATGAAAATAGTCAACATAATAACCCTA
ATAGTGTTAGGGTTAGGGTCAGGGTCCCGGTCCGGGTCGGGGTCCGGGTCCGGGGTCCGGGTCAGGGTGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAG
```

A BLASTN search of the human genome database was performed at the NCBI website using TTAGGGTTAGGGTTAGGG as query (i.e., three TTAGGG repeats). There were matches to hundreds of genomic scaffolds. This figure shows an example (NT_024477.14) assigned to the telomere of chromosome 12q.

Repetitive α-satellite DNA in centromeres



A consensus sequence for human α-satellite DNA (X07685) was compared to a BAC clone (AC125634) assigned to a pericentromeric region of chromosome 9q. BLASTN at NCBI was used, and a dotplot is shown.

Outline

Introduction

General features of eukaryotic genomes and chromosomes

C value paradox; organization; genome browsers

Analysis of chromosomes using BioMart and biomaRt

ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

Variation in chromosomal DNA

Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

Finding genes in eukaryotic DNA

Two of the biggest challenges in understanding any eukaryotic genome are

- defining what a gene is, and
- identifying genes within genomic DNA

Finding genes in eukaryotic DNA

Types of genes include

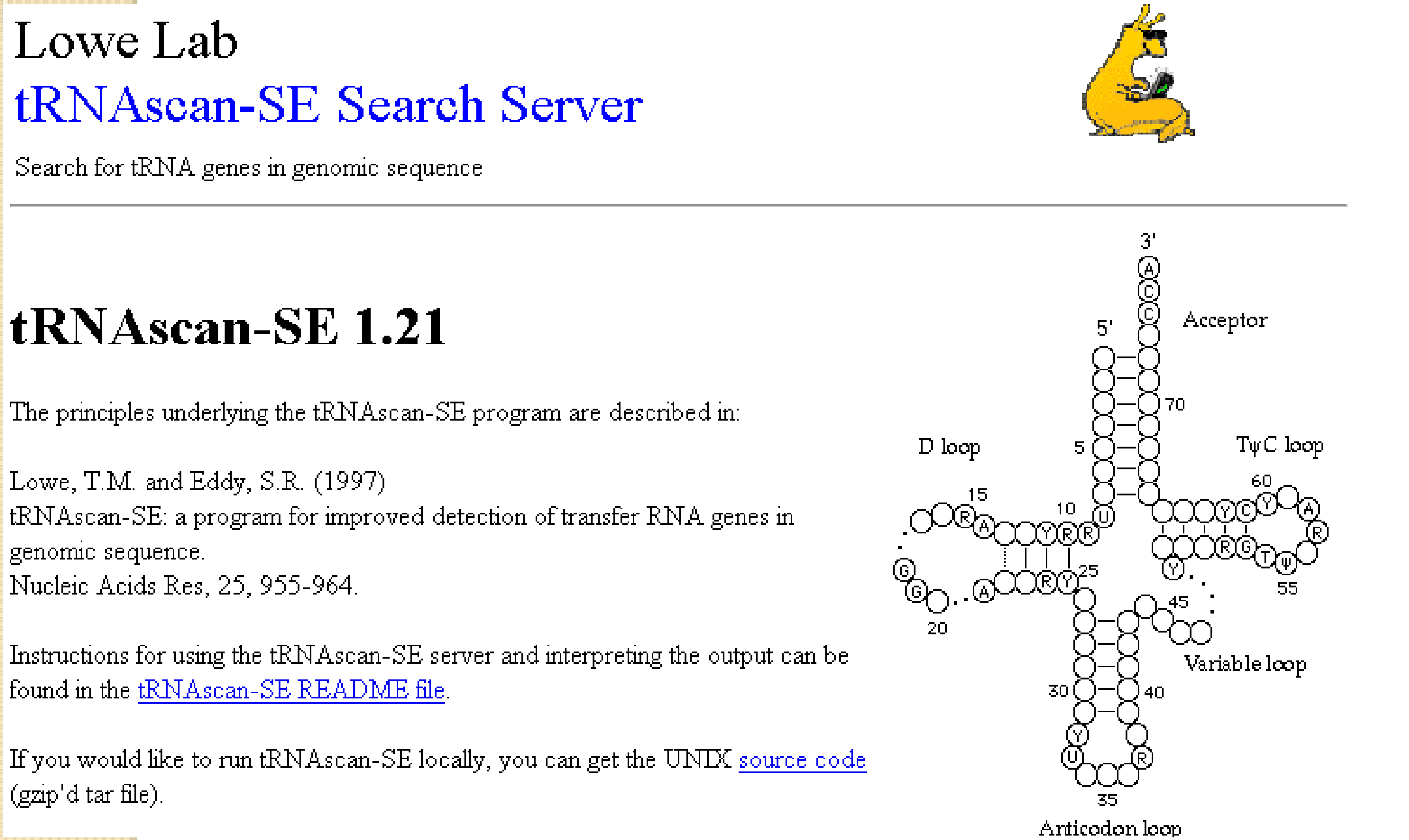
- protein-coding genes
- pseudogenes
- functional RNA genes

--tRNA	transfer RNA
--rRNA	ribosomal RNA
--snoRNA	small nucleolar RNA
--snRNA	small nuclear RNA
--miRNA	microRNA

Finding genes in eukaryotic DNA

RNA genes have diverse and important functions. However, they can be difficult to identify in genomic DNA, because they can be very small, and lack open reading frames that are characteristic of protein-coding genes.

tRNAscan-SE identifies 99 to 100% of tRNA molecules, with a rate of 1 false positive per 15 gigabases.
Visit **http://lowelab.ucsc.edu/tRNAscan-SE/**



Lowe Lab

tRNAscan-SE Search Server

Search for tRNA genes in genomic sequence

tRNAscan-SE 1.21

The principles underlying the tRNAscan-SE program are described in:

Lowe, T.M. and Eddy, S.R. (1997)
tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.
Nucleic Acids Res, 25, 955-964.

Instructions for using the tRNAscan-SE server and interpreting the output can be found in the tRNAscan-SE README file.

If you would like to run tRNAscan-SE locally, you can get the UNIX source code (gzip'd tar file).

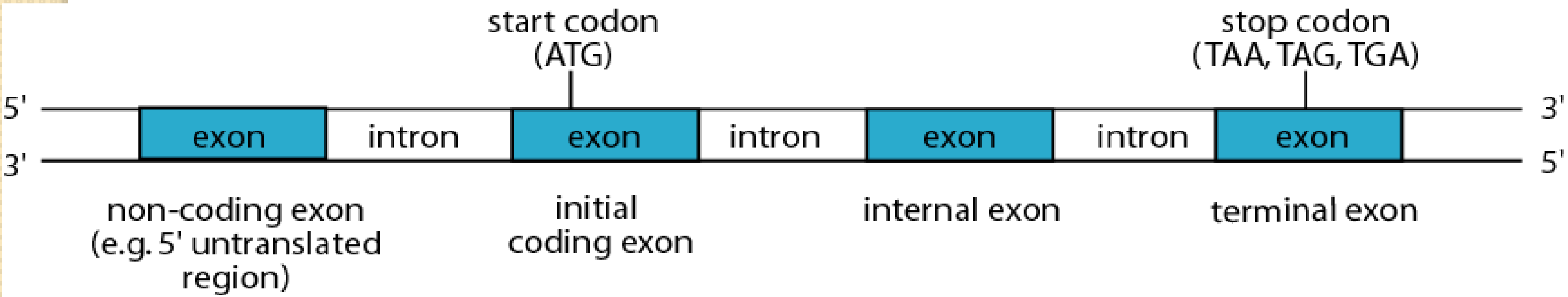
Finding genes in eukaryotic DNA

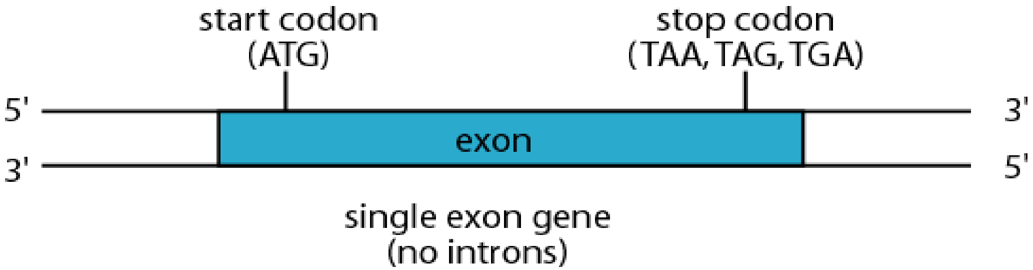
Protein-coding genes are relatively easy to find in prokaryotes, because the gene density is high (about one gene per kilobase). In eukaryotes, gene density is lower, and exons are interrupted by introns.

There are several kinds of exons:

-- noncoding
-- initial coding exons
-- internal exons
-- terminal exons
-- some single-exon genes are intronless

Eukaryotic gene prediction algorithms distinguish several kinds of exons





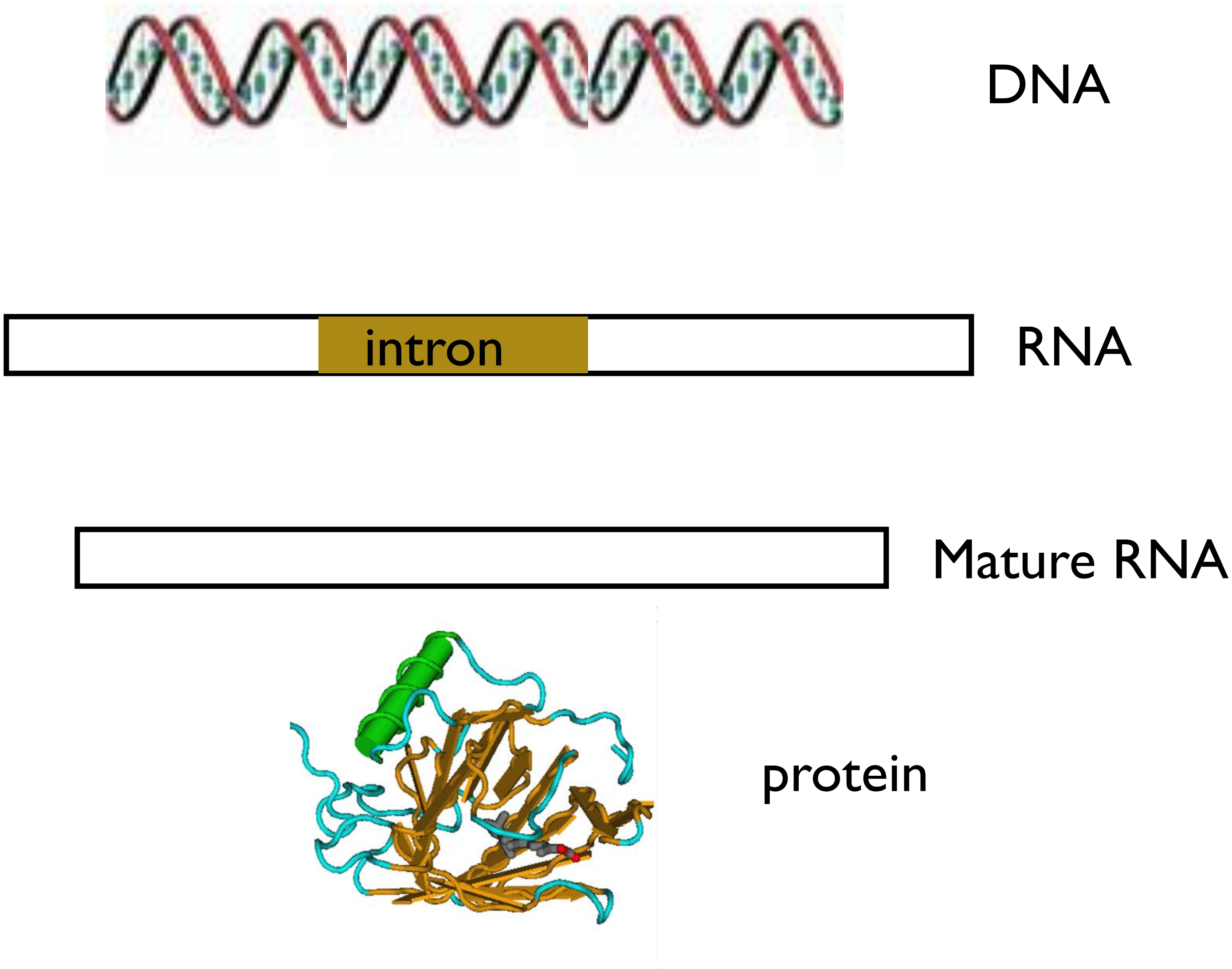
Gene-finding algorithms

Homology-based searches (“extrinsic”)

Rely on previously identified genes

Algorithm-based searches (“intrinsic”)

Investigate nucleotide composition, open-reading frames, and other intrinsic properties of genomic DNA



DNA

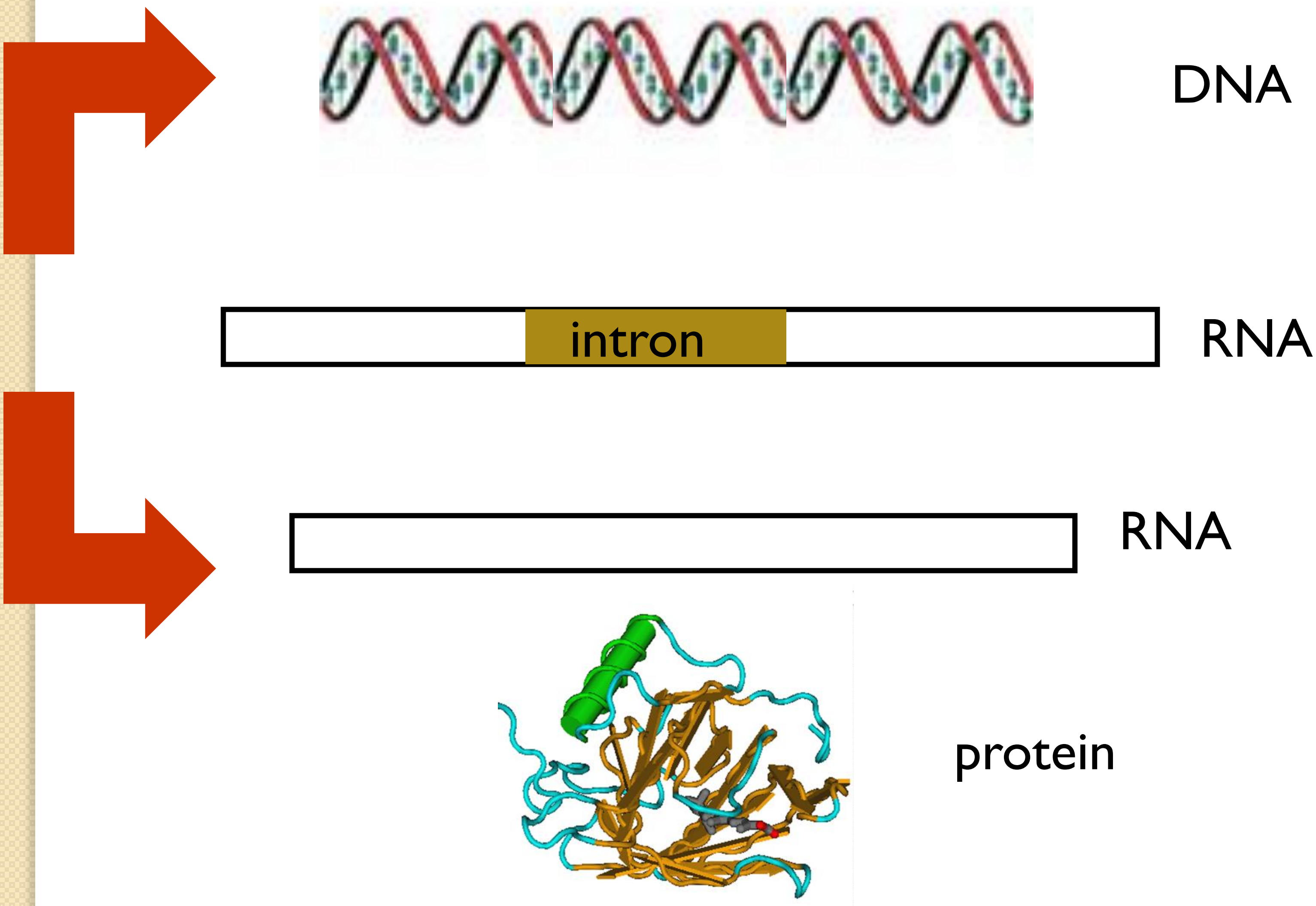
intron

RNA

Mature RNA

protein

Extrinsic, homology-based searching:
compare genomic DNA to expressed genes (ESTs)



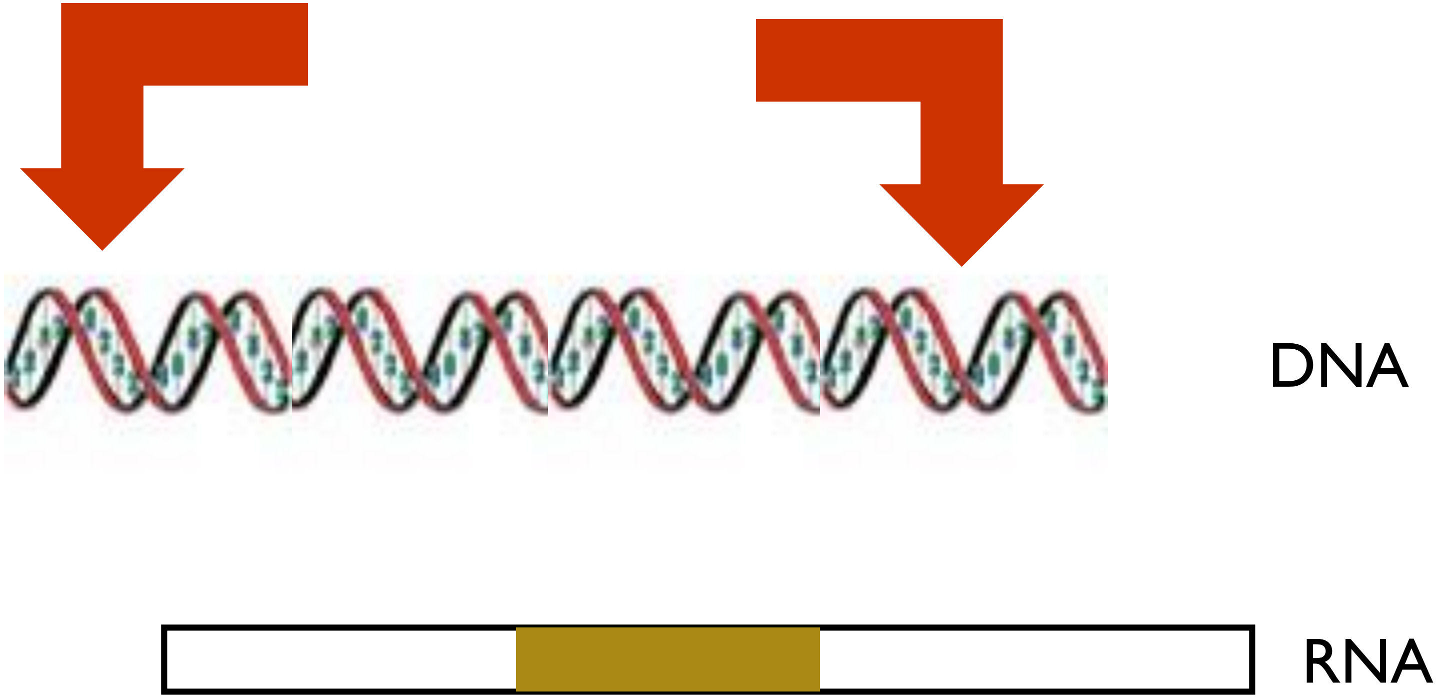
DNA

intron

RNA

RNA

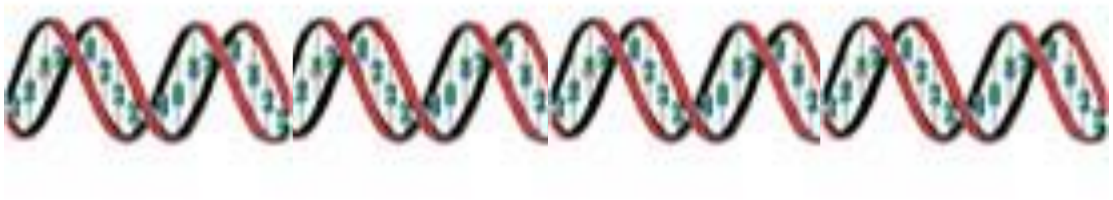
protein



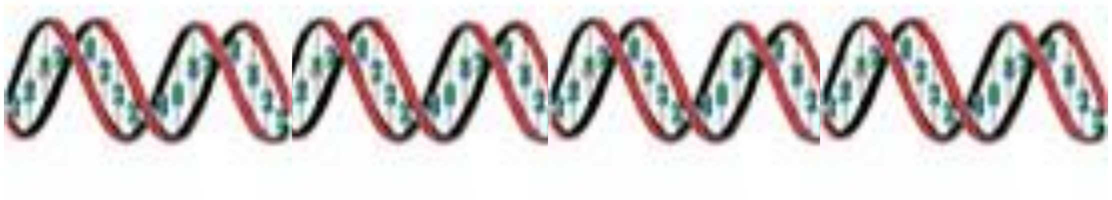
DNA

RNA

Intrinsic, algorithm-based searching:
Identify open reading frames (ORFs).
Compare DNA in exons (unique codon usage)
to DNA in introns (unique splices sites)
and to noncoding DNA.

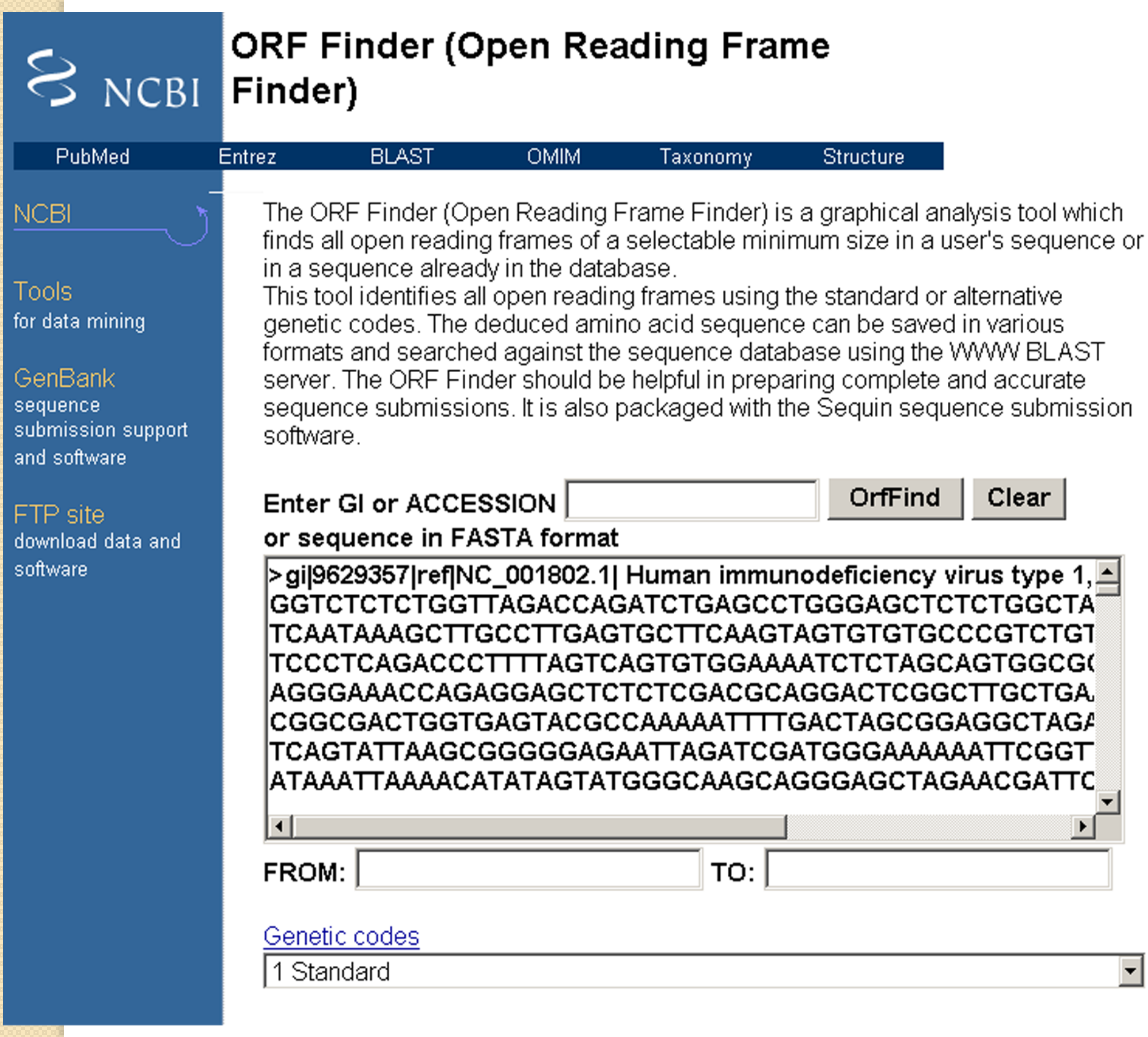


human DNA



chimpanzee DNA

Comparative genomics: Compare gene models between species. (For annotation of the chimpanzee genome reported in 2005, BLAT and BLASTZ searches were used to align the two genomes.)



ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION

or sequence in FASTA format

```
>gi|9629357|ref|NC_001802.1| Human immunodeficiency virus type 1,
GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTA
TCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGCCCGTCTGT
TCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGCG
AGGGAAACCAGAGGAGCTCTCTCGACGCAGGACTCGGCTTGCTGA
CGGCGACTGGTGAGTACGCCAAAAATTTTGACTAGCGGAGGCTAGA
TCAGTATTAAGCGGGGGAGAATTAGATCGATGGGAAAAAATTCGGT
ATAAATTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGATTC
```

FROM: **TO:**

1 Standard

Finding genes in eukaryotic DNA

While ESTs are very helpful in finding genes, beware of several caveats.

-- The quality of EST sequence is sometimes low

-- Highly expressed genes are disproportionately represented in many cDNA libraries

-- ESTs provide no information on genomic location

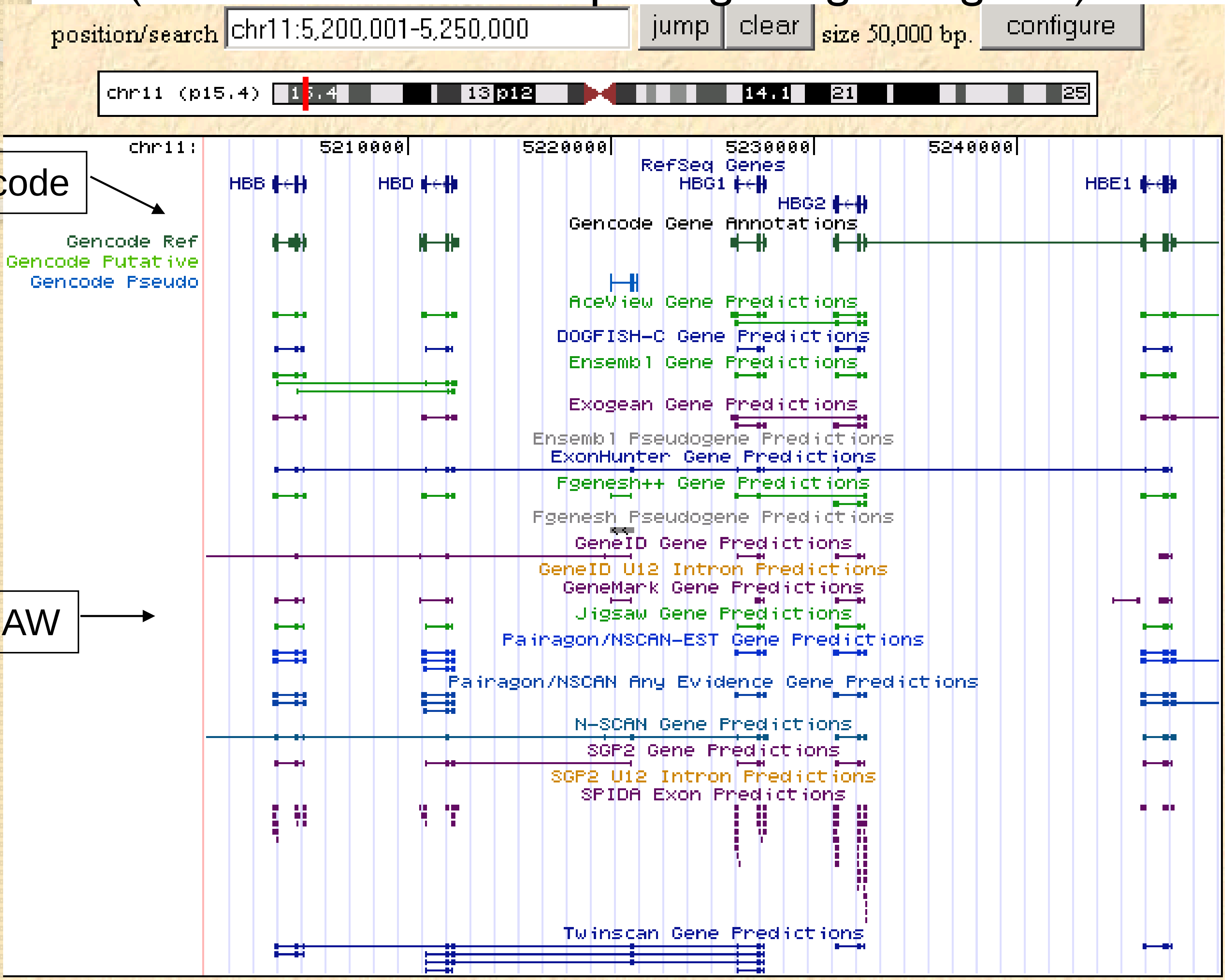
EGASP: the human ENCODE Genome Annotation Assessment Project

EGASP goals:

[1] **Assess of the accuracy of computational methods to predict protein coding genes**. 18 groups competed to make gene predictions, blind; these were evaluated relative to reference annotations generated by the GENCODE project.

[2] **Assess of the completeness of the current human genome annotations** as represented in the ENCODE regions.

UCSC: tracks for Gencode and for various gene prediction algorithms (focus on 50 kb encompassing five globin genes)



position/search chr11:5,200,001-5,250,000

size 50,000 bp.

Gencode

JIGSAW

EGASP: the human ENCODE Genome Annotation Assessment Project

"RESULTS: The best methods had at least one gene transcript correctly predicted for close to 70% of the annotated genes. Nevertheless, the multiple transcript accuracy, taking into account alternative splicing, reached only approximately 40% to 50% accuracy. At the coding nucleotide level, the best programs reached an accuracy of 90% in both sensitivity and specificity. Programs relying on mRNA and protein sequences were the most accurate in reproducing the manually curated annotations. Experimental validation shows that only a very small percentage (3.2%) of the selected 221 computationally predicted exons outside of the existing annotation could be verified."

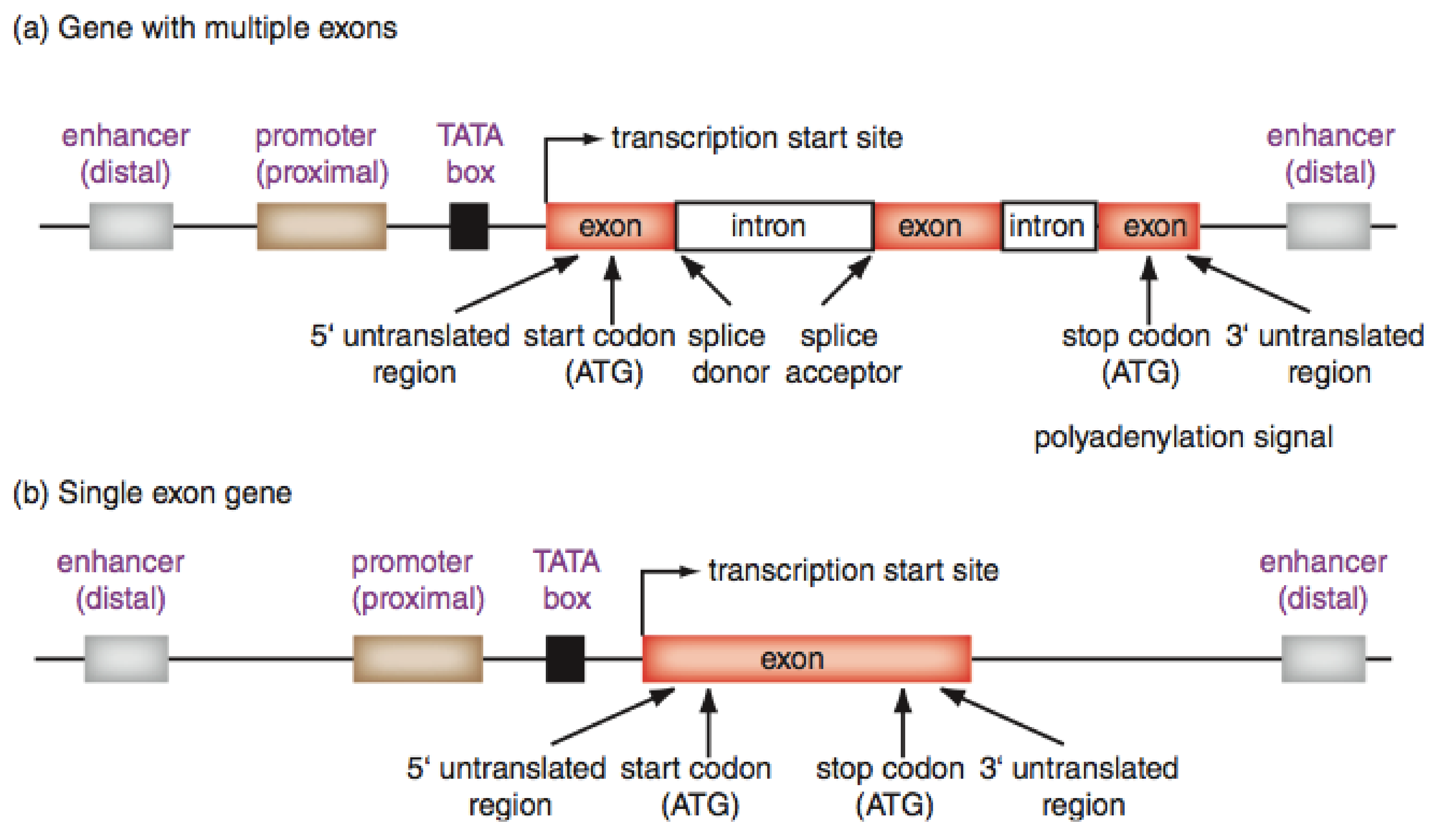
Guigo R et al., *Genome Biology* (2006) 7 Suppl 1: *S*2.1-31

Protein-coding genes in eukaryotic DNA: a new paradox

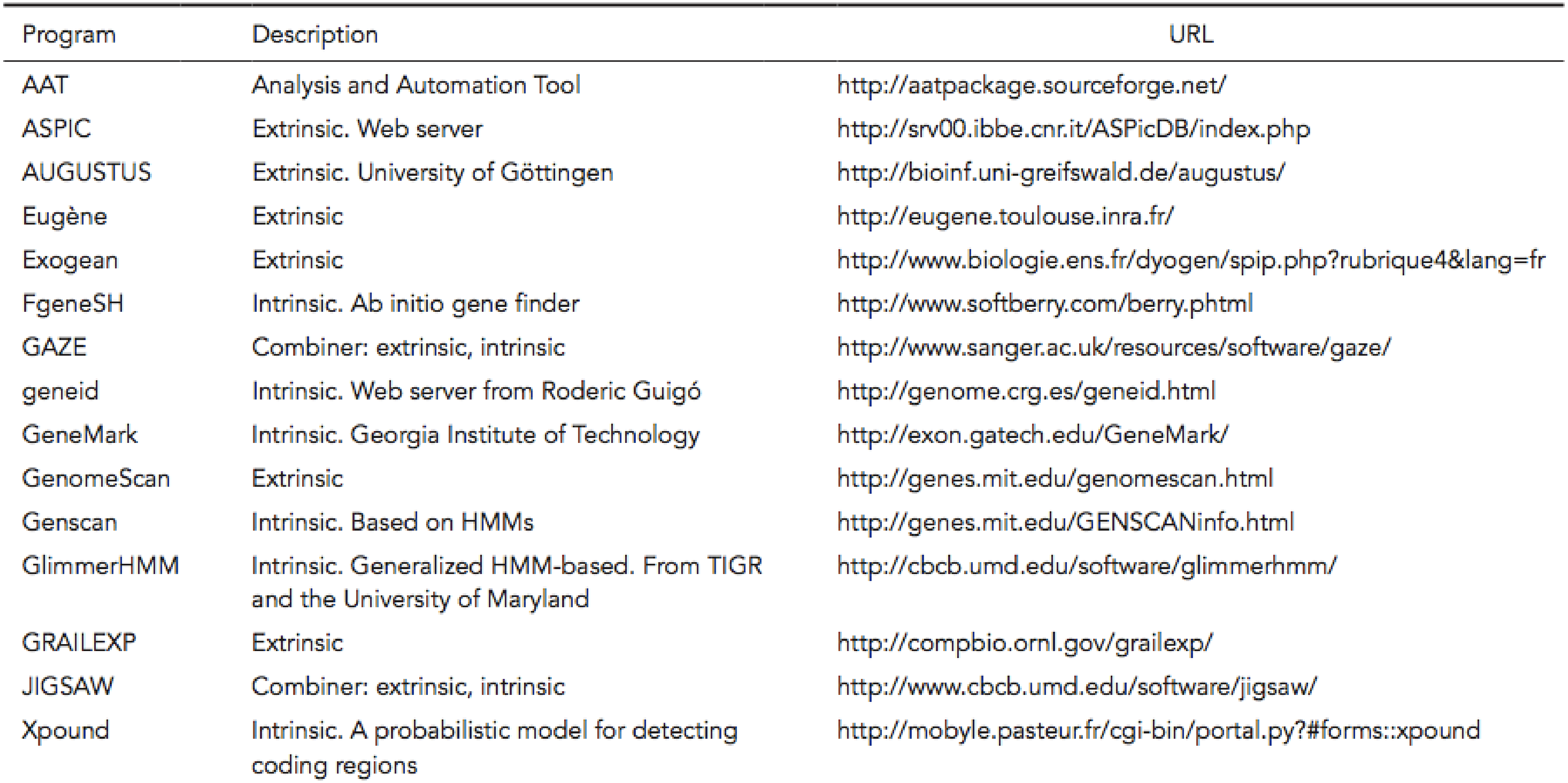
The *C* value paradox is answered by the presence of noncoding DNA.

Why are the number of protein-coding genes about the same for worms, flies, plants, and humans?

Eukaryotic gene prediction algorithms

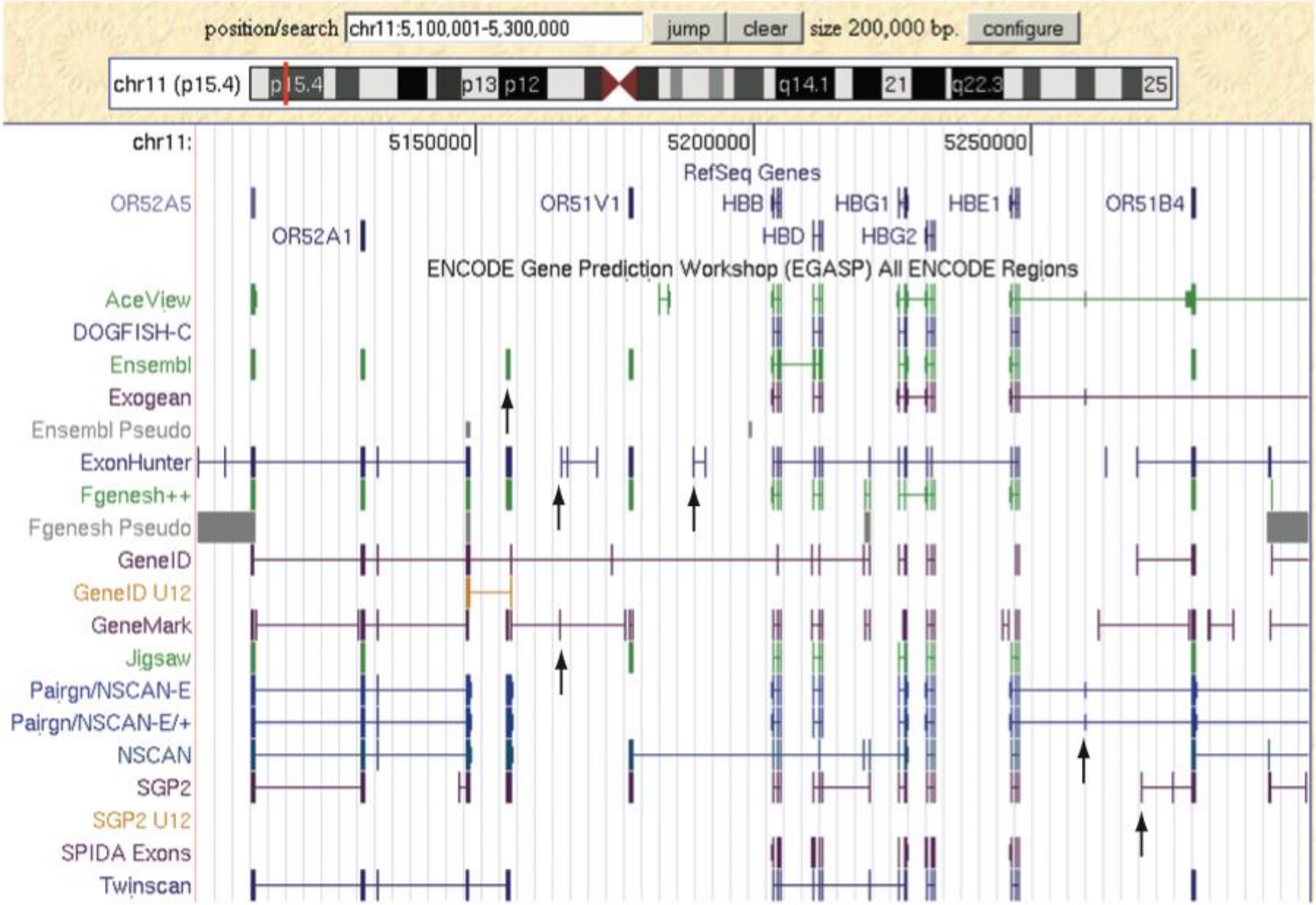


Algorithms for finding genes in eukaryotic DNA

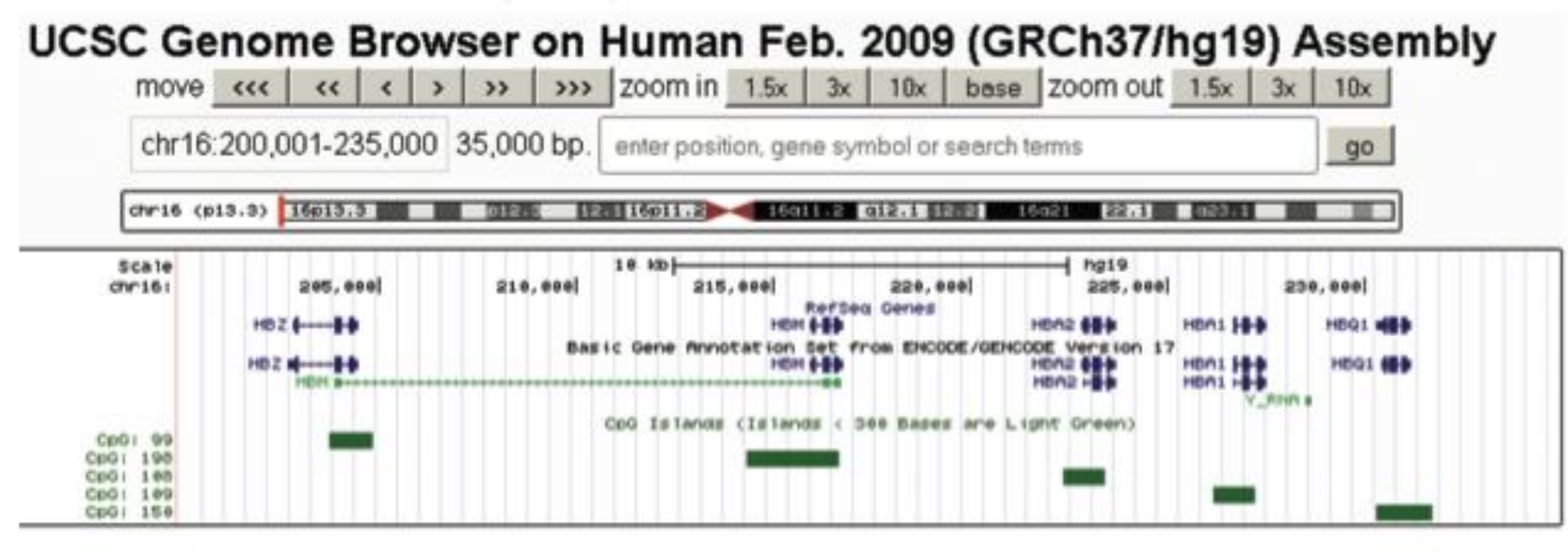


Program	Description	URL
AAT	Analysis and Automation Tool	http://aatpackage.sourceforge.net/
ASPIC	Extrinsic. Web server	http://srv00.ibbe.cnr.it/ASPicDB/index.php
AUGUSTUS	Extrinsic. University of Göttingen	http://bioinf.uni-greifswald.de/augustus/
Eugène	Extrinsic	http://eugene.toulouse.inra.fr/
Exogean	Extrinsic	http://www.biologie.ens.fr/dyogen/spip.php?rubrique4&lang=fr
FgeneSH	Intrinsic. Ab initio gene finder	http://www.softberry.com/berry.phtml
GAZE	Combiner: extrinsic, intrinsic	http://www.sanger.ac.uk/resources/software/gaze/
geneid	Intrinsic. Web server from Roderic Guigó	http://genome.crg.es/geneid.html
GeneMark	Intrinsic. Georgia Institute of Technology	http://exon.gatech.edu/GeneMark/
GenomeScan	Extrinsic	http://genes.mit.edu/genomescan.html
Genscan	Intrinsic. Based on HMMs	http://genes.mit.edu/GENSCANinfo.html
GlimmerHMM	Intrinsic. Generalized HMM-based. From TIGR and the University of Maryland	http://cbcb.umd.edu/software/glimmerhmm/
GRAILEXP	Extrinsic	http://compbio.ornl.gov/grailexp/
JIGSAW	Combiner: extrinsic, intrinsic	http://www.cbcb.umd.edu/software/jigsaw/
Xpound	Intrinsic. A probabilistic model for detecting coding regions	http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::xpound

EGASP: prediction and validation of genes



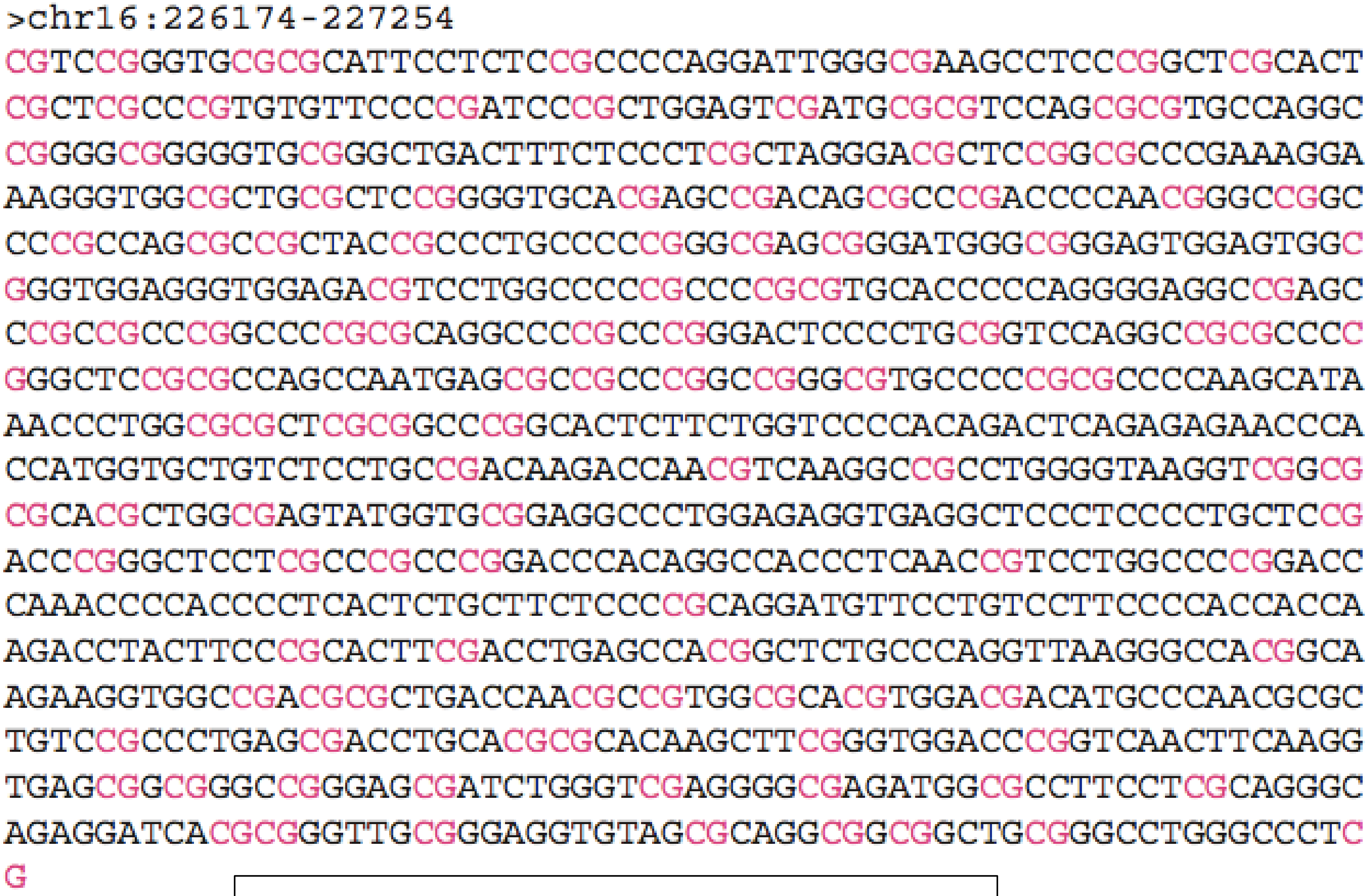
CpG islands are associated with the regulation of expression of many eukaryotic genes



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

CpG islands (green bars) in the human alpha globin gene cluster

CpG islands are associated with the regulation of expression of many eukaryotic genes



```
>chr16:226174-227254
CGTCCGGGTGCGCGCATTCCTCTCCGCCCCAGGATTGGGCGAAGCCTCCCGGCTCGCACT
CGCTCGCCCGTGTGTTCCCCGATCCCGCTGGAGTCGATGCGCGTCCAGCGCGTGCCAGGC
CGGGGCGGGGGTGCGGGCTGACTTTCTCCCTCGCTAGGGACGCTCCGGCGCCCGAAAGGA
AAGGGTGGCGCTGCGCTCCGGGGTGCACGAGCCGACAGCGCCCGACCCCAACGGGCCGGC
CCCGCCAGCGCCGCTACCGCCCTGCCCCCGGGCGAGCGGGATGGGCGGGAGTGGAGTGGC
GGGTGGAGGGTGGAGACGTCCTGGCCCCCGCCCCGCGTGCACCCCCAGGGGAGGCCGAGC
CCGCCGCCCGGCCCCGCGCAGGCCCCGCCCGGGACTCCCCTGCGGTCCAGGCCGCGCCCC
GGGCTCCGCGCCAGCCAATGAGCGCCGCCCGGCCGGGCGTGCCCCCGCGCCCCAAGCATA
AACCCTGGCGCGCTCGCGGCCCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCA
CCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCG
CGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCG
ACCCGGGCTCCTCGCCCGCCCGGACCCACAGGCCACCCTCAACCGTCCTGGCCCCGGACC
CAAACCCCACCCCTCACTCTGCTTCTCCCCGCAGGATGTTCCTGTCCTTCCCCACCACCA
AGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCA
AGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGC
TGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGG
TGAGCGGCGGGCCGGGAGCGATCTGGGTCGAGGGGCGAGATGGCGCCTTCCTCGCAGGGC
AGAGGATCACGCGGGTTGCGGGAGGTGTAGCGCAGGCGGCGGCTGCGGGCCTGGGCCCTC
G
```

CpG island associated with *HBA1*

Outline

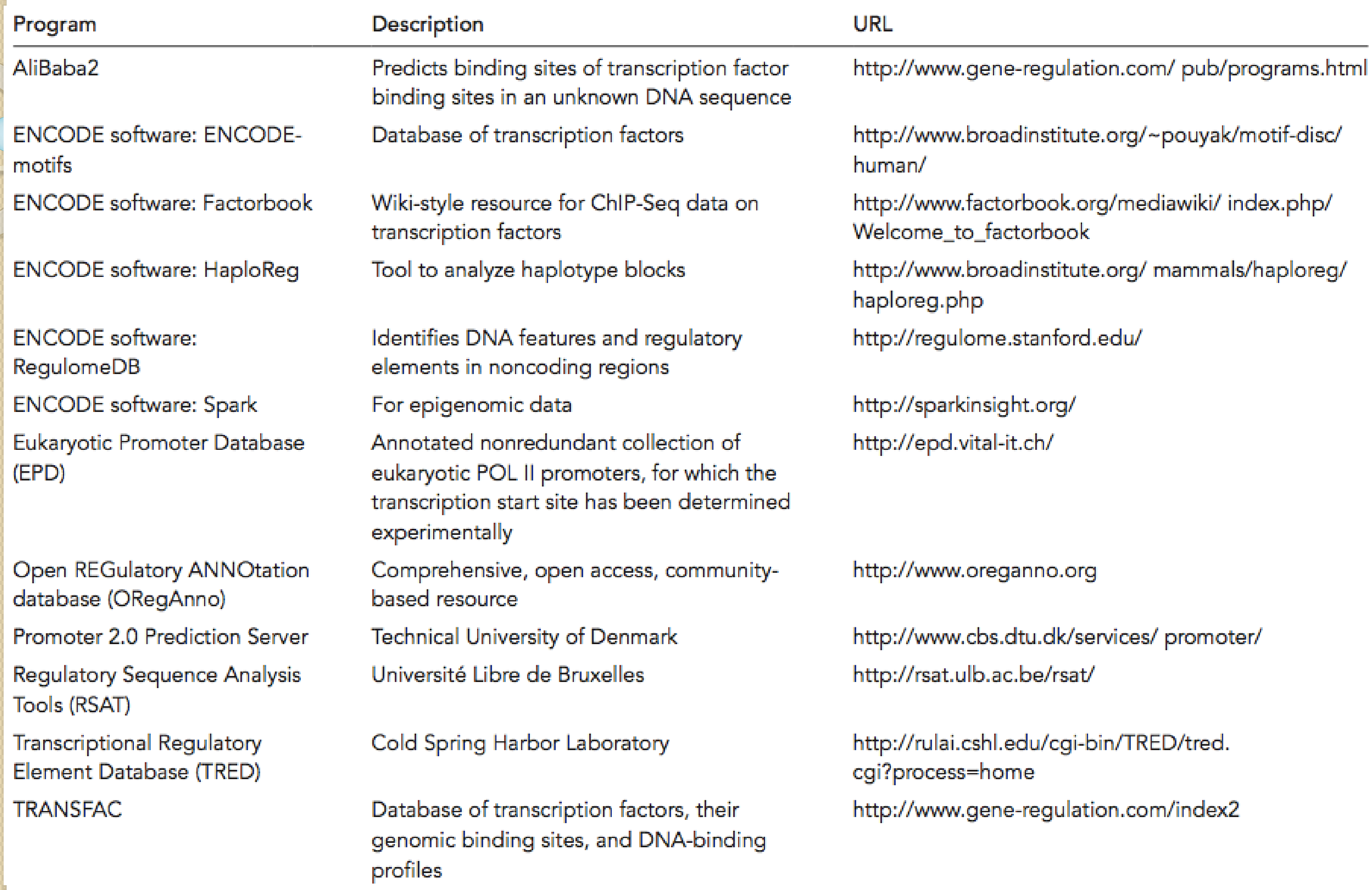
- Introduction
- General features of eukaryotic genomes and chromosomes
    - *C* value paradox; organization; genome browsers
    - Analysis of chromosomes using BioMart and biomaRt
    - ENCODE Project; critiques of ENCODE
- Repetitive DNA content of eukaryotic genomes
    - Noncoding and repetitive DNA sequences
- Gene content of eukaryotic chromosomes
    - Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE
- Regulatory regions of eukaryotic chromosomes
    - Databases of regulatory factors; ultraconserved elements; nonconserved elements
- Comparison of eukaryotic DNA
- Variation in chromosomal DNA
    - Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation
- Techniques to measure chromosomal change
- Perspective

Transcription factor databases

In addition to identifying repetitive elements and genes, it is also of interest to predict the presence of genomic DNA features such as promoter elements and GC content.

Many websites list predictions of transcription factor binding sites and related sequences.

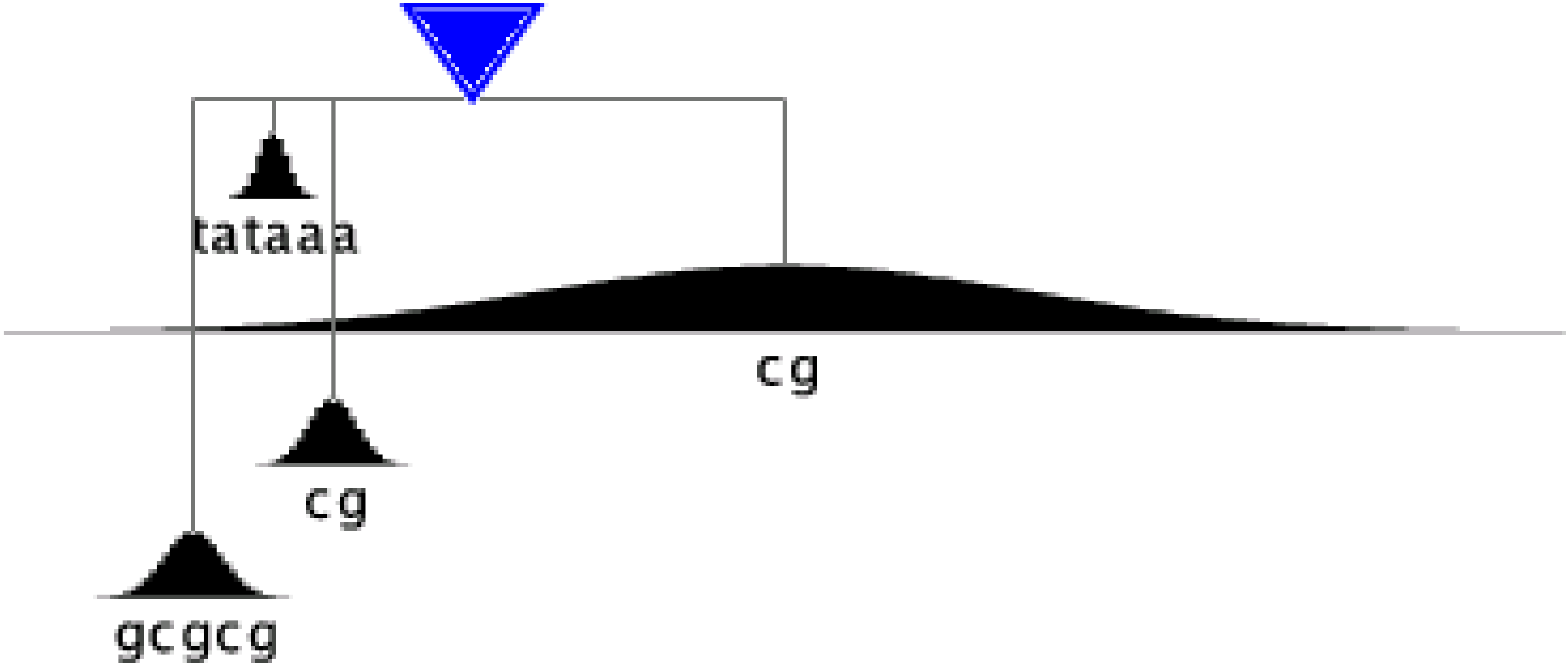
Software for identifying features of promoter regions



Program	Description	URL
AliBaba2	Predicts binding sites of transcription factor binding sites in an unknown DNA sequence	http://www.gene-regulation.com/ pub/programs.html
ENCODE software: ENCODE-motifs	Database of transcription factors	http://www.broadinstitute.org/~pouyak/motif-disc/ human/
ENCODE software: Factorbook	Wiki-style resource for ChIP-Seq data on transcription factors	http://www.factorbook.org/mediawiki/ index.php/ Welcome_to_factorbook
ENCODE software: HaploReg	Tool to analyze haplotype blocks	http://www.broadinstitute.org/ mammals/haploreg/ haploreg.php
ENCODE software: RegulomeDB	Identifies DNA features and regulatory elements in noncoding regions	http://regulome.stanford.edu/
ENCODE software: Spark	For epigenomic data	http://sparkinsight.org/
Eukaryotic Promoter Database (EPD)	Annotated nonredundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally	http://epd.vital-it.ch/
Open REGulatory ANNOtation database (ORegAnno)	Comprehensive, open access, community-based resource	http://www.oreganno.org
Promoter 2.0 Prediction Server	Technical University of Denmark	http://www.cbs.dtu.dk/services/ promoter/
Regulatory Sequence Analysis Tools (RSAT)	Université Libre de Bruxelles	http://rsat.ulb.ac.be/rsat/
Transcriptional Regulatory Element Database (TRED)	Cold Spring Harbor Laboratory	http://rulai.cshl.edu/cgi-bin/TRED/tred. cgi?process=home
TRANSFAC	Database of transcription factors, their genomic binding sites, and DNA-binding profiles	http://www.gene-regulation.com/index2

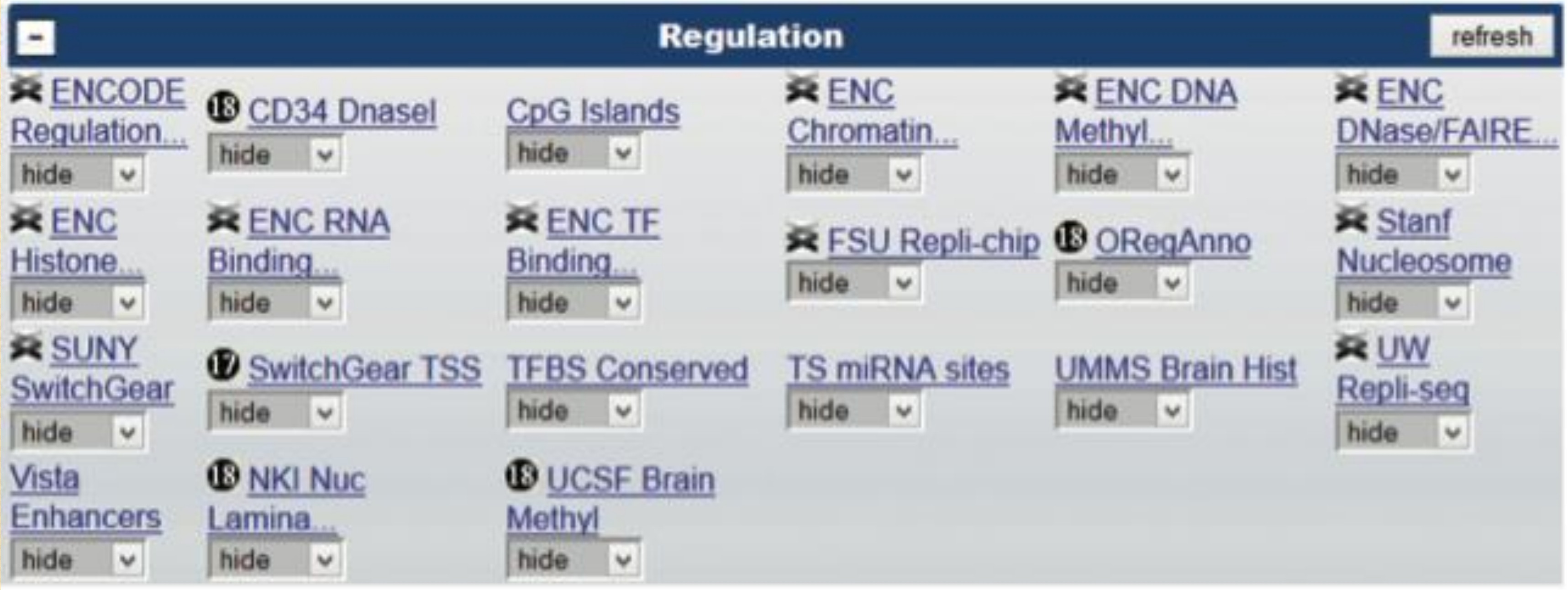
B&FG 3e
Table 8.9
Page 343

Eponine predicts transcription start sites in promoter regions. The algorithm uses a set **of DNA weight matrices** recognizing sequence **motifs** that are associated with a position distribution relative to the transcription start site. The model is as follows:



The specificity is good (~70%), and the positional accuracy is excellent. The program identifies ~50% of TSSs—although it does not always know the direction of transcription.

Regulatory regions in genomic DNA

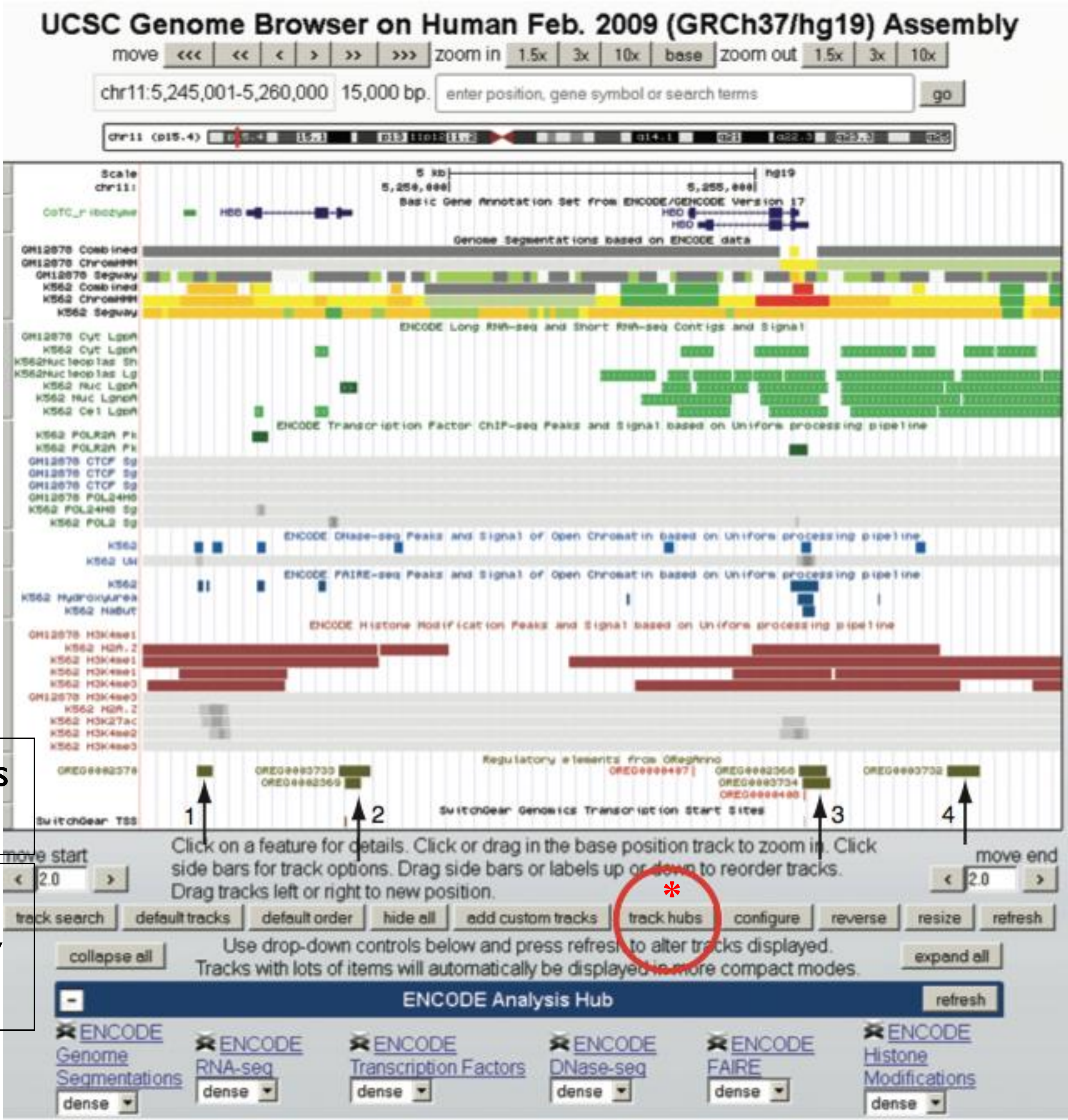


ENCODE Regulation... hide
⑱ CD34 DnaseI hide
CpG Islands hide
ENC Chromatin... hide
ENC DNA Methyl... hide
ENC DNase/FAIRE... hide
ENC Histone... hide
ENC RNA Binding... hide
ENC TF Binding... hide
FSU Repli-chip hide
⑱ ORegAnno hide
Stanf Nucleosome hide
SUNY SwitchGear hide
⑰ SwitchGear TSS hide
TFBS Conserved hide
TS miRNA sites hide
UMMS Brain Hist hide
UW Repli-seq hide
Vista Enhancers hide
⑱ NKI Nuc Lamina... hide
⑱ UCSF Brain Methyl hide

The UCSC Genome (and Table) Browser includes two dozen annotation tracks in the "regulation" category. Explore these!

Regulatory regions in genomic DNA

Beta globin, delta globin region (15 kb at chr11:5,245,001–5,260,000



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr11:5,245,001-5,260,000 15,000 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
enter position, gene symbol or search terms
go
chr11 (p15.4)
Scale
chr11:
5 kb
hg19
5,250,000
5,255,000
Basic Gene Annotation Set from ENCODE/GENCODE Version 17
HBB
HBD
Genome Segmentations based on ENCODE data
GM12878 Combined
GM12878 ChromHMM
GM12878 Segway
K562 Combined
K562 ChromHMM
K562 Segway
ENCODE Long RNA-seq and Short RNA-seq Contigs and Signal
GM12878 Cyt LgpA
K562 Cyt LgpA
K562 Nuc LgpA
K562 Cel LgpA
ENCODE Transcription Factor ChIP-seq Peaks and Signal based on Uniform processing pipeline
K562 POLR2A Pk
GM12878 CTCF Sg
GM12878 POL24H8
K562 POL2 Sg
ENCODE DNase-seq Peaks and Signal of Open Chromatin based on Uniform processing pipeline
K562
K562 UW
ENCODE FAIRE-seq Peaks and Signal of Open Chromatin based on Uniform processing pipeline
K562
K562 Hydroxyurea
K562 NaBut
ENCODE Histone Modification Peaks and Signal based on Uniform processing pipeline
GM12878 H3K4me1
K562 H2A.Z
K562 H3K4me1
K562 H3K4me3
GM12878 H3K4me3
K562 H3K27ac
K562 H3K4me2
Regulatory elements from ORegAnno
SwitchGear Genomics Transcription Start Sites
SwitchGear TSS
1
2
3
4
move start
2.0
move end
2.0
track search
default tracks
default order
hide all
add custom tracks
track hubs
*
configure
reverse
resize
refresh
collapse all
expand all
ENCODE Analysis Hub
refresh
ENCODE Genome Segmentations
dense
ENCODE RNA-seq
dense
ENCODE Transcription Factors
dense
ENCODE DNase-seq
dense
ENCODE FAIRE
dense
ENCODE Histone Modifications
dense

Four regulatory features from ORegAnno

Track hub offers access to ENCODE regulatory data (*)

Outline

- Introduction
- General features of eukaryotic genomes and chromosomes
  - *C* value paradox; organization; genome browsers
  - Analysis of chromosomes using BioMart and biomaRt
  - ENCODE Project; critiques of ENCODE
- Repetitive DNA content of eukaryotic genomes
  - Noncoding and repetitive DNA sequences
- Gene content of eukaryotic chromosomes
  - Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE
- Regulatory regions of eukaryotic chromosomes
  - Databases of regulatory factors; ultraconserved elements; nonconserved elements
- Comparison of eukaryotic DNA
- Variation in chromosomal DNA
  - Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation
- Techniques to measure chromosomal change
- Perspective

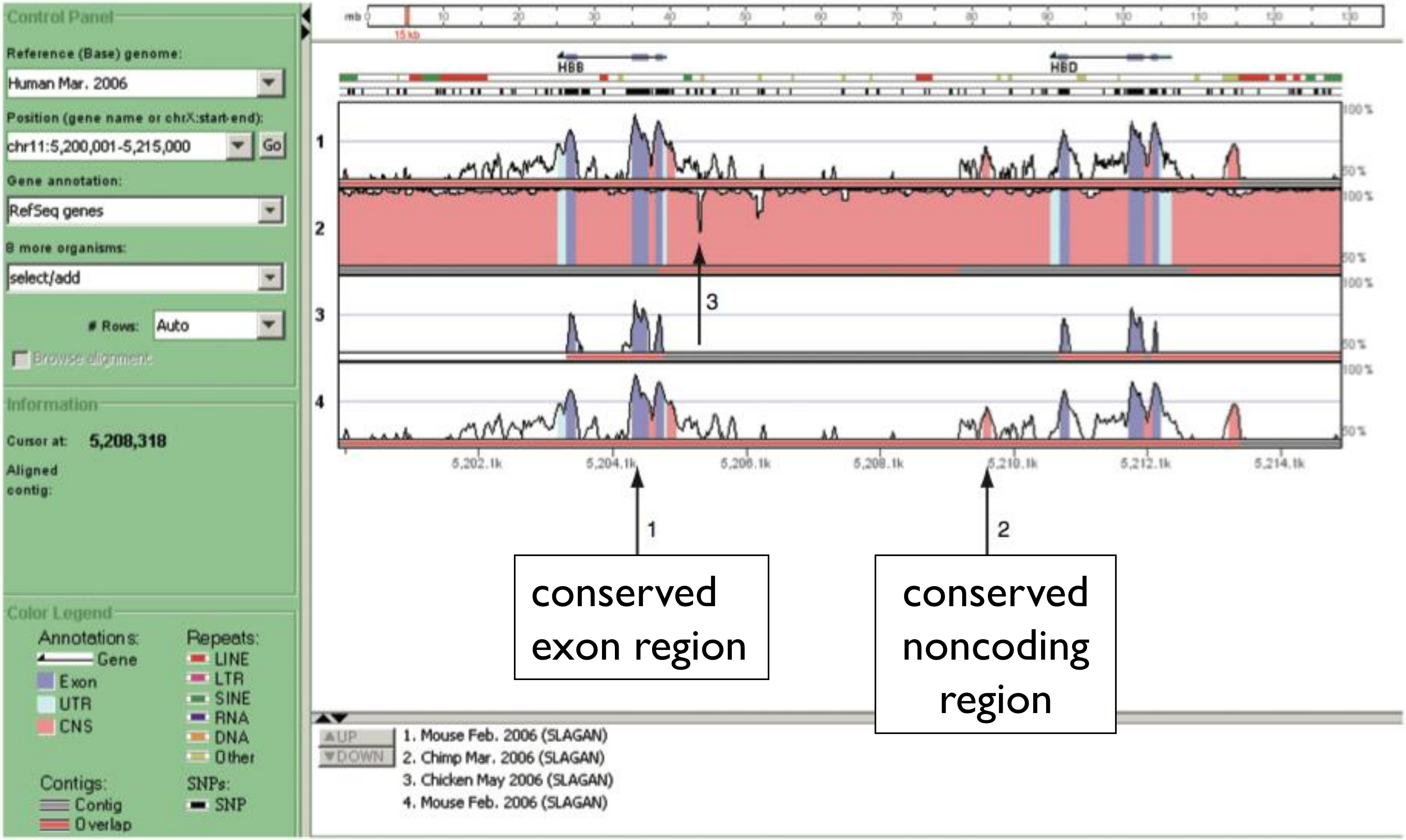
Comparison of eukaryotic DNA: PipMaker and VISTA

In studying genomes, it is important to align large segments of DNA.

PipMaker and VISTA are two tools for sequence alignment and visualization. They show conserved segments, including the order and orientation of conserved elements. They also display large-scale genomic changes (inversions, rearrangements, duplications).

Try VISTA (**http://www-gsd.lbl.gov/vista**) or PipMaker (http://bio.cse.psu.edu/pipmaker) with genomic DNA from Hs10 and Mm19 (containing RBP4).

VISTA for aligning genomic sequences



Control Panel

Reference (Base) genome: Human Mar. 2006

Position (gene name or chrX:start-end): chr11:5,200,001-5,215,000 Go

Gene annotation: RefSeq genes

8 more organisms: select/add

# Rows: Auto

□ Browse alignment

Information

Cursor at: 5,208,318

Aligned contig:

Color Legend

Annotations: Gene, Exon, UTR, CNS

Repeats: LINE, LTR, SINE, RNA, DNA, Other

Contigs: Contig, Overlap

SNPs: SNP

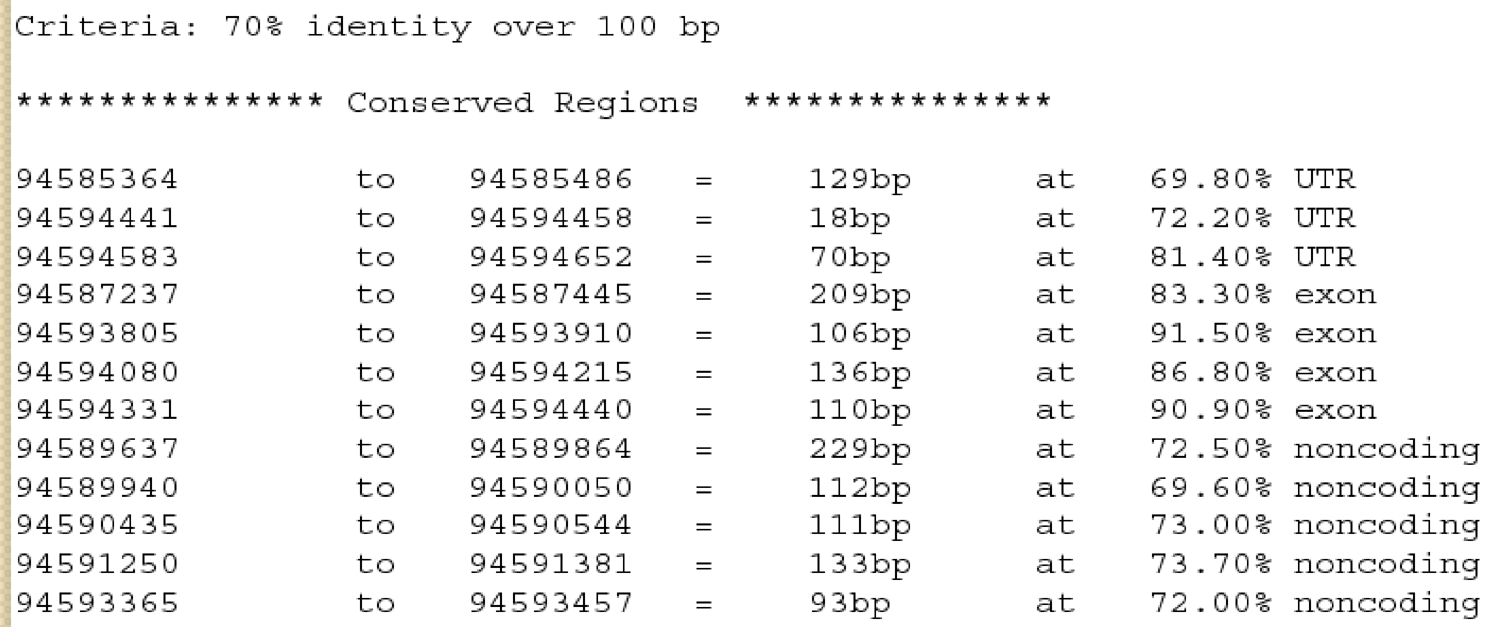
▲UP ▼DOWN

1. Mouse Feb. 2006 (SLAGAN)
2. Chimp Mar. 2006 (SLAGAN)
3. Chicken May 2006 (SLAGAN)
4. Mouse Feb. 2006 (SLAGAN)

conserved exon region

conserved noncoding region

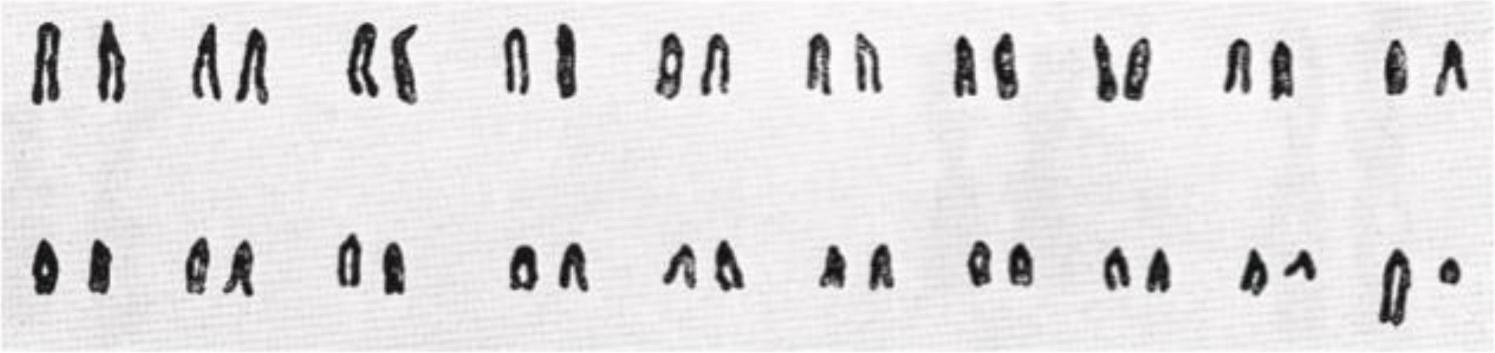
VISTA output for an alignment of human and mouse genomic DNA (including RBP4)



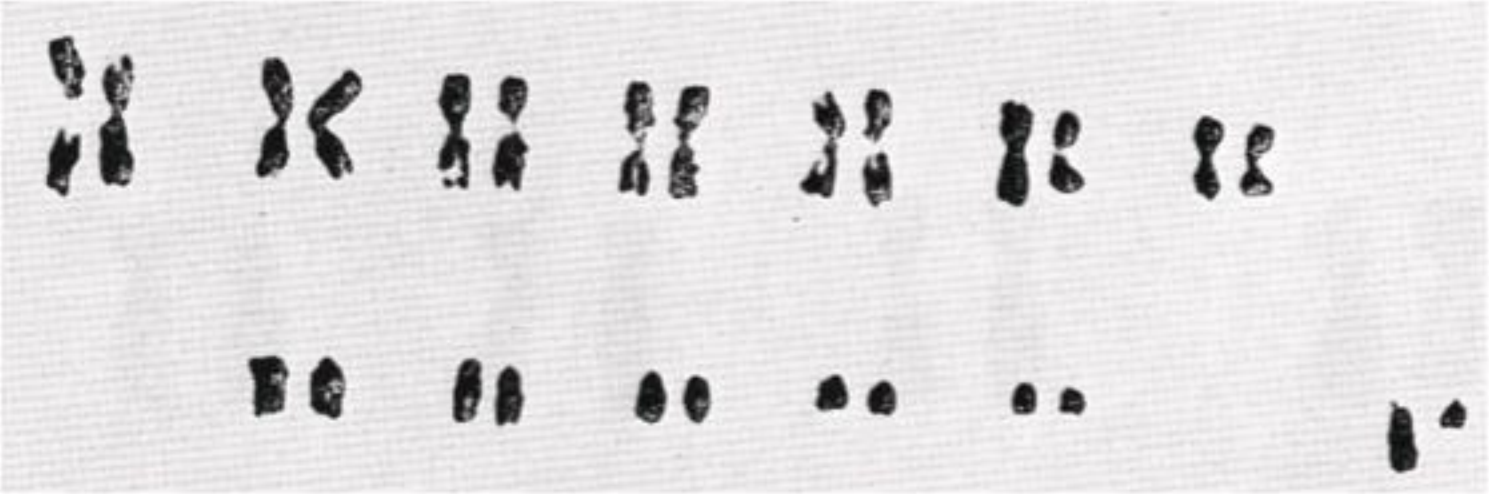
```
Criteria: 70% identity over 100 bp

*************** Conserved Regions  ***************

94585364         to    94585486   =     129bp      at    69.80% UTR
94594441         to    94594458   =     18bp       at    72.20% UTR
94594583         to    94594652   =     70bp       at    81.40% UTR
94587237         to    94587445   =     209bp      at    83.30% exon
94593805         to    94593910   =     106bp      at    91.50% exon
94594080         to    94594215   =     136bp      at    86.80% exon
94594331         to    94594440   =     110bp      at    90.90% exon
94589637         to    94589864   =     229bp      at    72.50% noncoding
94589940         to    94590050   =     112bp      at    69.60% noncoding
94590435         to    94590544   =     111bp      at    73.00% noncoding
94591250         to    94591381   =     133bp      at    73.70% noncoding
94593365         to    94593457   =     93bp       at    72.00% noncoding
```

Robertsonian fusion: creation of one metacentric chromosome by fusion of two acrocentrics

Ordinary male house mouse (*Mus musculus*, 2*n* = 40)



Male tobacco mouse (*Mus poschiavinus*, 2*n* = 26)



Outline

Introduction

General features of eukaryotic genomes and chromosomes

- *C* value paradox; organization; genome browsers
- Analysis of chromosomes using BioMart and biomaRt
- ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

- Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

- Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

- Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

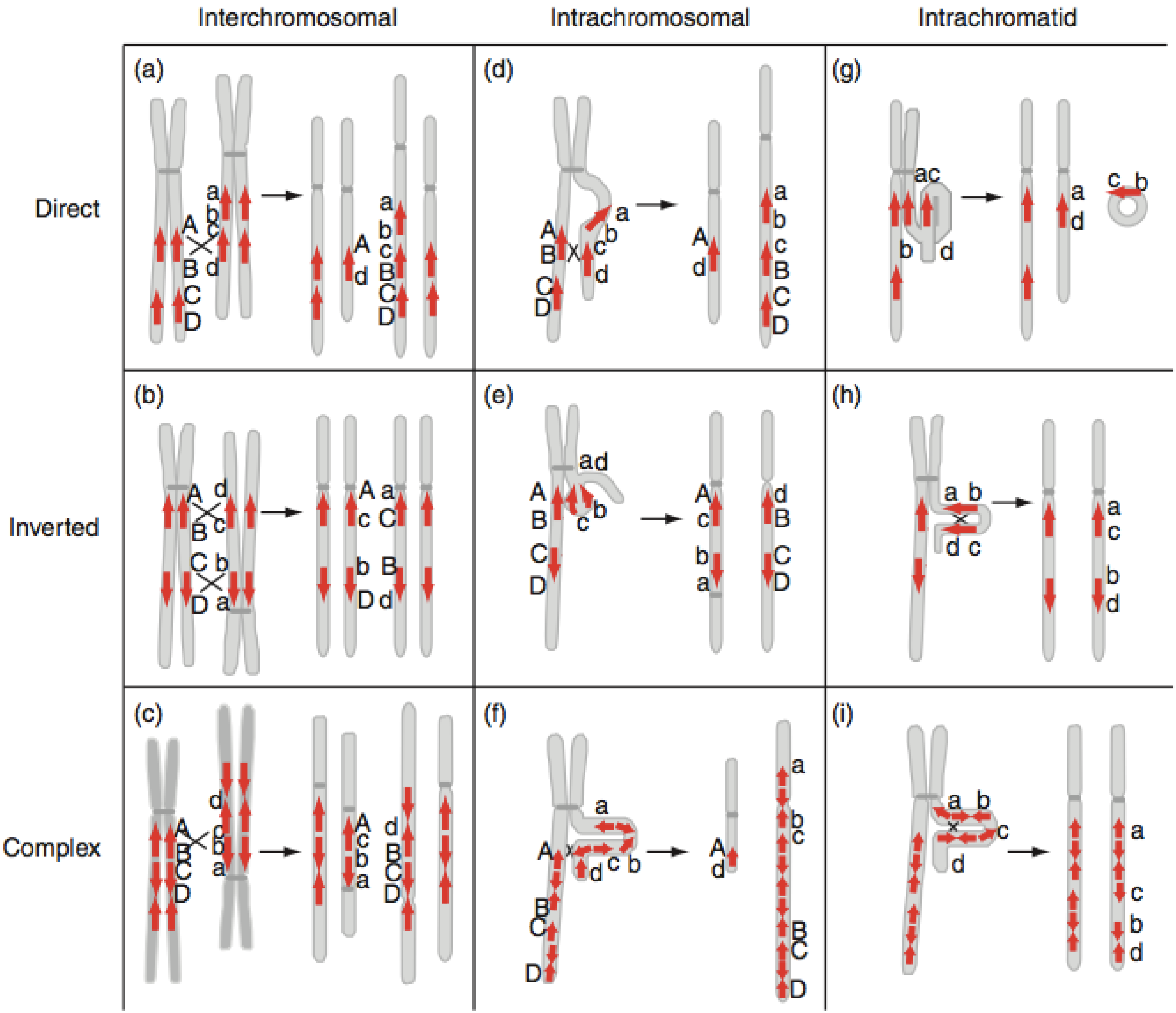
Variation in chromosomal DNA

- Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

Mechanisms of creating genomic rearrangements

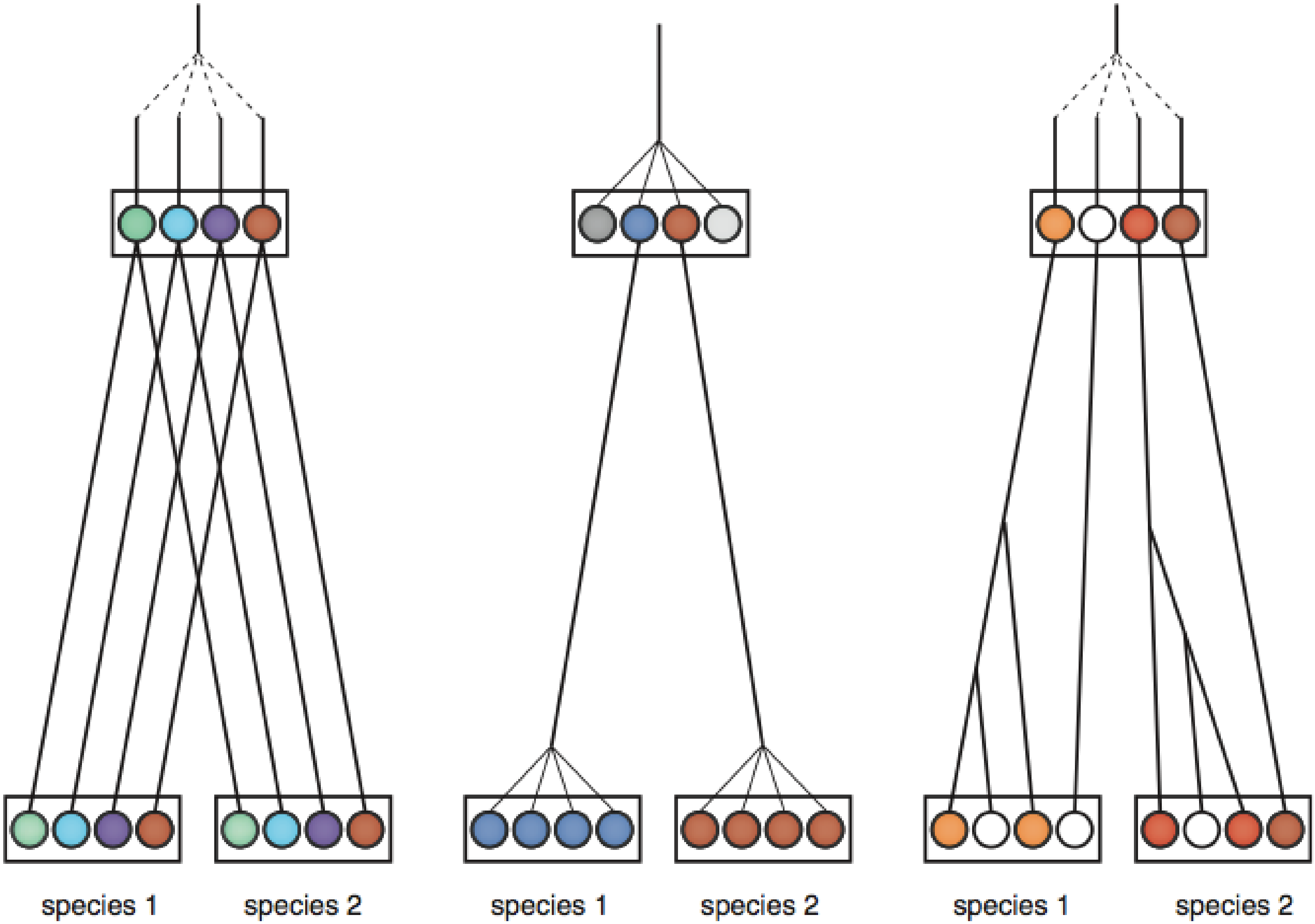


Models for creation of duplicate genes

divergent evolution

concerted evolution

birth-and-death evolution



Eukaryotic chromosomes can be dynamic

- Whole genome duplication (autopolyploidy) can occur, as in yeast and some plants.
- The genomes of two distinct species can merge, as in the mule (male donkey, 2n = 62 and female horse, 2n = 64)
- An individual can acquire an extra copy of a chromosome (e.g. Down syndrome, TS13, TS18)
- Chromosomes can fuse; e.g. human chromosome 2 derives from a fusion of two ancestral primate chromosomes
- Chromosomal regions can be inverted (hemophilia A)
- Portions of chromosomes can be deleted (e.g. del 11q syndrome)
- Segmental and other duplications occur
- Chromatin diminution can occur (*Ascaris*)

Inversions in chromosome evolution

Chromosomal inversions occur when a fragment of a chromosome breaks at two places, inverts, and is reinserted. This is a useful mechanism for producing a sterility barrier during speciation. An example is in **deer mice**; another example is in ***Anopheles gambiae***.

Ohno (1970) p. 42

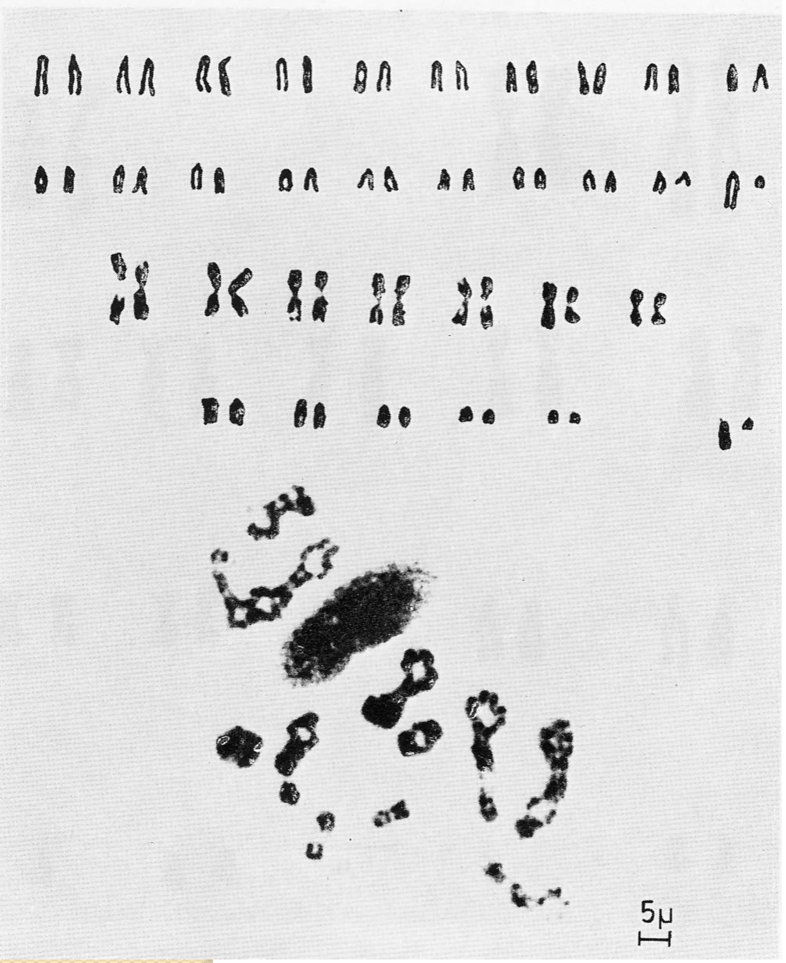
The eukaryotic chromosome: Robertsonian fusion creates one metacentric by fusion of two acrocentrics

Translocations occur when chromosomal material is exchanged between two non-homologous chromosomes. Roberstonian fusion, which often accompanies speciation, is the creation of one metacentric chromosome by the centric fusion of two acrocentrics.

Robertsonian fusions are often tolerated and may sometimes be considered selectively neutral. An example is the house mouse (*Mus musculus*, 2n = 40) and a small group of tobacco mice in Switzerland (*Mus poschiavinus*, 2n = 26). *Mus poschiavinus* is homozygous for seven Robertsonian fusions.

Page 675;
Ohno (1970) p. 43

The eukaryotic chromosome: Robertsonian fusion creates one metacentric by fusion of two acrocentrics



ordinary male house mouse (*Mus musculus*, 2n = 40)

male tobacco mouse (*Mus poschiavinus*, 2n = 26)

Male first meiotic metaphase from an interspecific F1-hbrid. Note seven trivalents (each from one *poschiavinus (Tobacco mouse)* metacentric and two *musculus* acrocentrics)

Ohno (1970) Plate II

Diploidization of the tetraploid

- A species can become tetraploid. All loci are duplicated, and what was formerly the diploid chromosome complement is now the haploid set of the genome.

Polyploid evolution occurs commonly in plants. For example, in the cereal plant *Sorghum*
S. versicolor (diploid) 2n = 2 x 5; 10 chromosomes
S. sudanense (tetraploid) 4n = 4 x 5; 20 chromosomes
S. halepense (octoploid) 8n = 8 x 5; 40 chromosomes

In plants, the male sex organ (stamen) and female organ (pistil or carpel) is present in the same flower; they are hermaphroditic.

Ohno (1970) pp 98- 101

Trisomy and polysomy

Nondisjunction results in two chromatids of one chromosome moving to the same division pole. In diploid species, one daughter cell receives three homologous chromosomes (trisomy). If this occurs in germ cells, the progeny may be trisomic.

In the Jimson weed (*Datura stramonium*) trisomy for each of the 12 chromosomes was observed by Blakeslee (1930). A mating between trisomic individuals may produce tetrasomic progeny having two homologous chromosomes (thus duplicating an entire chromosome).

Ohno (1970) p. 107

Trisomy and polysomy

For vertebrates, this mechanism is too severe. Generally, only trisomy of chromosomes 13, 18, or 21 are compatible with postnatal survival in humans.

In rainbow trout that have become tetraploid, trisomy (i.e. from four to five copies) and monosomy (i.e. from four to three copies) may be tolerated.

Ohno (1970) p. 107

Outline

Introduction

General features of eukaryotic genomes and chromosomes

C value paradox; organization; genome browsers

Analysis of chromosomes using BioMart and biomaRt

ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

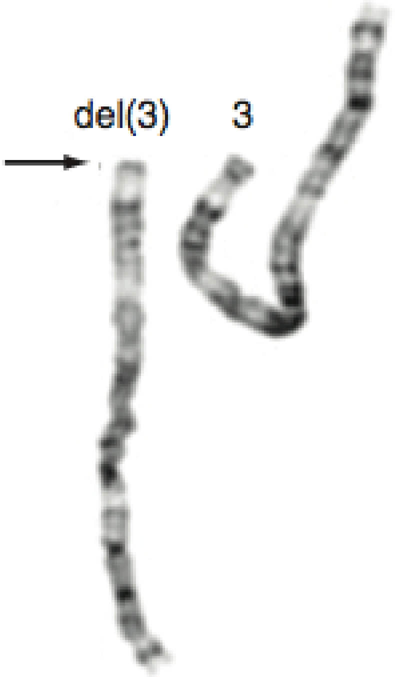
Variation in chromosomal DNA

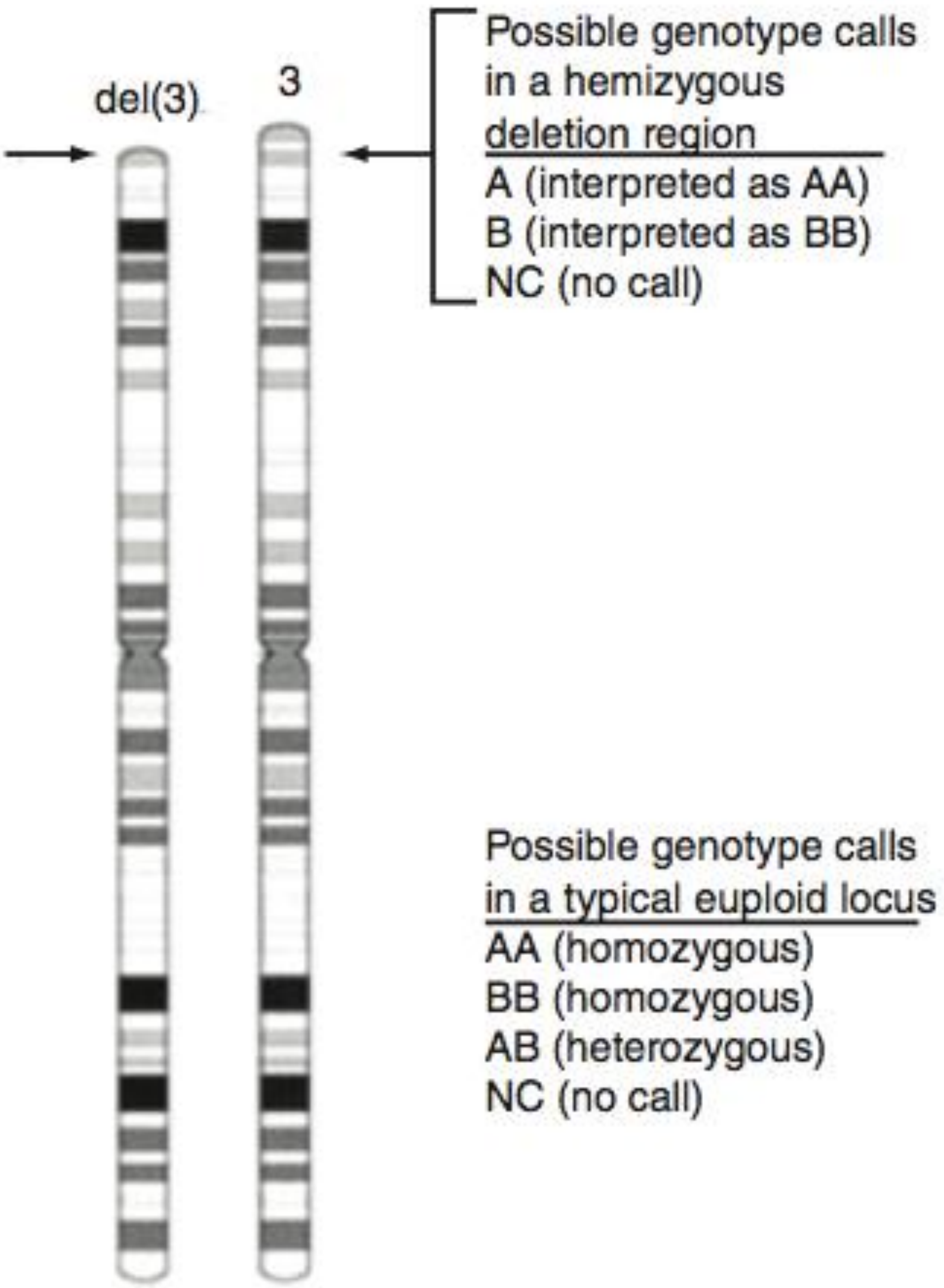
Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

SNP microarrays provide information about chromosomal copy number (e.g. a deletion on 3p) and genotype (e.g. occurrence of homozygous calls).





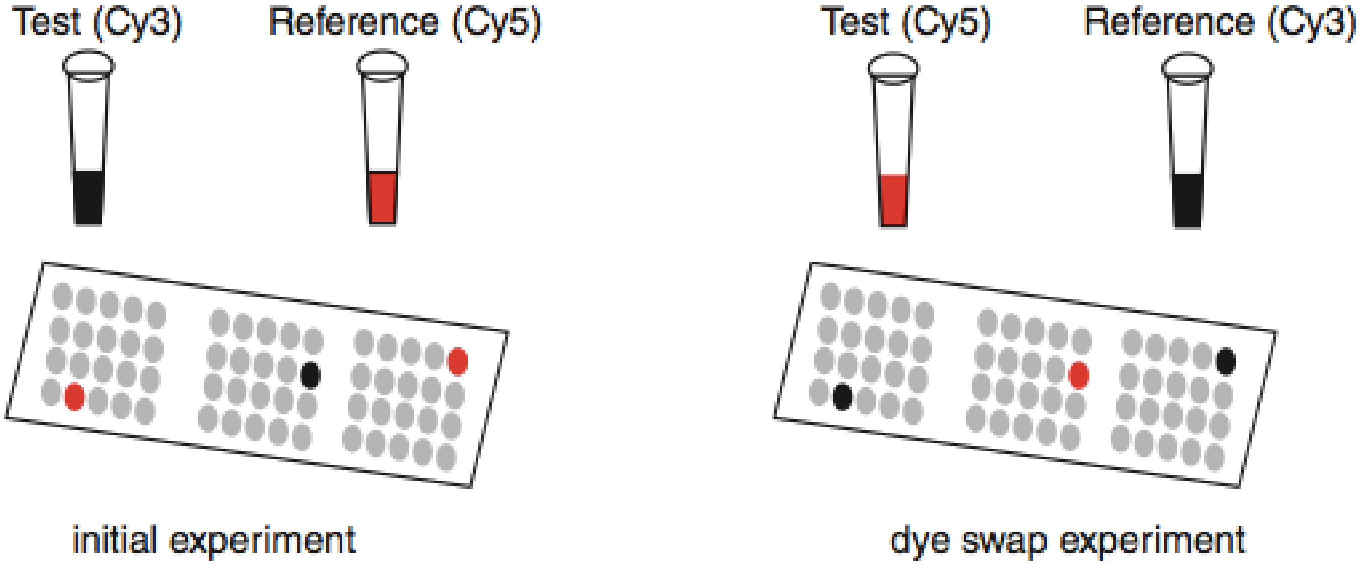
Possible genotype calls in a hemizygous deletion region

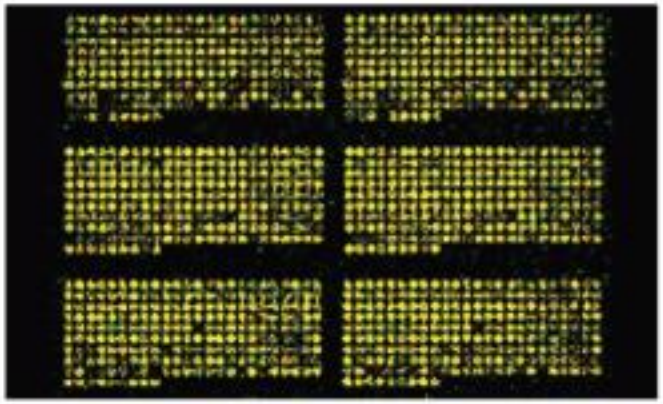
A (interpreted as AA)
B (interpreted as BB)
NC (no call)

Possible genotype calls in a typical euploid locus

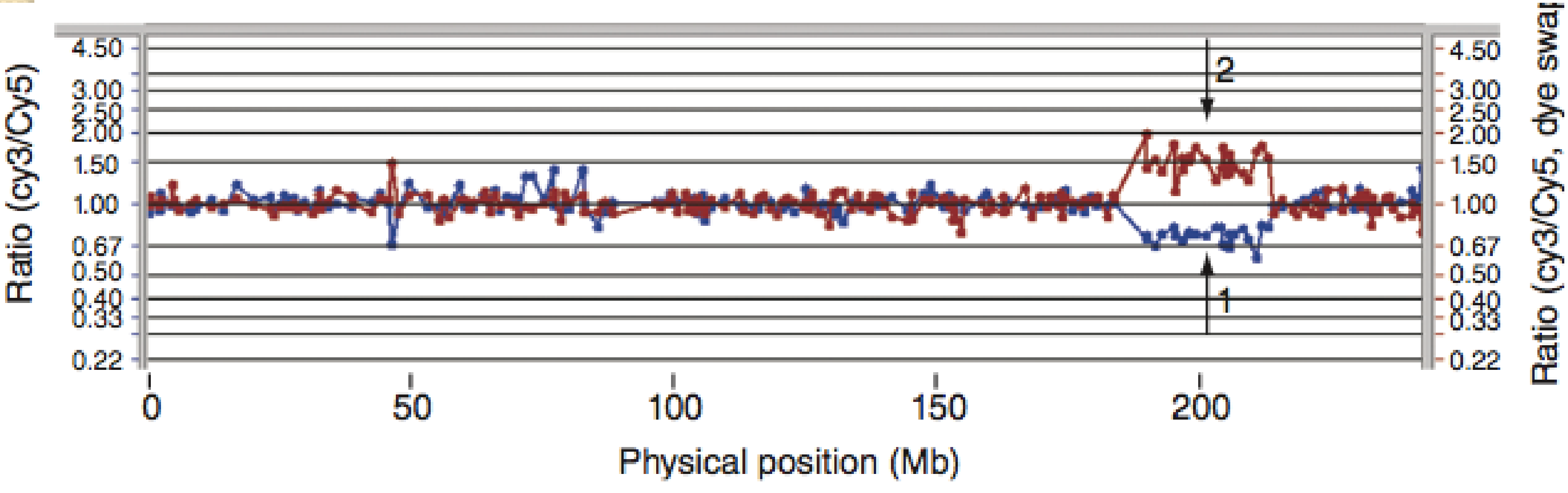
AA (homozygous)
BB (homozygous)
AB (heterozygous)
NC (no call)

Array comparative genome hybridization

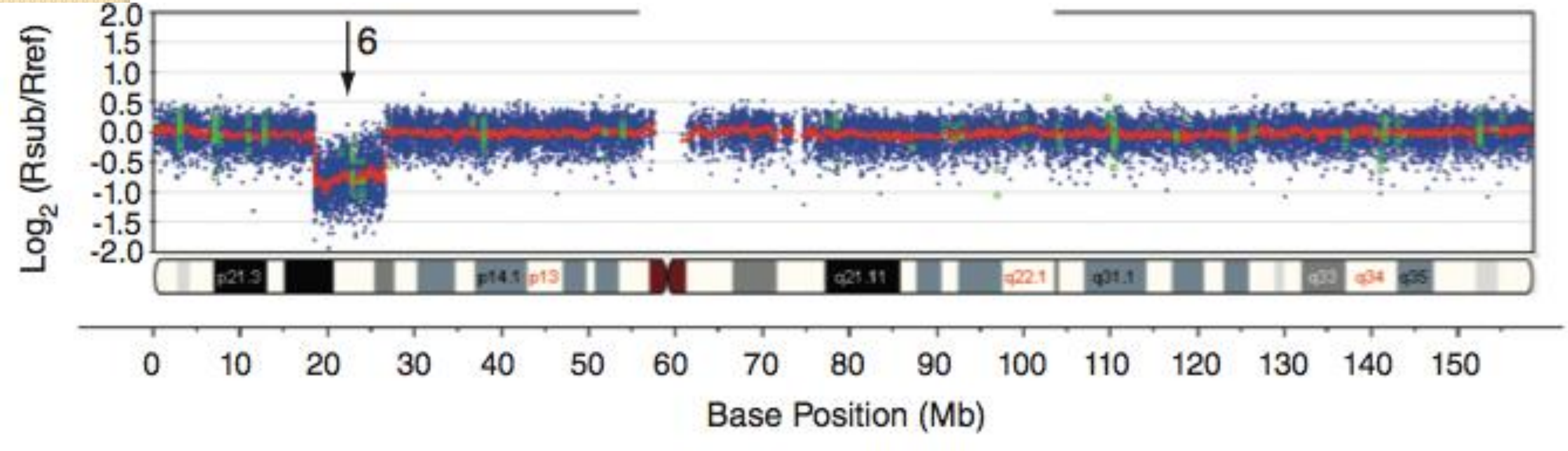




Array comparative genome hybridization

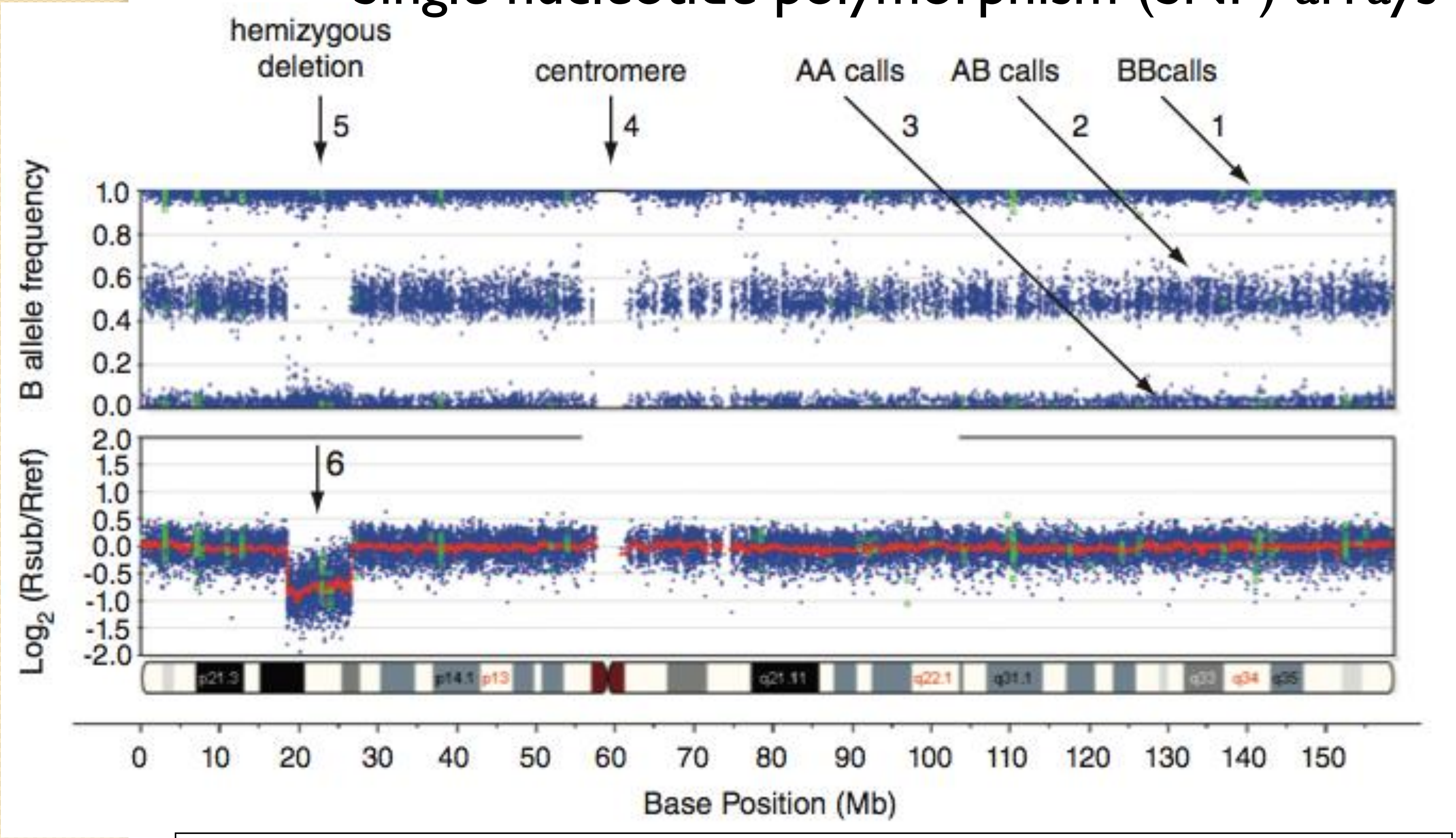


Single nucleotide polymorphism (SNP) arrays



A SNP array provides information on copy number, based on signal intensity. The x-axis is genomic position (in megabases) along a chromosome. The y-axis shows the log2 ratio of signal from a subject relative to a reference, and ordinarily has a value of zero. Here in a deletion region (arrow 6) there is 1 chromosomal copy instead of 2. Each dot is a data point from a SNP array. There is an absence of signal in the region of the centromere (~60 Mb).

Single nucleotide polymorphism (SNP) arrays



SNP arrays also provide genotype information (upper panel): calls are AA (0% B allele), heterozygous (AB; 50% B allele), or BB (100% B allele). Note the loss of heterozygous calls in a hemizygous deletion region (arrow 5).

Outline

Introduction

General features of eukaryotic genomes and chromosomes

C value paradox; organization; genome browsers

Analysis of chromosomes using BioMart and biomaRt

ENCODE Project; critiques of ENCODE

Repetitive DNA content of eukaryotic genomes

Noncoding and repetitive DNA sequences

Gene content of eukaryotic chromosomes

Definition of gene; finding genes; EGASP; RefSeq, UCSC genes, and GENCODE

Regulatory regions of eukaryotic chromosomes

Databases of regulatory factors; ultraconserved elements; nonconserved elements

Comparison of eukaryotic DNA

Variation in chromosomal DNA

Dynamic nature of chromosomes; variation in individual genomes; six types of structural variation

Techniques to measure chromosomal change

Perspective

Perspective

- One of the broadest goals of biology is to understand the nature of each species of life: what are the mechanisms of development, metabolism, homeostasis, reproduction, and behavior? Sequencing of a genome does not answer these questions directly. Instead, we must first try to annotate the genome sequence in order to estimate its contents, and then we try to interpret the function of these parts in a variety of physiological and evolutionary processes.

- As complete genomes are sequenced, we are becoming aware of the nature of non-coding and coding DNA, and repetitive DNA. Genome browsers and various bioinformatics tools are useful to explore and tabulate chromosomal features, we also appreciate the dynamic, complex nature of chromosomes as exquisite biological objects.