Chapter 20:
The human genome

Jonathan Pevsner, Ph.D.
pevsner@kennedykrieger.org

Bioinformatics and Functional Genomics
(Wiley-Liss, 3rd edition, 2015)

Learning objectives

After studying this chapter you should be able to:

■ describe the main features of the human genome;
■ provide an overview of all the human chromosomes, giving a general description of their size, number of genes, and key features; and
■ explain the purpose and main conclusions of several key human genome efforts including the HapMap Project and 1000 Genomes Projects.

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Introduction

- The human genome is the complete set of DNA in *Homo sapiens*.
- Its initial sequencing (2003) came 50 years after the publication of the double-stranded helical structure of DNA by Crick and Watson (1953).
- In 2001 the sequencing extensive draft versions of the human genome were reported separately by the International Human Genome Sequencing Consortium (IHGSC) and by Celera Genomics.
- In this chapter we follow the outline of the 62 page article in *Nature* by IHGSC (2001), providing updated information. We then survey the human chromosomes and end by discussing variation in the human genome.

B&FG 3e
Page 957

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

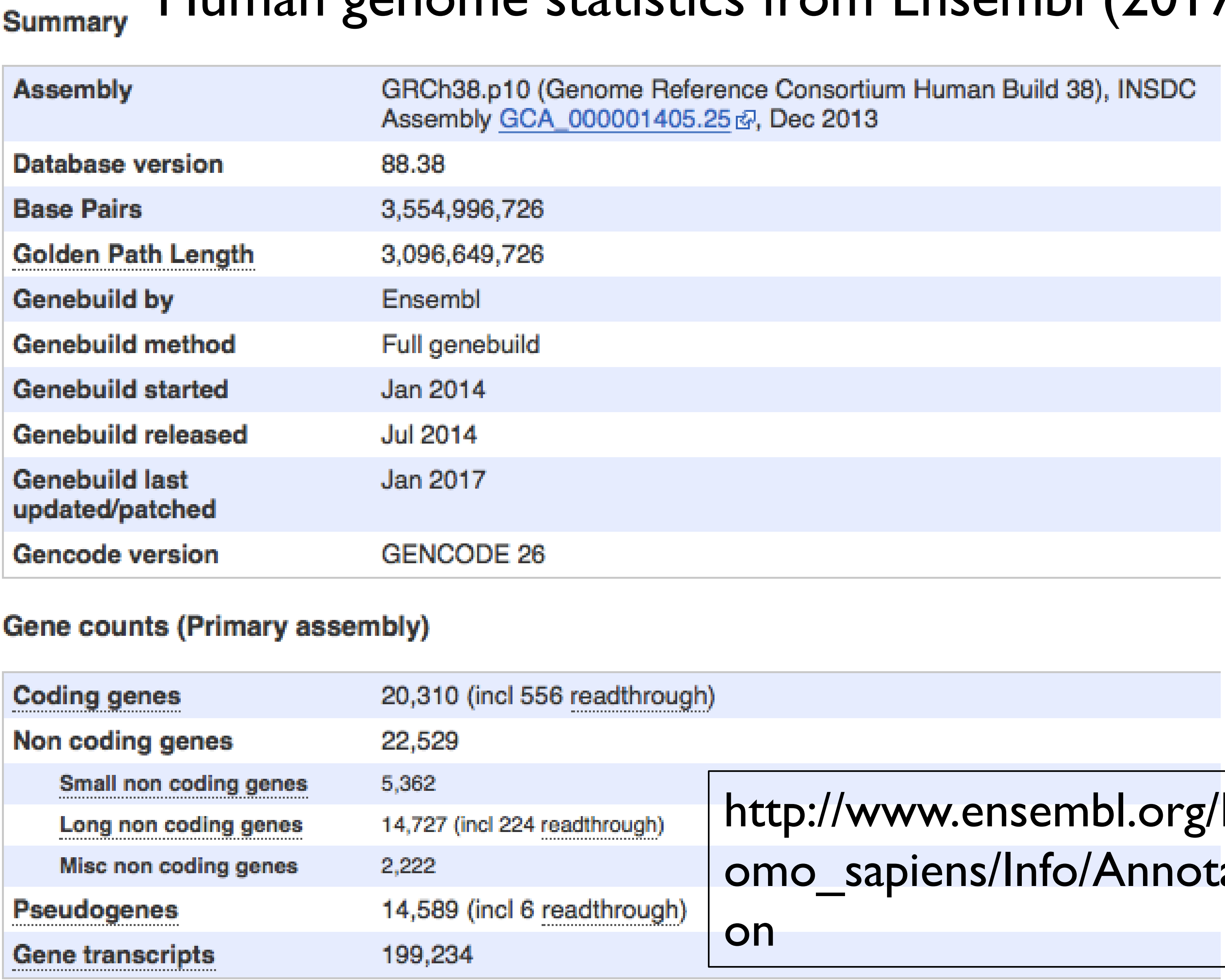
Main conclusions of human genome project

1. There are about 20,310 human protein-coding genes. This number is far smaller than earlier estimates.

   In 2001 the public consortium estimated 31,000, while Celera estimated 38,500.

B&FG 3e
Page 958

Human genome statistics from Ensembl (2017)



Summary

Assembly	GRCh38.p10 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.25, Dec 2013
Database version	88.38
Base Pairs	3,554,996,726
Golden Path Length	3,096,649,726
Genebuild by	Ensembl
Genebuild method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jan 2017
Gencode version	GENCODE 26

Gene counts (Primary assembly)

Coding genes	20,310 (incl 556 readthrough)
Non coding genes	22,529
Small non coding genes	5,362
Long non coding genes	14,727 (incl 224 readthrough)
Misc non coding genes	2,222
Pseudogenes	14,589 (incl 6 readthrough)
Gene transcripts	199,234

http://www.ensembl.org/Homo_sapiens/Info/Annotation

Main conclusions of human genome project

1. We now appreciate that human have about the same number of protein-coding genes as fish and plants, and not that many more genes than worms and flies.

Fugu rubripes (pufferfish):	31,000 to 38,000
	18,500*
Arabidopsis thaliana (thale cress):	27,600**
Caenorhabditis elegans (worm):	20,300*
Drosophila melanogaster (fly):	13,900*

* 2017 estimate from Ensembl

** 2017 estimate from TAIR

Main conclusions of human genome project

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

Vertebrates have a more complex mixture of protein domain architectures. Additionally, the human genome displays greater complexity in its processing of mRNA transcripts by alternative splicing.

B&FG 3e
Page 958

Main conclusions of human genome project

3. Hundreds of human genes were acquired from bacteria by lateral gene transfer, according to the initial report.

Evidence: compare the proteomes of human, fly, worm, yeast, *Arabidopsis*, eukaryotic parasites, and all completed prokaryotic genomes. Find some genes shared exclusively by humans and bacteria—but according to TIGR, only about 40 of these genes (or fewer?) were acquired by LGT.

Reasons for artifactually high estimates include:
-- gene loss
-- small sample size of species

B&FG 3e
Page 958

Main conclusions of human genome project

4. 98% of the genome does not code for genes

   >50% of the genome consists of repetitive DNA derived from transposable elements:

   LINEs (20%)
   SINEs (13%)
   LTR retrotransposons (8%)
   DNA transposons (3%)

   There has been a decline in activity of some of these elements in the human lineage.

B&FG 3e
Page 958

Main conclusions of human genome project

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

-- tandem duplications (rare)
-- retrotransposition (intronless paralogs)
-- segmental duplications (common)

B&FG 3e
Page 958

Main conclusions of human genome project

6. There are >1 million *Alu* repeats in the human genome.

These are about 300 base pairs and contain an *Alu*I restriction enzyme site. They occupy ~10% of the genome. We saw an example of an *Alu* repeat in Chapter 16.

Their distribution is non-random: they are retained in GC-rich regions and may confer some benefit.

B&FG 3e
Page 959

Main conclusions of human genome project

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

B&FG 3e
Page 959

Main conclusions of human genome project

8. More than 1.4 million single nucleotide polymorphisms (SNPs; single base pair changes) were identified.

Celera initially identified 2.1 million SNPs.

Currently (2017), dbSNP at NCBI (build 150*) has about 324 million SNPs. An individual genome has about 3-4 million SNPs. Fewer than 1% of SNPs alter protein sequence.

* See https://www.ncbi.nlm.nih.gov/news/04-11-2017-human-snp-build-150/

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Three gateways to access the human genome

NCBI map viewer
www.ncbi.nlm.nih.gov

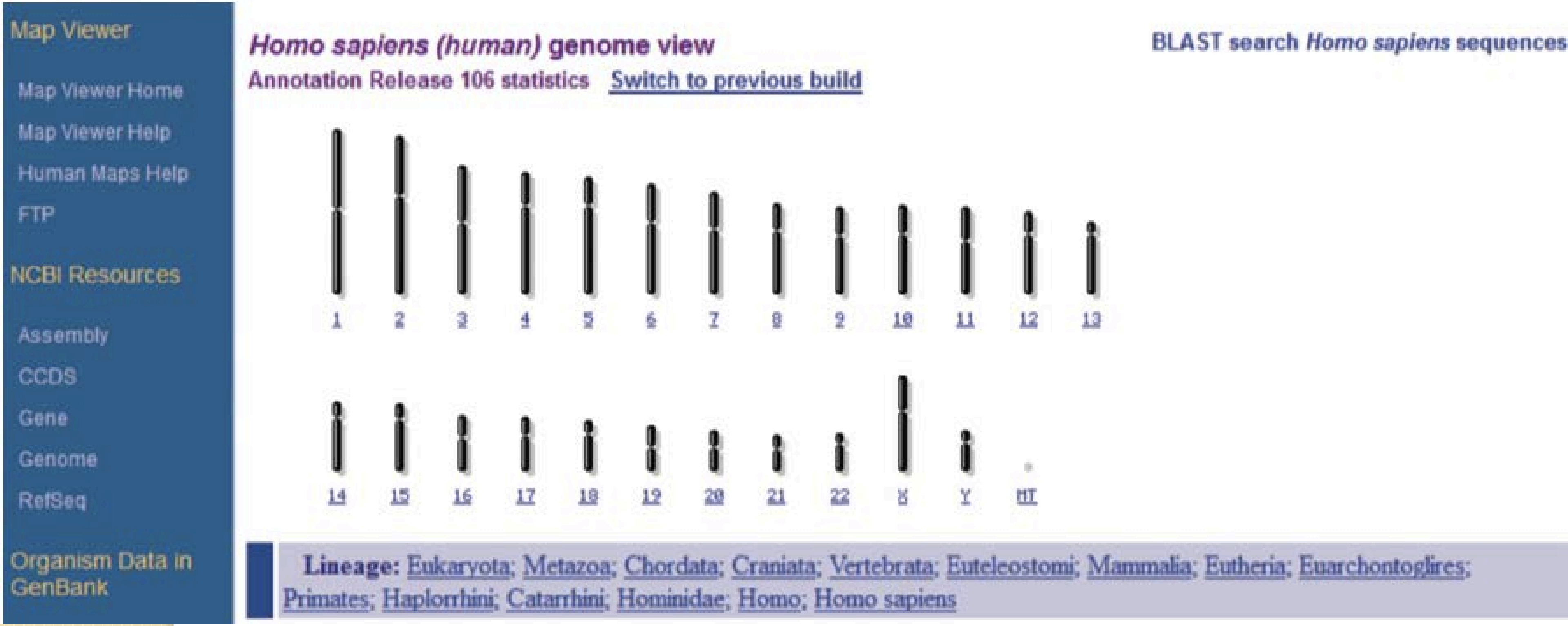
Ensembl Project (EBI/Sanger Institute)
www.ensembl.org

UCSC (Golden Path)
www.genome.ucsc.edu

Each of these three sites provides essential resources to study the human genome (and other genomes)

B&FG 3e
Page 959

Human Map Viewer from NCBI



Homo sapiens (human) genome view

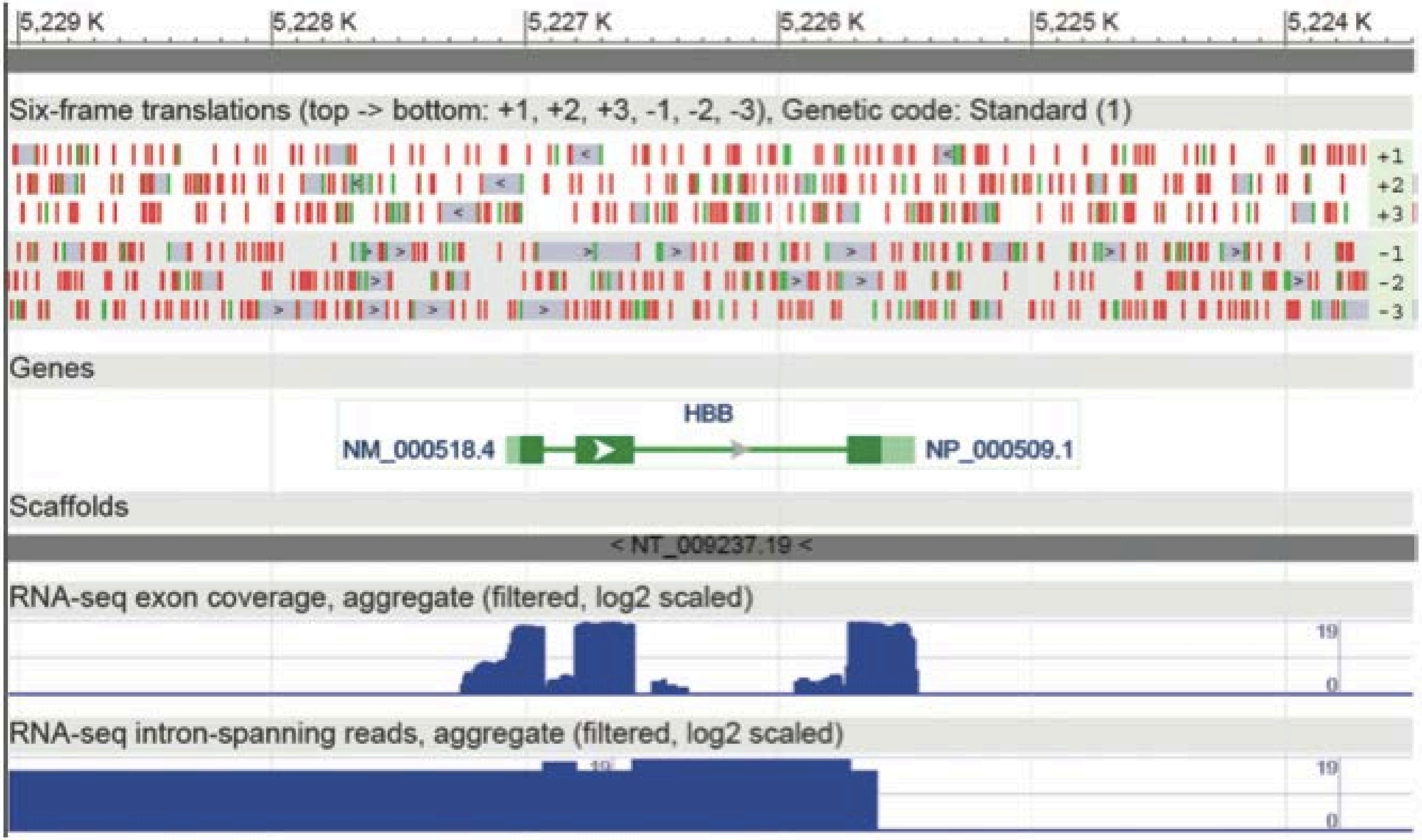
Annotation Release 106 statistics Switch to previous build

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens

B&FG 3e
Fig. 20.1
Page 960

The Human Map Viewer is accessible from NCBI. This resource displays cytogenetic, genetic, physical, and radiation hybrid maps of human genome sequence.

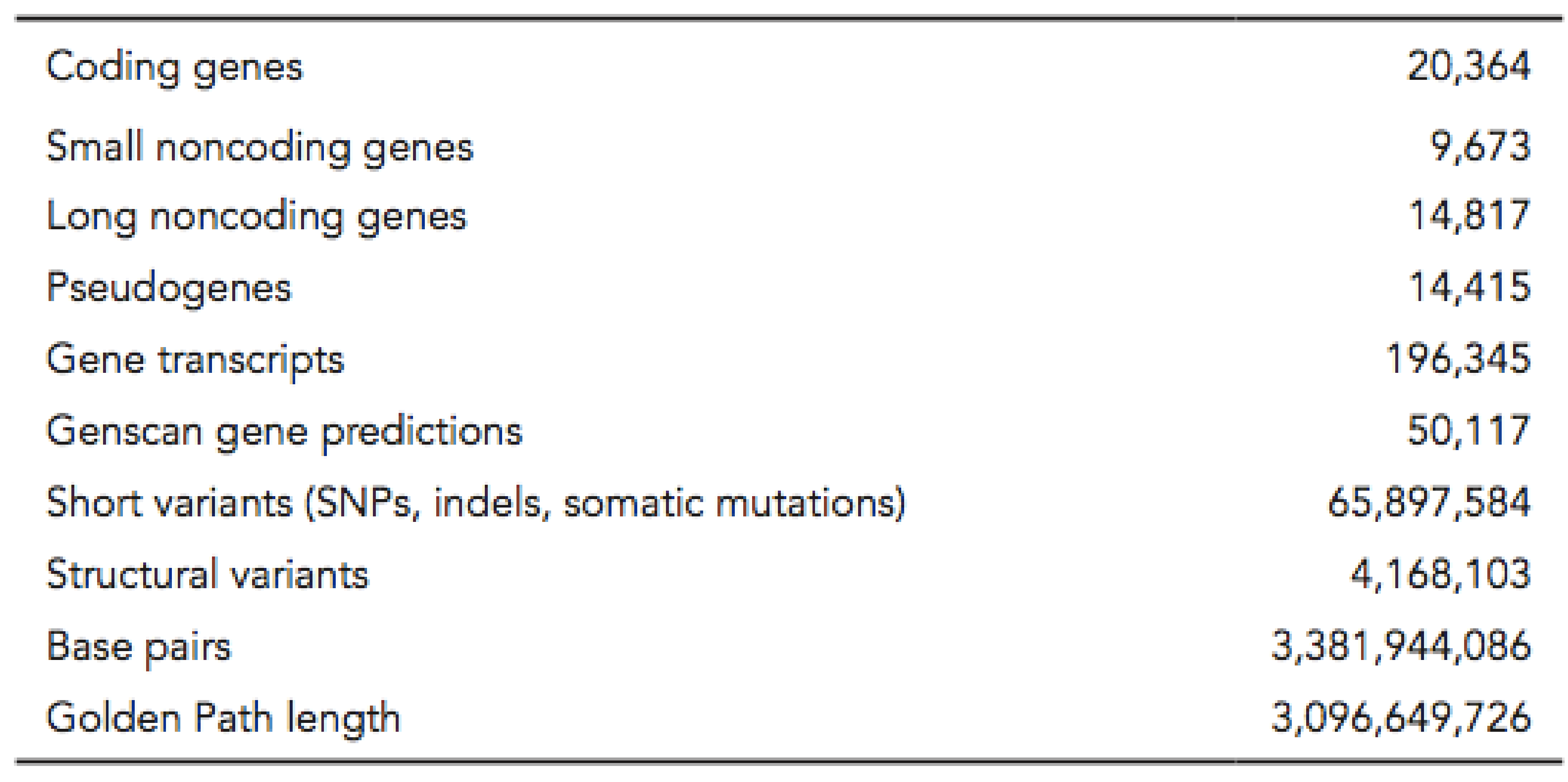
NCBI Gene sequence viewer


5,229 K
5,228 K
5,227 K
5,226 K
5,225 K
5,224 K
Six-frame translations (top -> bottom: +1, +2, +3, -1, -2, -3), Genetic code: Standard (1)
+1
+2
+3
-1
-2
-3
Genes
HBB
NM_000518.4
NP_000509.1
Scaffolds
< NT_009237.19 <
RNA-seq exon coverage, aggregate (filtered, log2 scaled)
19
0
RNA-seq intron-spanning reads, aggregate (filtered, log2 scaled)
19
19
0

The sequence viewer shows a region of chromosome 11 containing the beta globin gene. A variety of tracks can be added; shown here are six-frame translations, scaffolds, and RNA-seq data.

B&FG 3e
Fig. 20.2
Page 960

Human genome statistics from Ensembl



Coding genes	20,364
Small noncoding genes	9,673
Long noncoding genes	14,817
Pseudogenes	14,415
Gene transcripts	196,345
Genscan gene predictions	50,117
Short variants (SNPs, indels, somatic mutations)	65,897,584
Structural variants	4,168,103
Base pairs	3,381,944,086
Golden Path length	3,096,649,726

Ensembl Release 75

Ensembl is a comprehensive, authoritative resource for information about the human genome as well as other genomes.

B&FG 3e
Table 20.1
Page 961

The UCSC human genome browser

The University of California at Santa Cruz (UCSC) offers a genome browser with the "golden path" annotation of the human genome.

The browser features searches by keyword, gene name, or other text searches. UCSC offers the lightning fast BLAT BLAST-like tool.

A key feature of this browser is its customizable annotation tracks. About half of these tracks are offered by users of the site throughout the world.

Visit http://genome.ucsc.edu

B&FG 3e
Page 961

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Background of the human genome project

The Human Genome Project (HGP) was first proposed by the U.S. National Research Council in 1988. The goals were to create genetic, physical, and sequence maps of the human genome. In parallel, genomes of model organisms were to be studied.

(You can read this report on-line via http://www.nap.edu)

B&FG 3e
Page 964

Eight goals of Human Genome Project (1998–2003)

[1] Human DNA sequence

[2] Develop sequencing technology

[3] Identify human genome sequence variation

[4] Functional genomics technology

[5] Comparative genomics

[6] ELSI: ethical, legal, and social issues

[7] Bioinformatics and computational biology

[8] Training and manpower

ELSI: Who owns genetic information? Who has access to it?
To what extent do genes determine behavior?
What is the relation between genes and race?

B&FG 3e
Table 20.2
Page 965

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Strategic issues:
Hierarchical / shotgun sequencing

The human genome was sequenced in parallel by
a public consortium (IHGSC) and by Celera Genomics.
These groups applied alternative sequencing strategies.

B&FG 3e
Page 966

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Human genome project: strategies

Whole genome shotgun sequencing (Celera)

-- Given the computational capacity, this approach is far faster than hierarchical shotgun sequencing

-- The approach was validated using *Drosophila*

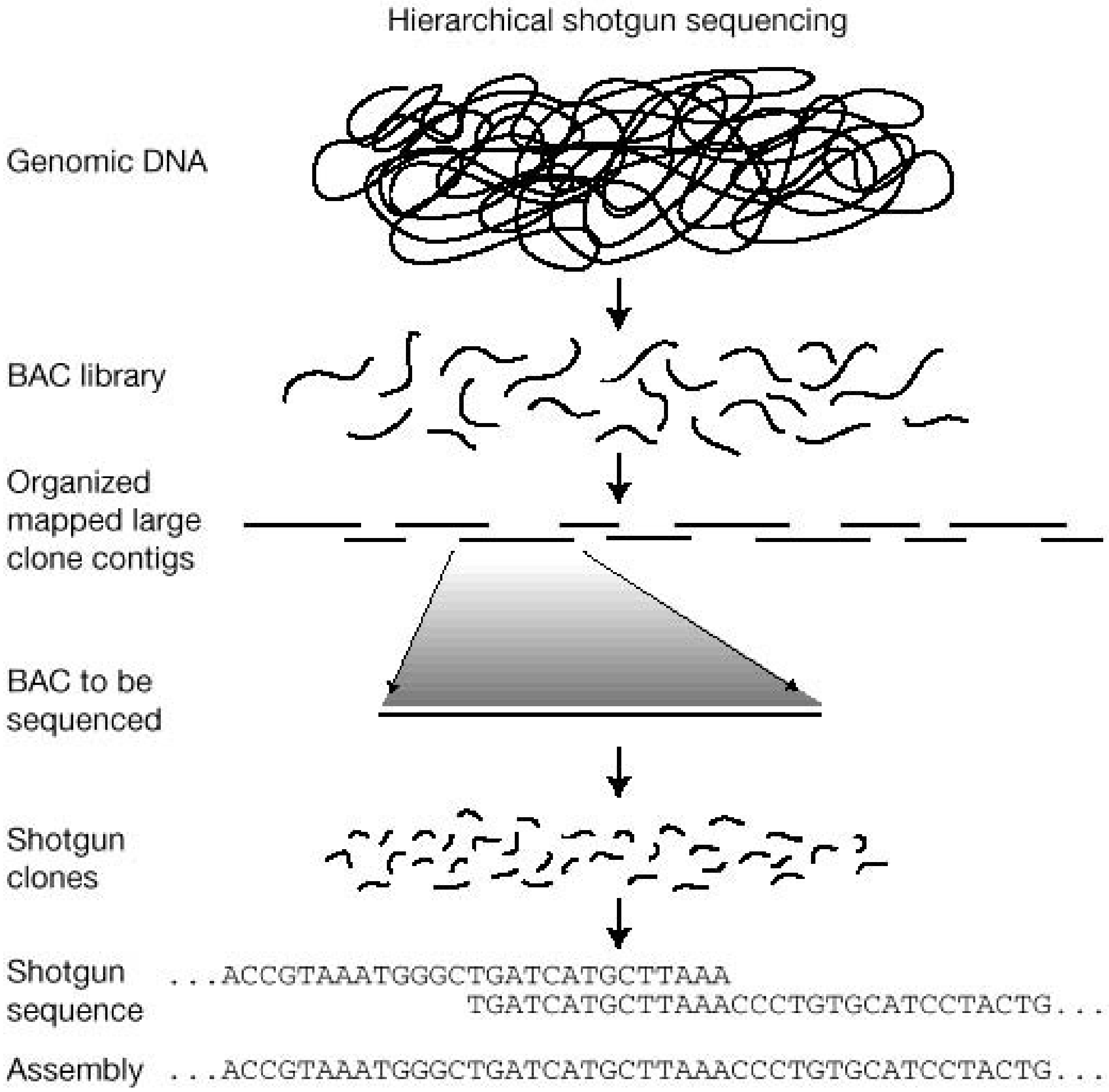
Hierarchical shotgun sequencing (public consortium)

-- 29,000 BAC clones

-- 4.3 billion base pairs

-- It is helpful to assign chromosomal loci to sequenced fragments, especially in light of the large amount of repetitive DNA in the genome

-- Individual chromosomes assigned to centers


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

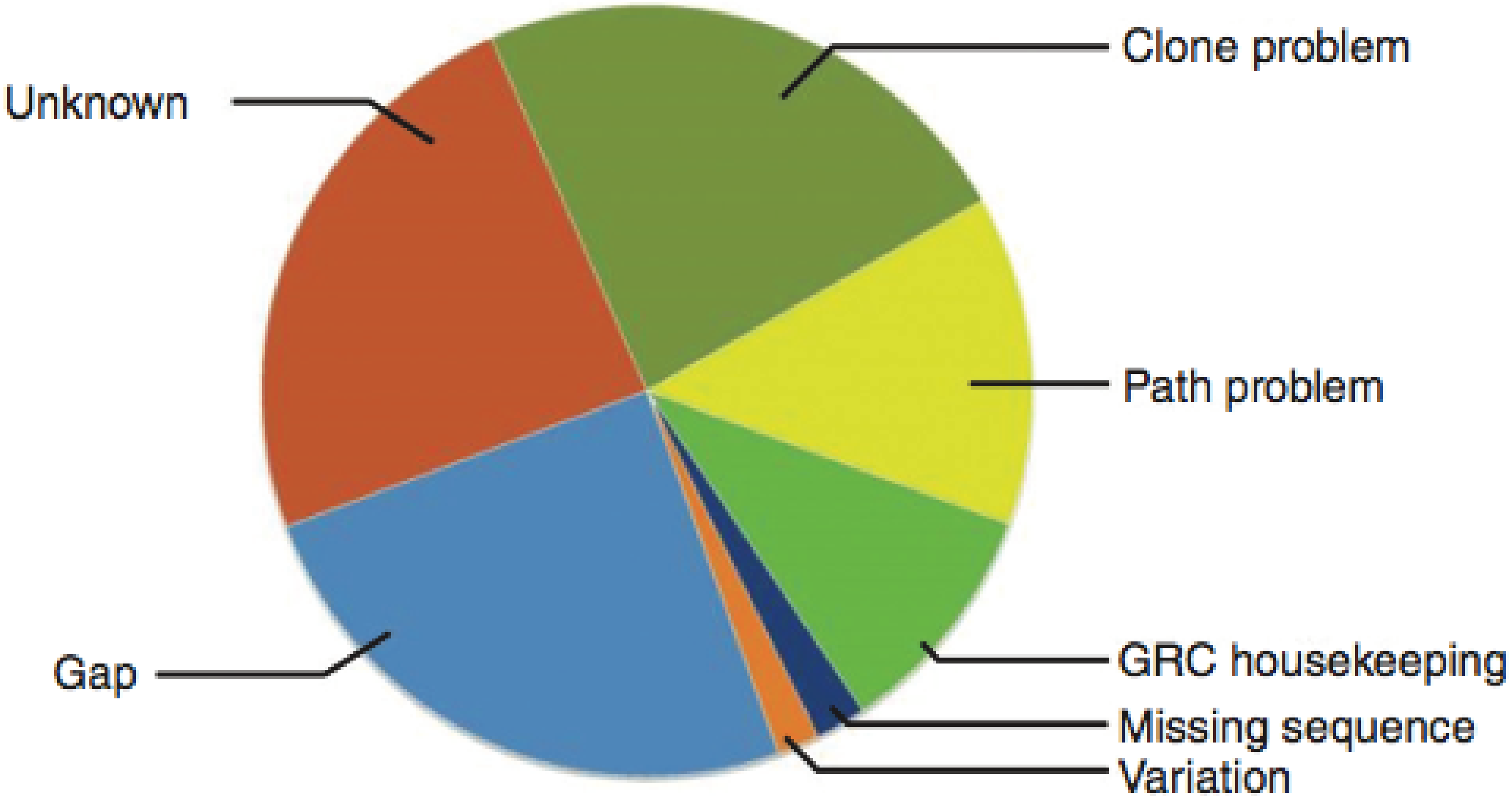
Source: IHGSC (2001)

Genome Reference Consortium (GRC)

- The Genome Reference Consortium (GRC) is responsible for coordinating new assemblies for human, mouse, and zebrafish genomes.
- Every few years a new genome assembly is released.
- The current human assembly is Genome Reference Consortium Human Build 38 (abbreviated GRCh38).
- Each new genome build must be annotated (a very complex process) and this is a main reason that GRCh37 (sometimes called hg19) has remained more popular than GRCh38.
- Investigators must choose which build to use for a variety of applications.

B&FG 3e
Page 967

Issues addressed by the Genome Reference Consortium (GRC) in the releases leading up to GRCh38 (Dec. 2013)



Categories of issues. Clone problem: a single clone has a single-nucleotide difference or misassembly. Path problem: the tiling path is incorrect and must be updated. GRC housekeeping: the tiling path must be regularized.

B&FG 3e
Fig. 20.5
Page 967

Features of the genome sequence

The genome sequence (in 2001) included a mixture of finished, draft, and pre-draft data. The N50 length describes the largest length *L* such that 50% of all nucleotides are contained in contigs or scaffolds of at least size *L*. For the 2001 version of the human genome, N50 is at least 8.4 Mb.

B&FG 3e
Page 968

Features of the genome sequence

The quality of genome sequence is assessed by counting the number of gaps and by measuring the nucleotide accuracy. About 91% of the unfinished draft sequence had an error rate less than 1 per 10,000 bases.
This corresponds to a PHRAP score >40 (i.e. an error probability of $10^{-40/10}$, or 99.99% accuracy).

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Broad genomic landscape

The broad genomic landscape includes the following features:

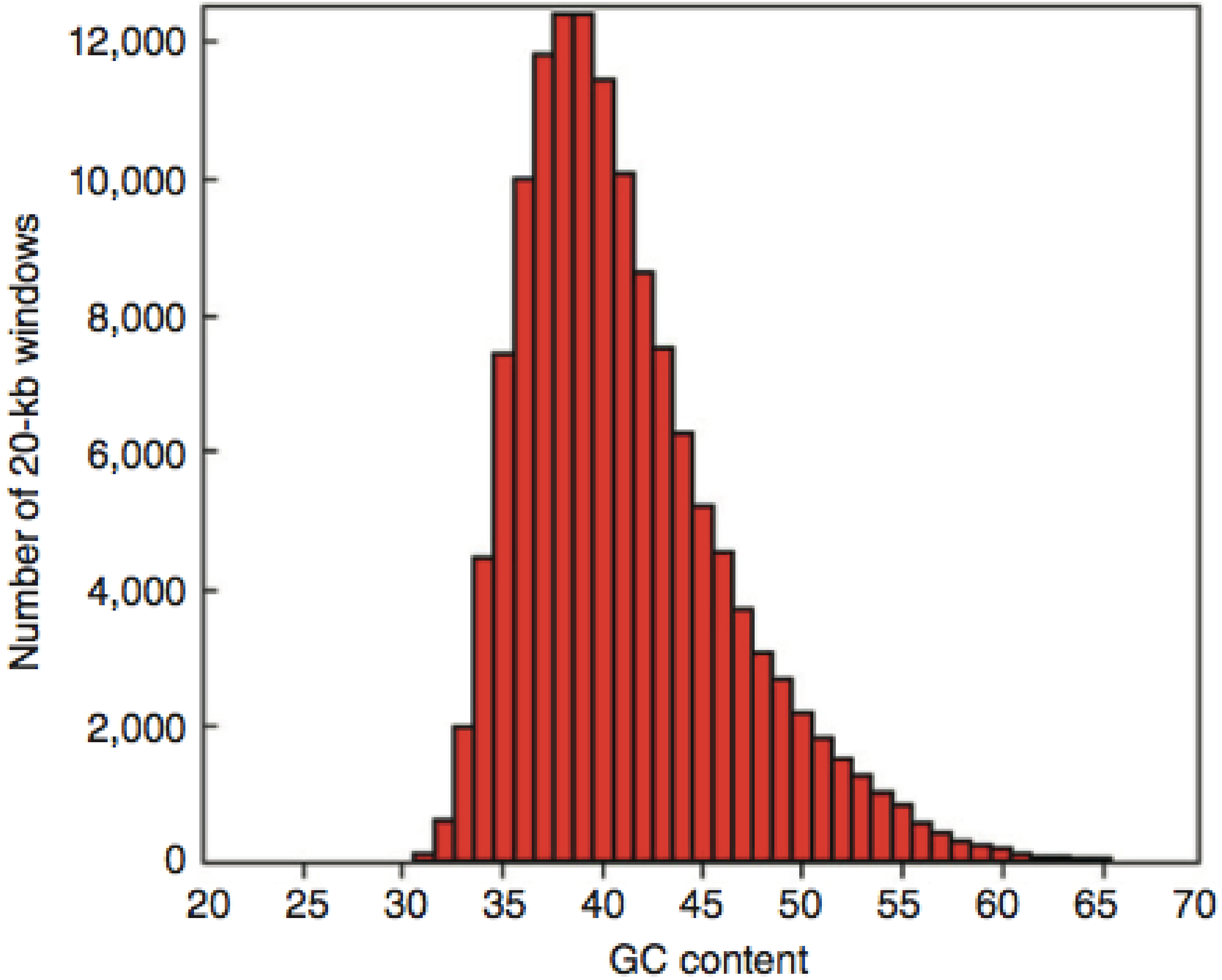
- Long-range variation in GC content
- CpG islands
- Comparison of genetic and physical distance
- The repeat content of the human genome
- The protein-coding gene content of the human genome
- Non-coding genes

B&FG 3e
Page 968

Broad genomic landscape: GC content

The overall GC content of the human genome is 41%. A plot of GC content versus number of 20 kb windows shows a broad profile with skewing to the right.

Histogram of percent GC content versus the number of 20 kb windows in the draft human genome sequence



B&FG 3e
Fig. 20.6
Page 969

The distribution is skewed to the right, with a mean GC content of 41%

Broad genomic landscape: GC content

Some genomic regions are GC-rich, while some are GC-poor. Giorgio Bernardi and colleagues described "isochores" which are large DNA segments (e.g. >300 kb) that are fairly homogeneous compositionally and can be divided into GC-rich and GC-poor subtypes.

B&FG 3e
Page 969

Broad genomic landscape: CpG islands

Dinucleotides of CpG are under-represented in genomic DNA, occuring at one fifth the expected frequency. CpG dinucleotides are often methylated on cytosine (and subsequently may be deamination to thymine).

Methylated CpG residues are often associated with house-keeping genes in the promoter and exonic regions.

Methyl-CpG binding proteins recruit histone deacetylases and are thus responsible for transcriptional repression. They have roles in gene silencing, genomic imprinting, and X-chromosome inactivation.

B&FG 3e
Page 969

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Repeat content of the human genome

The human genome contains >50% repetitive DNA. In Chapter 8 we discussed five classes of repetitive DNA in humans:

[1] interspersed repeats (transposon-derived)
[2] processed pseudogenes
[3] simple sequence repeats (micro-, minisatellites)
[4] segmental duplications
[5] blocks of tandem repeats (e.g. at centromeres)

B&FG 3e
Page 971

Human genome: interspersed repeats

Four main classes of interspersed repeats:

[1] LINEs (21% of human genome)

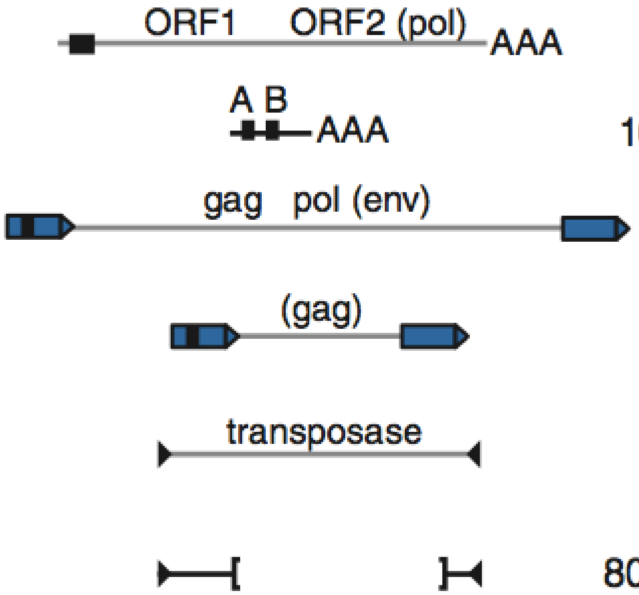
[2] SINEs (13%)

[3] Long terminal repeat transposons (8%)

[4] DNA transposons (3%)

B&FG 3e
Page 972

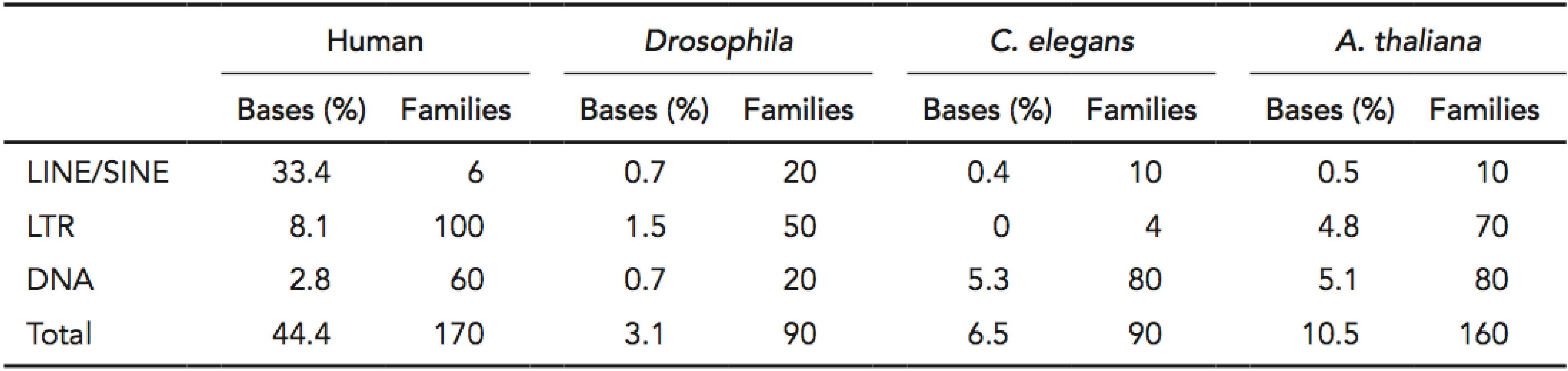
Four types of transposable elements in the human genome: LINEs, SINEs, LTR transposons, and DNA transposons



		Length	Copy number	Fraction of genome
LINEs	Autonomous	6-8 kb	850,000	21%
SINEs	Non-autonomous	100-300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	6-11 kb	450,000	8%
	Non-autonomous	1.5-3 kb		
DNA transposon fossils	Autonomous	2-3 kb	300,000	3%
	Non-autonomous	80-3,000 bp		

B&FG 3e
Fig. 20.9
Page 972

Interspersed repeats occupy a far greater proportion of the human genome than in other eukaryotic genomes

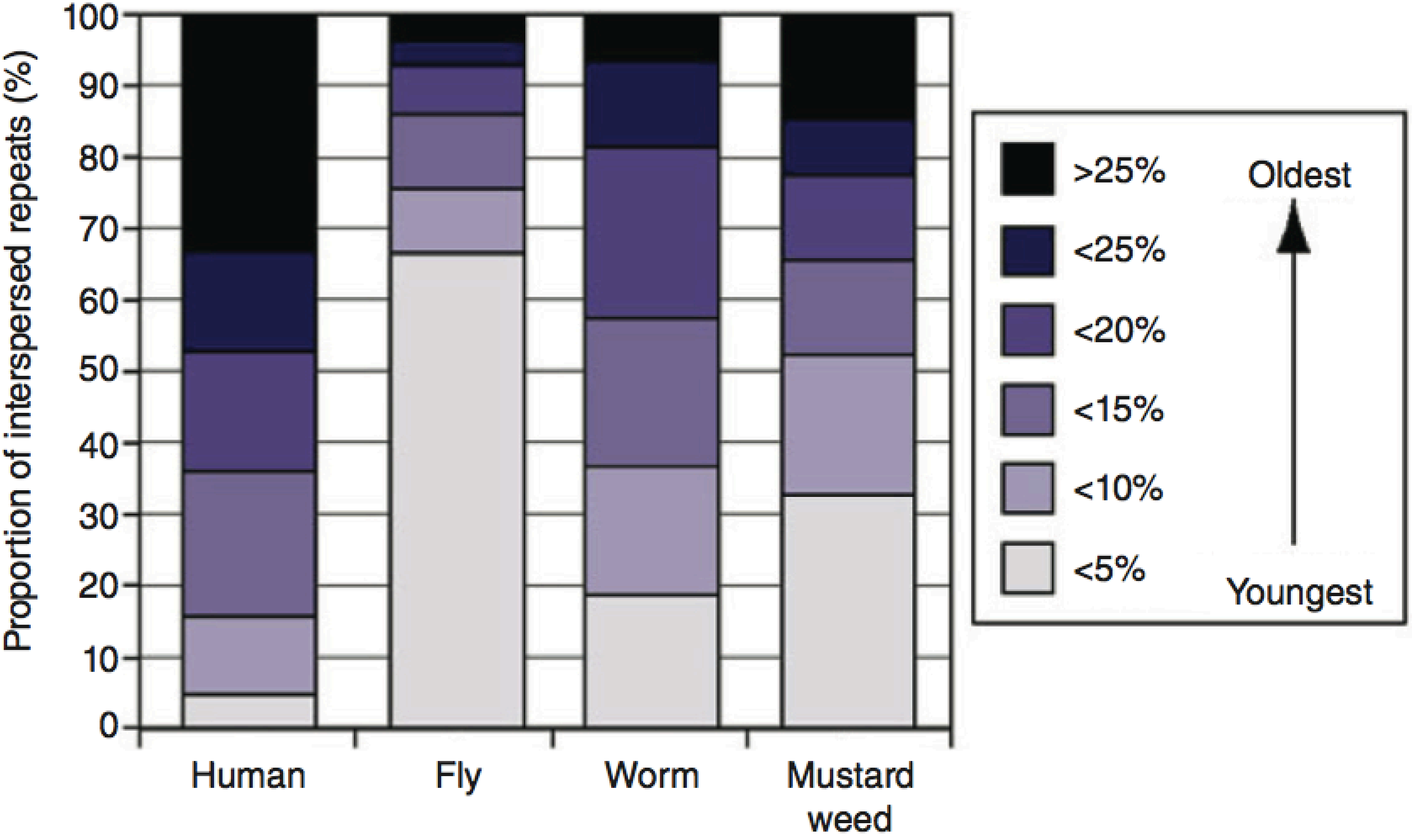


	Human		*Drosophila*		*C. elegans*		*A. thaliana*	
	Bases (%)	Families	Bases (%)	Families	Bases (%)	Families	Bases (%)	Families
LINE/SINE	33.4	6	0.7	20	0.4	10	0.5	10
LTR	8.1	100	1.5	50	0	4	4.8	70
DNA	2.8	60	0.7	20	5.3	80	5.1	80
Total	44.4	170	3.1	90	6.5	90	10.5	160

B&FG 3e
Table 20.5
Page 973

“Bases” refers to percentage of bases in the genome, “families” to approximate number of families in the genome.

Comparison of the age of interspersed repeats in four eukaryotic genomes



B&FG 3e
Fig. 20.10
Page 973

Humans have few recent interspersed repeats.

Human genome: simple sequence repeats

Simple sequence repeats (SSR) are perfect (or slightly imperfect) tandem repeats of *k*-mers. Microsatellites have *k* = 1 to 12, while minisatellites have k from about a dozen to 500 base pairs.

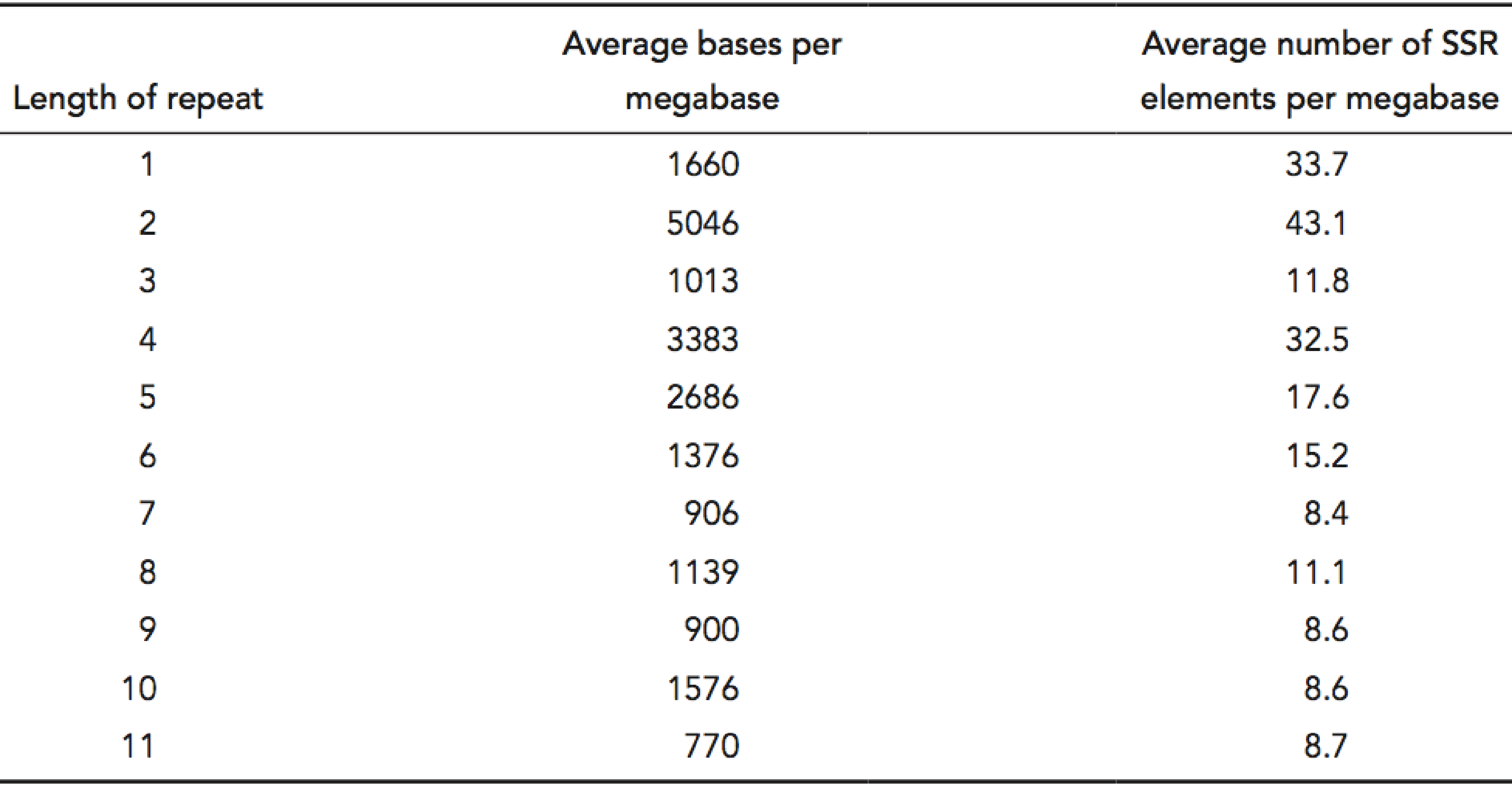
Micro- and minisatellites comprise 3% of the genome.

AC, AT, and AG are the most common dinucleotide repeats.

Length=	1	33.7 elements/megabase
	2	43.1
	3	11.8
	4	32.5

B&FG 3e
Page 973

Simple sequence repeats (SSRs; microsatellites) in the human genome



Length of repeat	Average bases per megabase	Average number of SSR elements per megabase
1	1660	33.7
2	5046	43.1
3	1013	11.8
4	3383	32.5
5	2686	17.6
6	1376	15.2
7	906	8.4
8	1139	11.1
9	900	8.6
10	1576	8.6
11	770	8.7

B&FG 3e
Table 20.6
Page 974

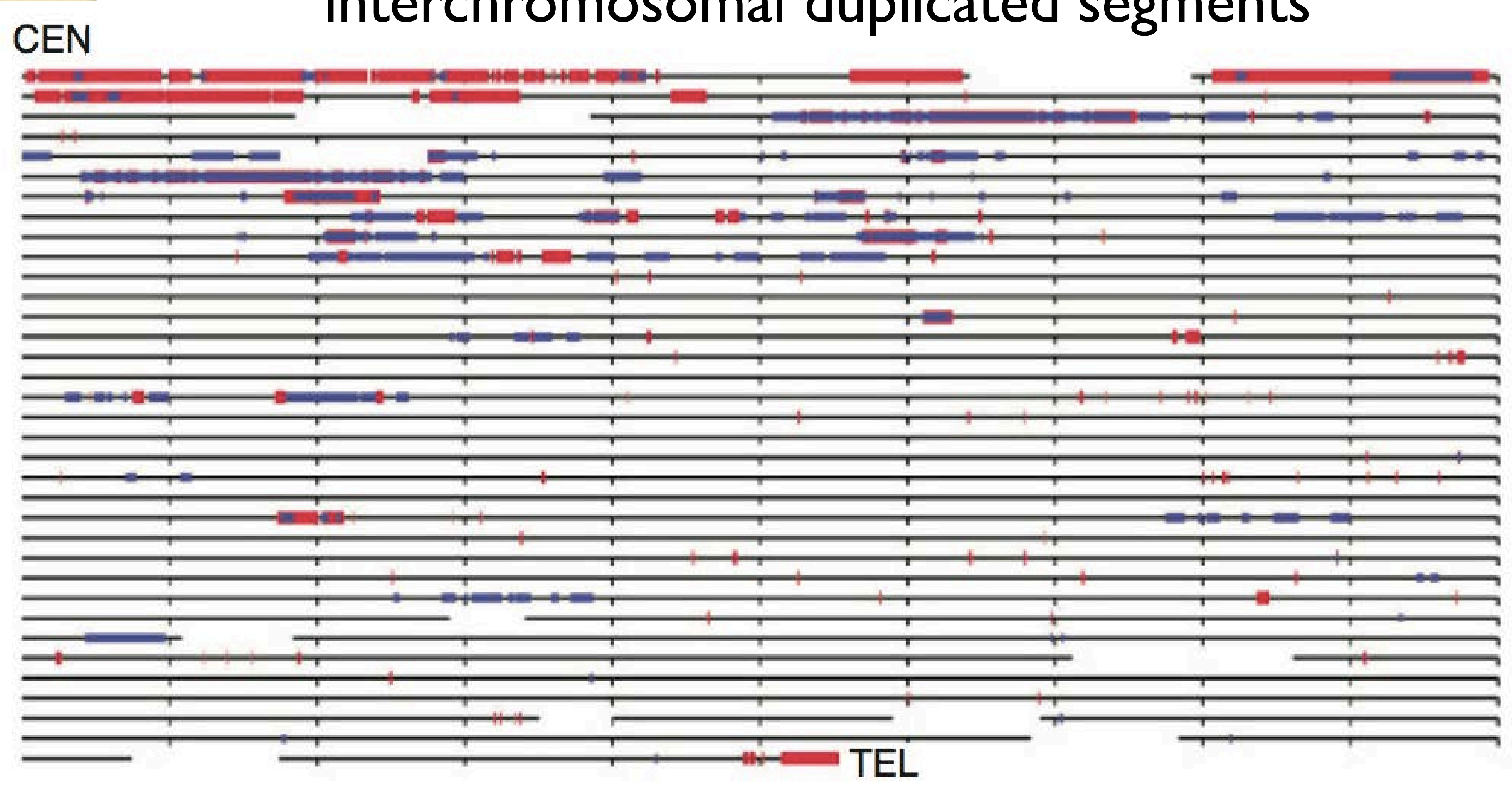
Human genome: segmental duplications

About 5% of the finished human genome sequence consists of segmental duplications, typically 10-50 kb.

Centromeres contain particularly large amounts of interchromosomally duplicated DNA.

B&FG 3e
Page 973

Centromeres consist of large amounts of interchromosomal duplicated segments



The size and location of intrachromosomal (black) and interchromosomal (red) segmental duplications are indicated. Each horizontal line represents 1 Mb of chromosome 22q; the tick marks indicate 100 kb intervals. The centromere is at top left, and the telomere is at the lower right.

B&FG 3e
Fig. 20.11
Page 974

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Human genome: gene content

As for any eukaryotic genome, gene prediction is difficult.

- For the human genome, the average exon is only 150 nucleotides. Thus they are hard to identify.
- Exon/intron borders can be difficult to assign.
- Introns may be many kilobases in length.
- Pseudogenes may be difficult to identify.
- Noncoding RNAs are also difficult to identify.

According to Ensembl (2017), there are ~20,300 protein coding genes.

Human genome: noncoding RNA

Noncoding RNAs include the following:

tRNA
rRNA
snoRNA
snRNA
miRNA

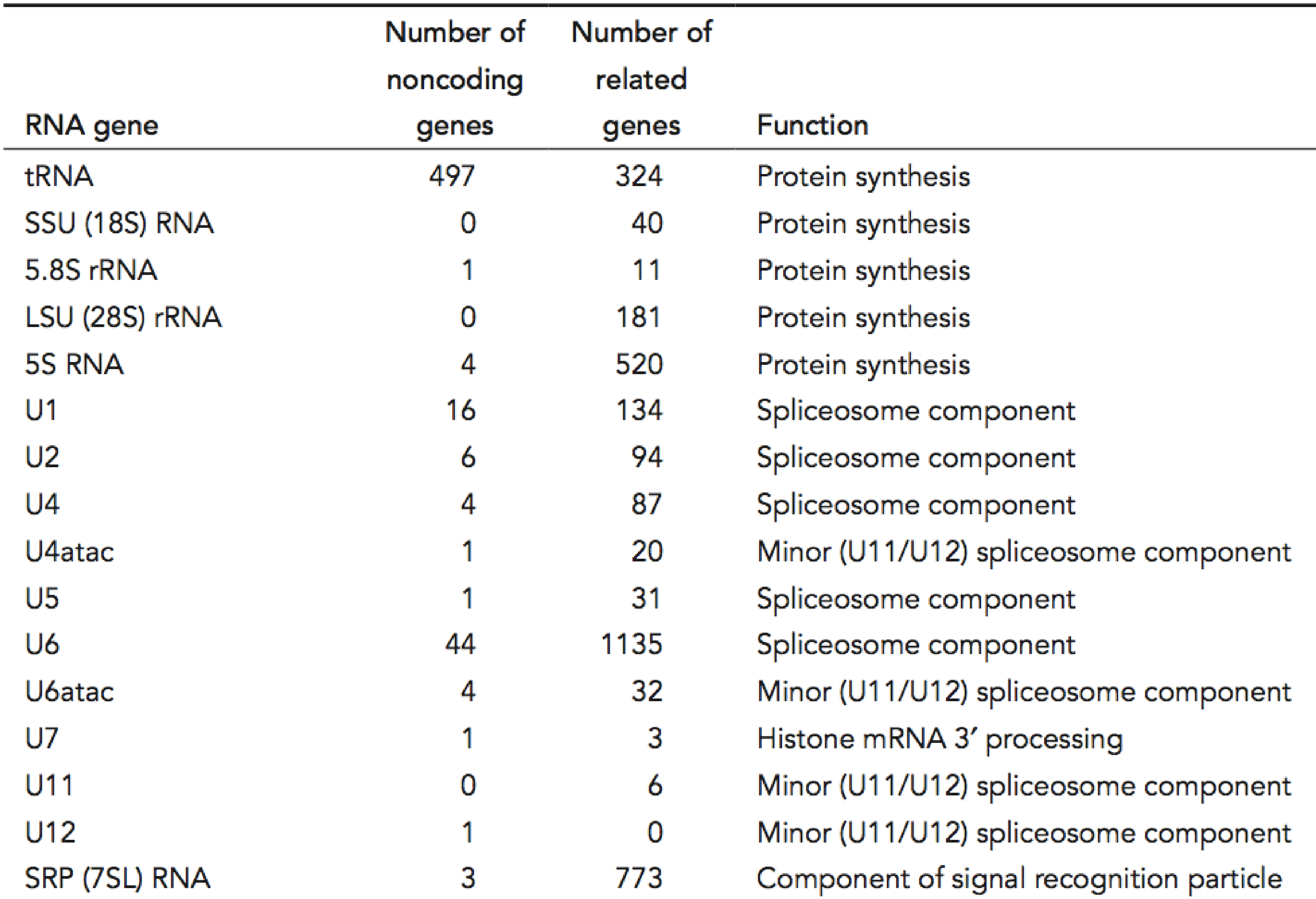
They can be difficult to identify because

-- they lack open reading frames
-- they may be extremely short
-- they are not polyadenylated

They have been identified using BLAST and various other tools (see Chapter 10).

B&FG 3e
Page 975

Noncoding genes in the human genome



RNA gene	Number of noncoding genes	Number of related genes	Function
tRNA	497	324	Protein synthesis
SSU (18S) RNA	0	40	Protein synthesis
5.8S rRNA	1	11	Protein synthesis
LSU (28S) rRNA	0	181	Protein synthesis
5S RNA	4	520	Protein synthesis
U1	16	134	Spliceosome component
U2	6	94	Spliceosome component
U4	4	87	Spliceosome component
U4atac	1	20	Minor (U11/U12) spliceosome component
U5	1	31	Spliceosome component
U6	44	1135	Spliceosome component
U6atac	4	32	Minor (U11/U12) spliceosome component
U7	1	3	Histone mRNA 3′ processing
U11	0	6	Minor (U11/U12) spliceosome component
U12	1	0	Minor (U11/U12) spliceosome component
SRP (7SL) RNA	3	773	Component of signal recognition particle

partial list

B&FG 3e
Table 20.7
Page 976

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

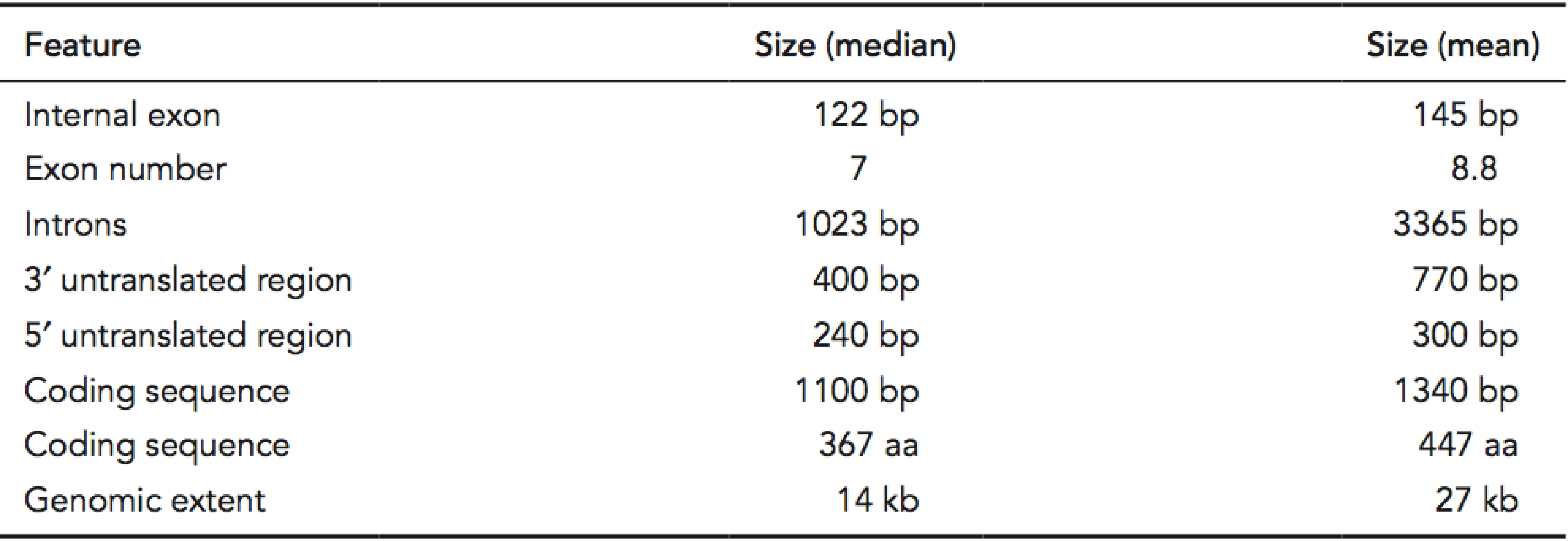
Human genome: protein coding genes

When RefSeq genes (manually curated) were compared to the draft human genome sequence, 92% could be aligned at high stringency over part of their length, and 85% could be aligned at ≥ half their length.

Some RefSeq genes had high stringency matches to multiple genomic locations. This could be due to paralogs, pseudogenes, or misassembly of the genome sequence.

Characteristics of human genes

- The basic characteristics of human genes include the following:

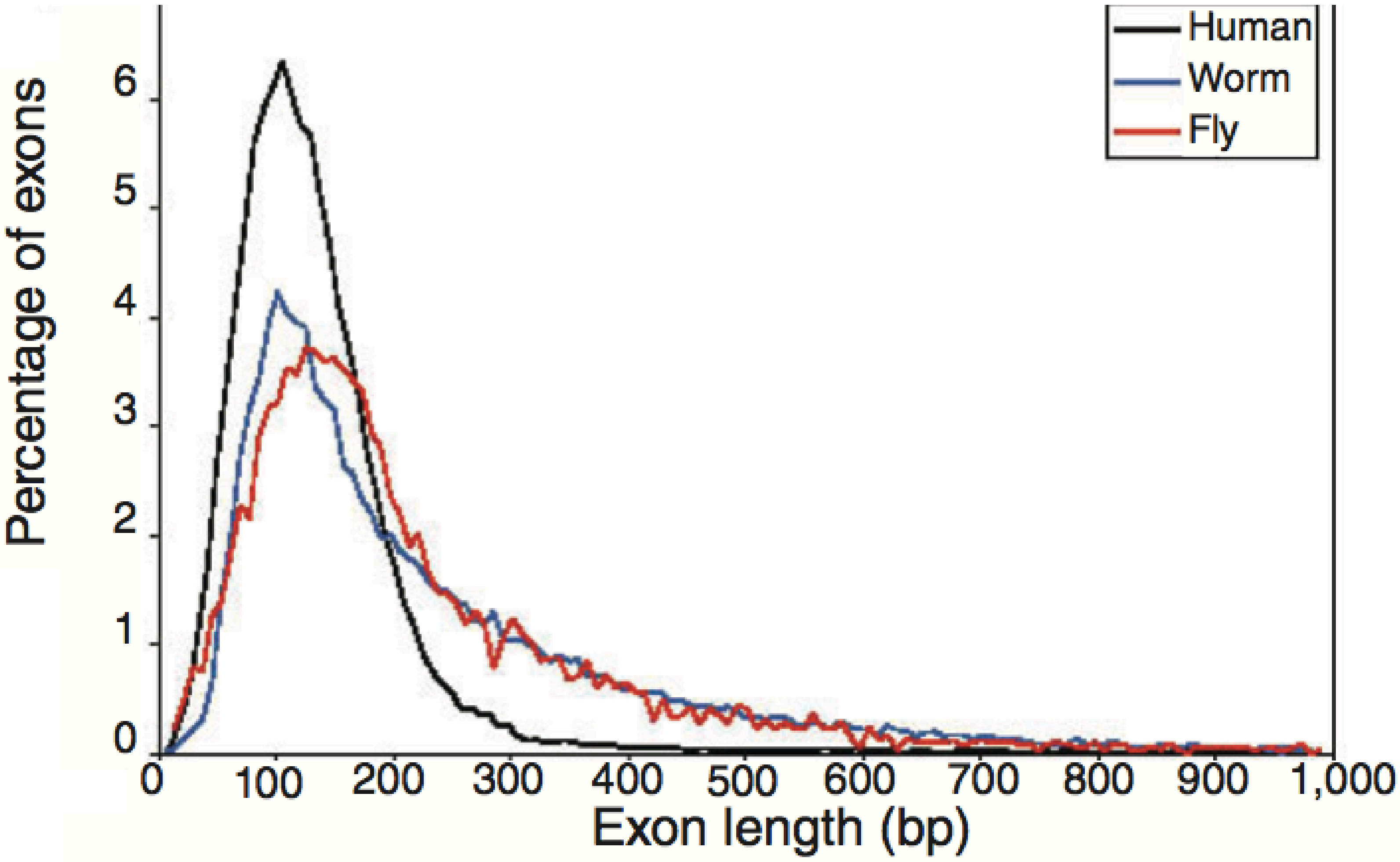


Feature	Size (median)	Size (mean)
Internal exon	122 bp	145 bp
Exon number	7	8.8
Introns	1023 bp	3365 bp
3′ untranslated region	400 bp	770 bp
5′ untranslated region	240 bp	300 bp
Coding sequence	1100 bp	1340 bp
Coding sequence	367 aa	447 aa
Genomic extent	14 kb	27 kb

B&FG 3e
Table 20.8
Page 976

aa: amino acids; bp: base pairs; kb: kilo base pairs

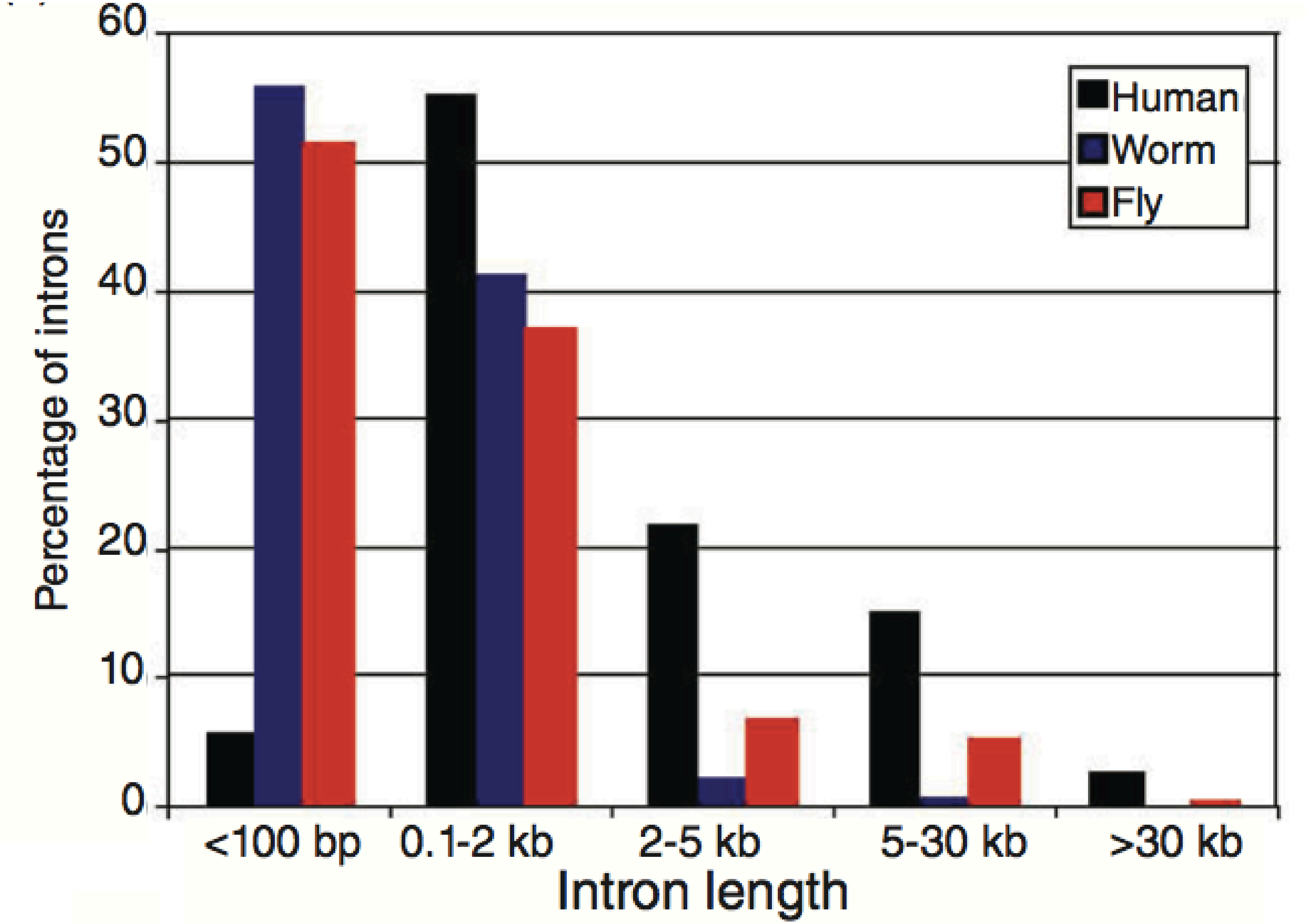
Size distribution of exons, introns, and short introns in human, worm, and fly



The average coding sequence for human genes is 1340 bp, comparable to the size of an average coding sequence in nematode and *Drosophila*. Most internal exons are about 50–200 bp in length in all three species (plot above), although worm and fly have a greater proportion of longer exons.

B&FG 3e
Fig. 20.12
Page 977

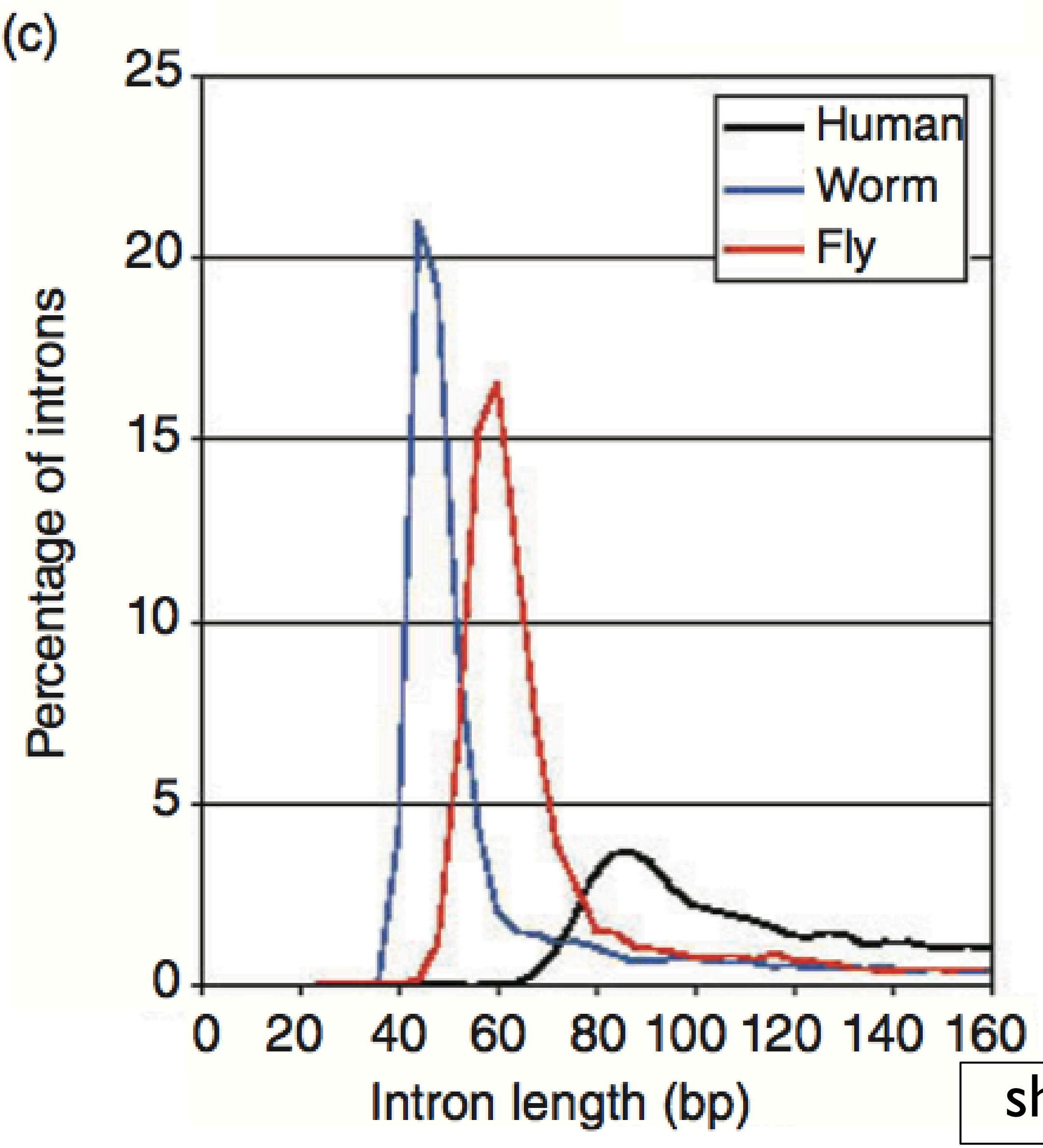
Size distribution of exons, introns, and short introns in human, worm, and fly



B&FG 3e
Fig. 20.12
Page 977

The size of human introns is far more variable than in fly and worm.

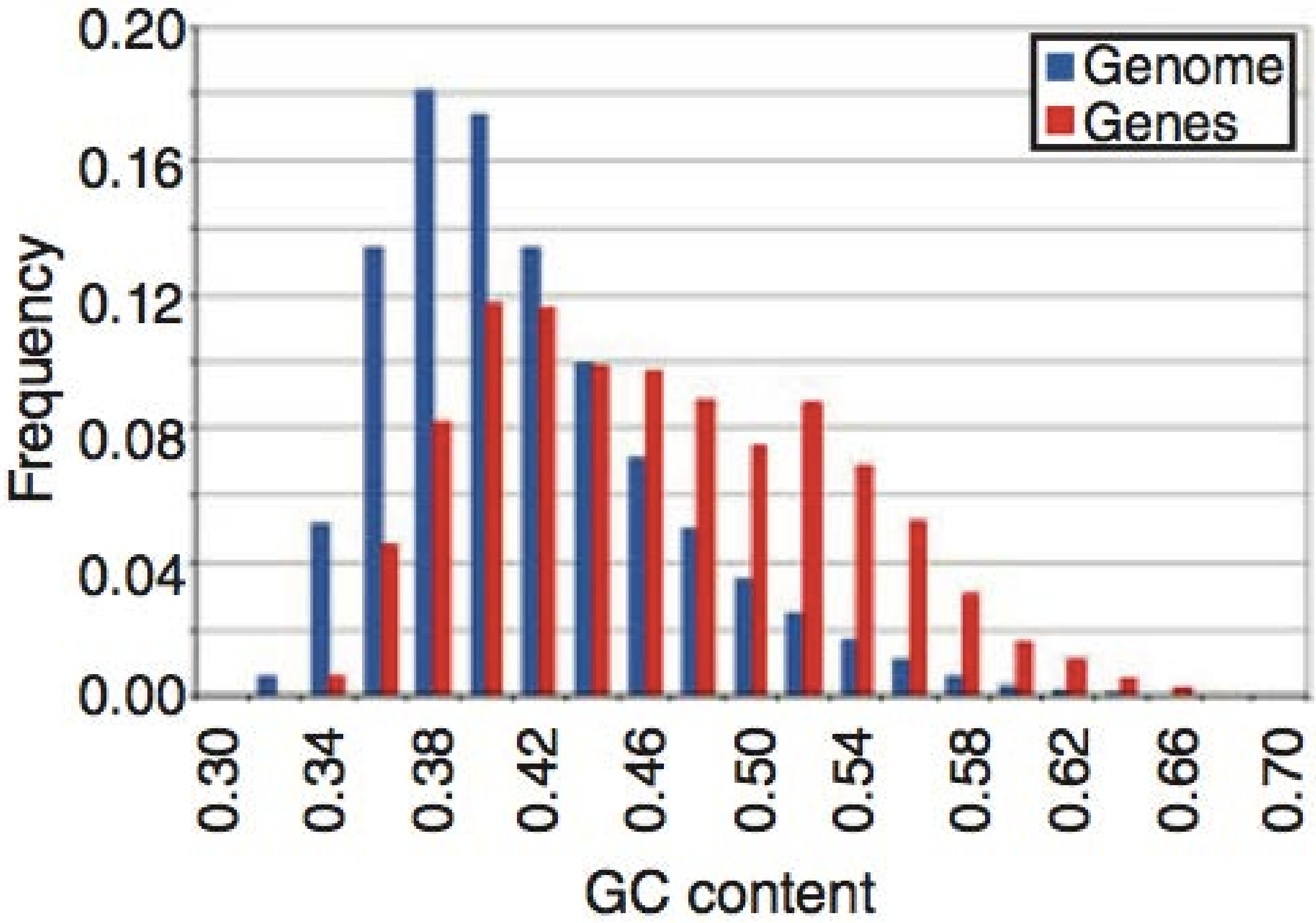
Size distribution of exons, introns, and short introns in human, worm, and fly



short introns

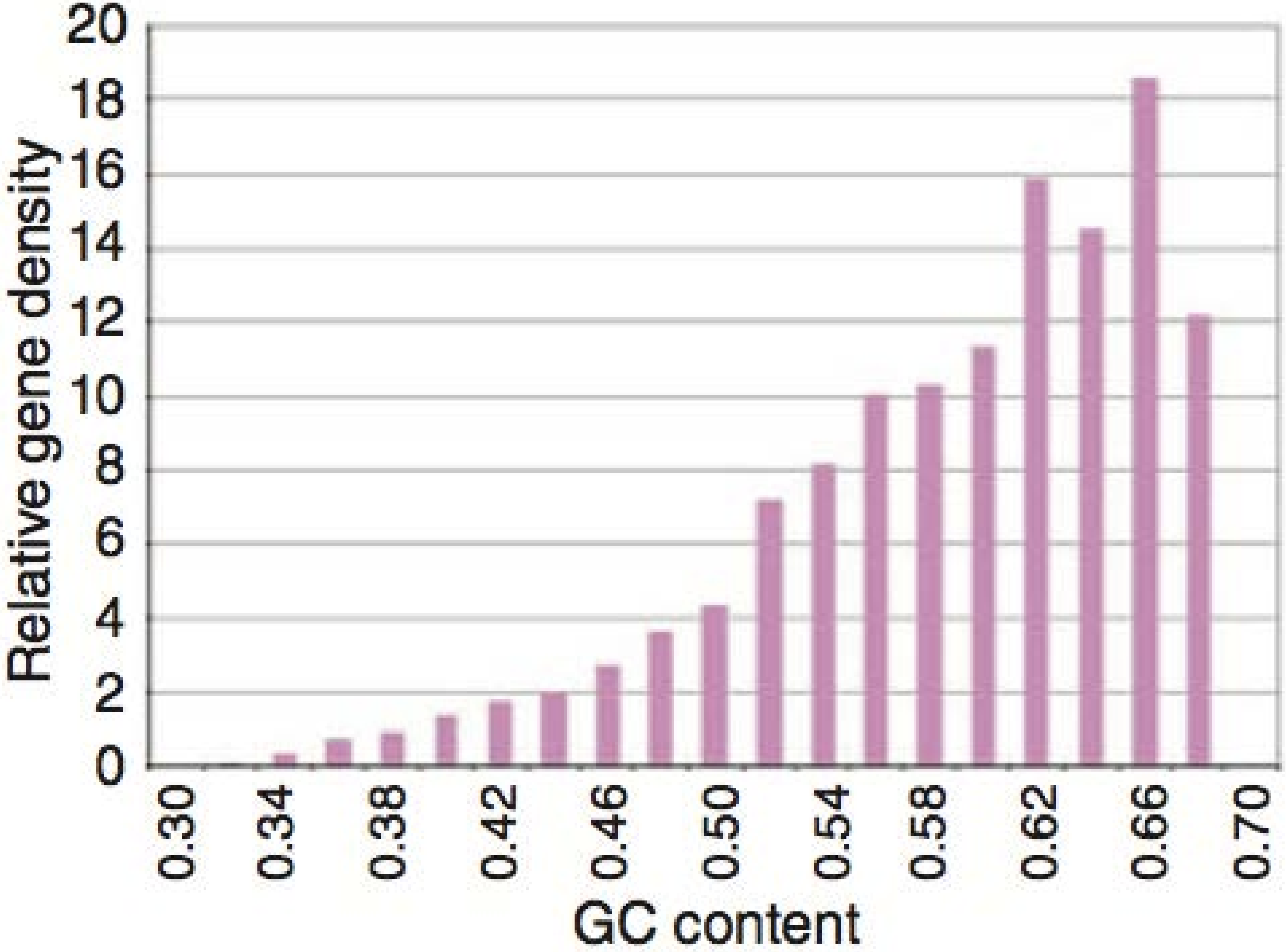
B&FG 3e
Fig. 20.12
Page 977

Distribution of GC content in genes and in the genome: protein-coding genes are associated with higher GC content



B&FG 3e
Fig. 20.13
Page 978

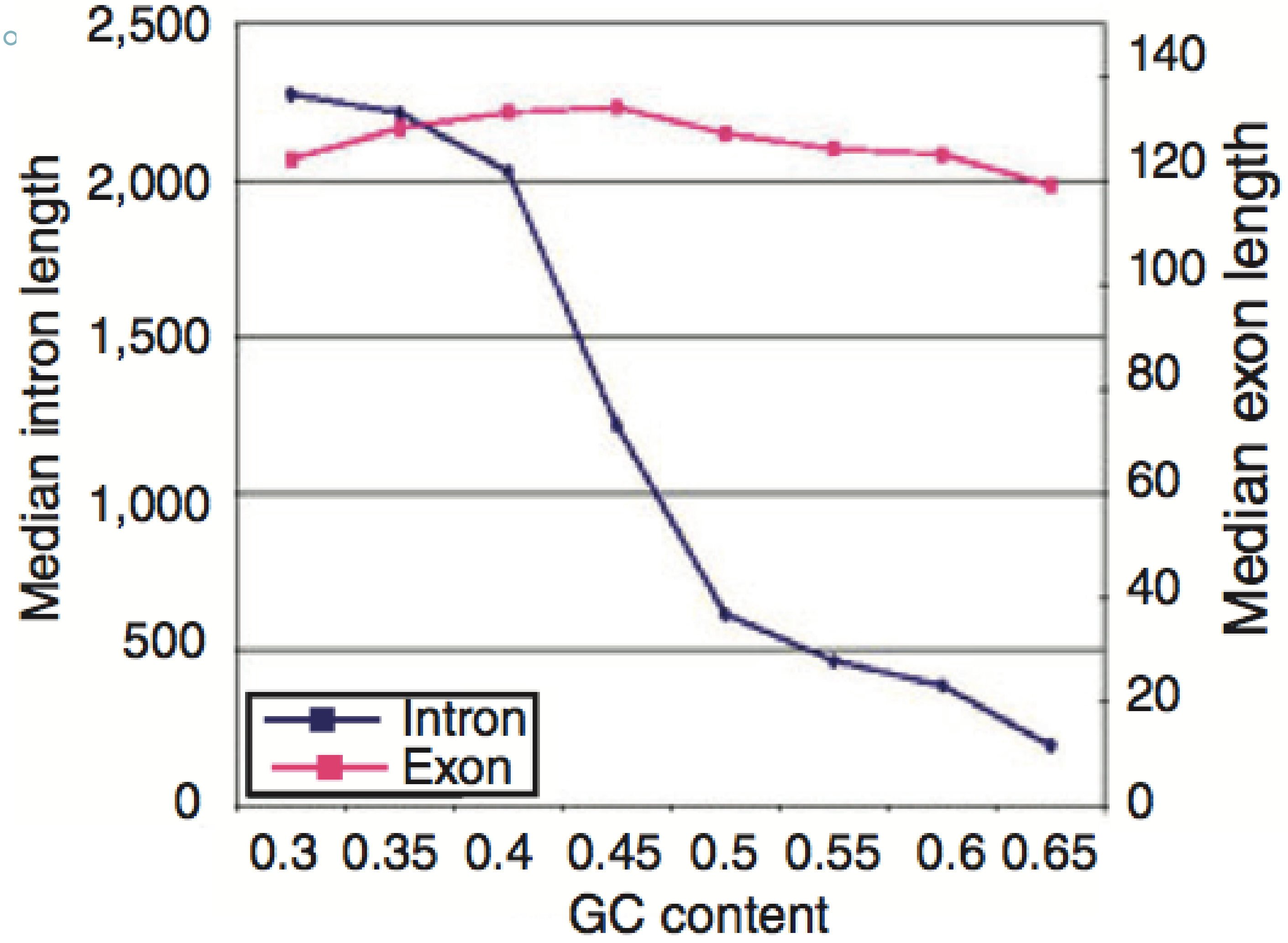
Gene density plotted as a function of the GC content



B&FG 3e
Fig. 20.13
Page 978

As GC content rises, relative gene density increases dramatically.

Median exon length is unaffected by GC content, but introns are far shorter as GC content rises



B&FG 3e
Fig. 20.13
Page 978

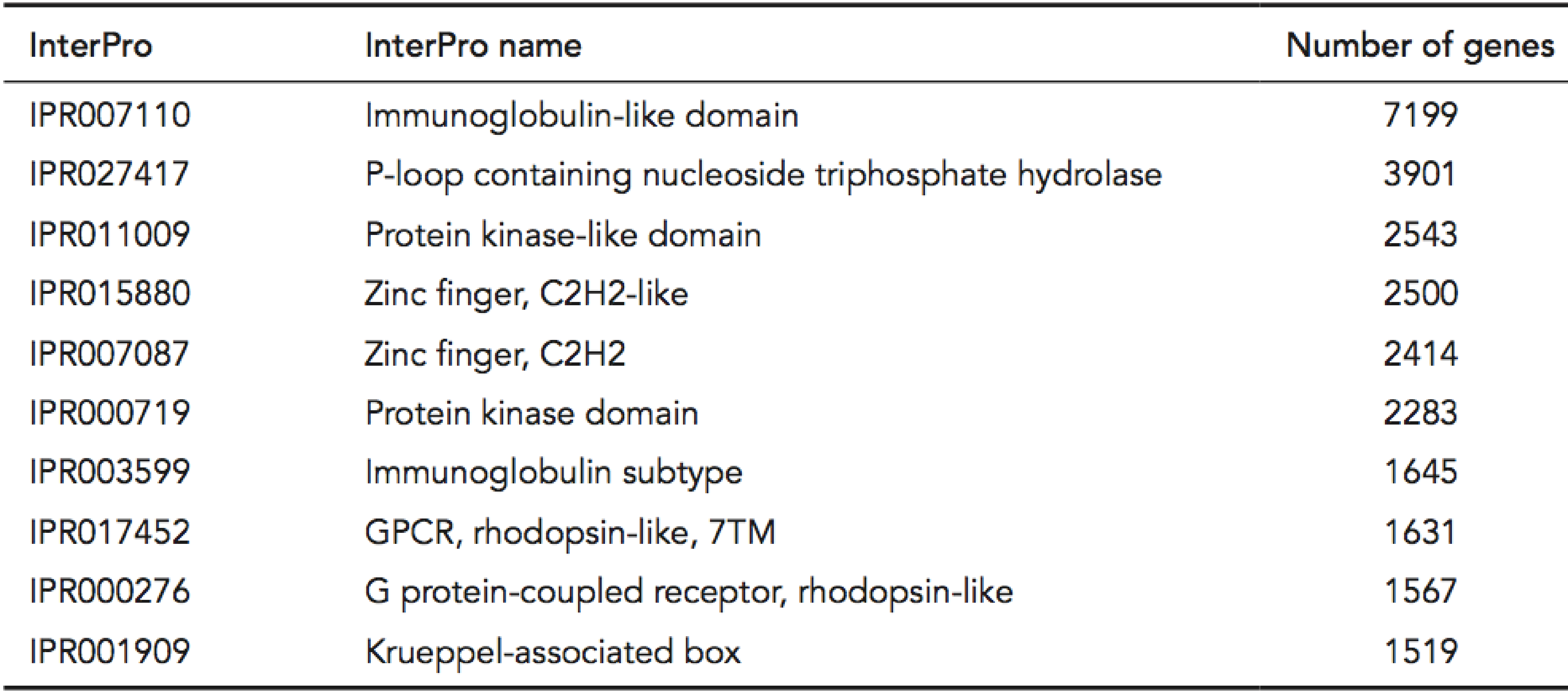
The human proteome

The number of protein-coding genes in humans is comparable to the number of genes in other metazoans and plants and only three-fold greater than the number in unicellular fungi.

While humans do not have more protein-coding (or noncoding) genes than other "lower" organisms, the human proteome may be more complex.

B&FG 3e
Page 978

Ten most common InterPro hits for *Homo sapiens*



InterPro	InterPro name	Number of genes
IPR007110	Immunoglobulin-like domain	7199
IPR027417	P-loop containing nucleoside triphosphate hydrolase	3901
IPR011009	Protein kinase-like domain	2543
IPR015880	Zinc finger, C2H2-like	2500
IPR007087	Zinc finger, C2H2	2414
IPR000719	Protein kinase domain	2283
IPR003599	Immunoglobulin subtype	1645
IPR017452	GPCR, rhodopsin-like, 7TM	1631
IPR000276	G protein-coupled receptor, rhodopsin-like	1567
IPR001909	Krueppel-associated box	1519

B&FG 3e
Tab. 20.9
Page 978

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

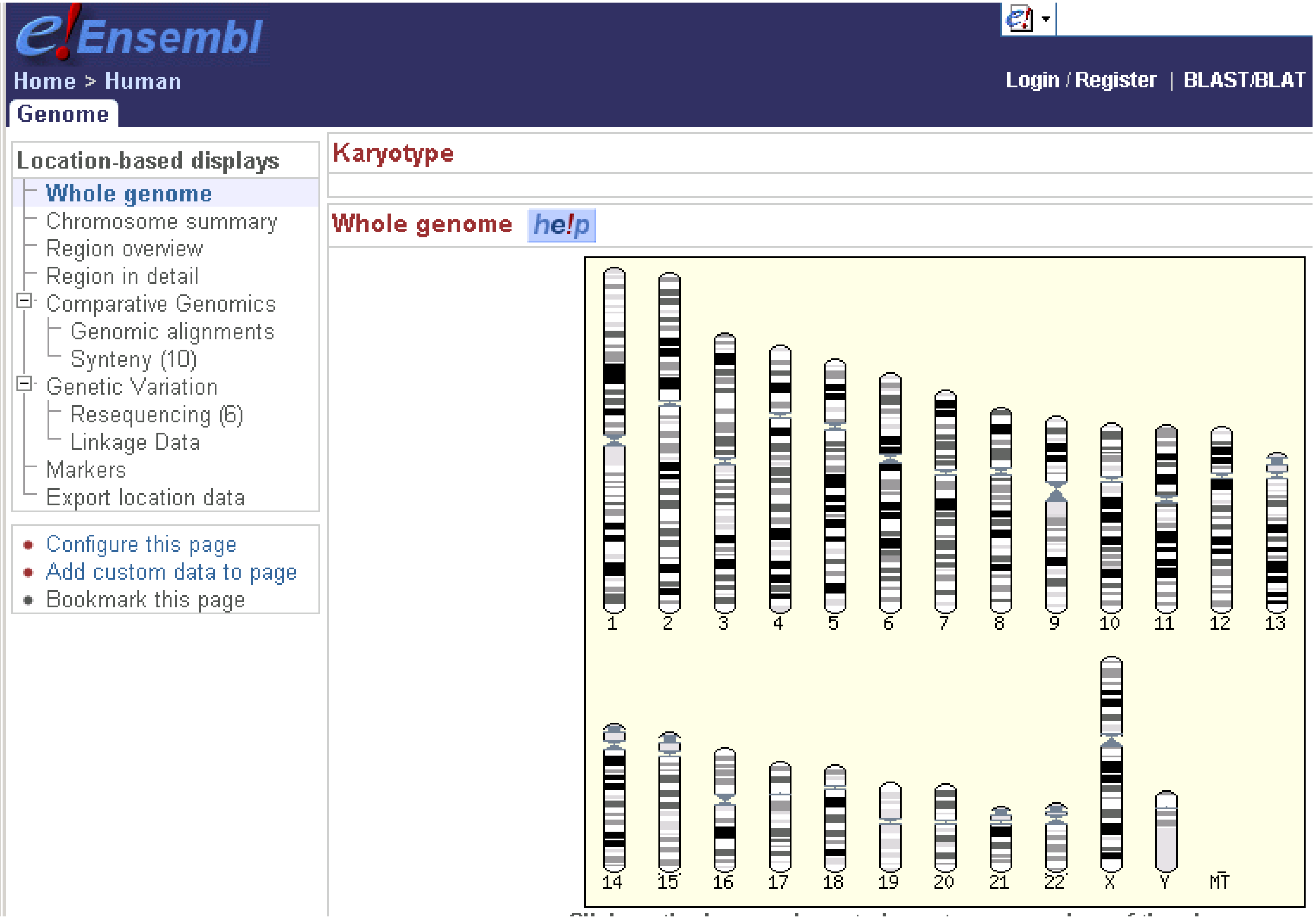
- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Human karyotype from Ensembl



e!Ensembl

Home > Human

Login / Register | BLAST/BLAT

Genome

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Comparative Genomics
  - Genomic alignments
  - Synteny (10)
- Genetic Variation
  - Resequencing (6)
  - Linkage Data
- Markers
- Export location data

- Configure this page
- Add custom data to page
- Bookmark this page

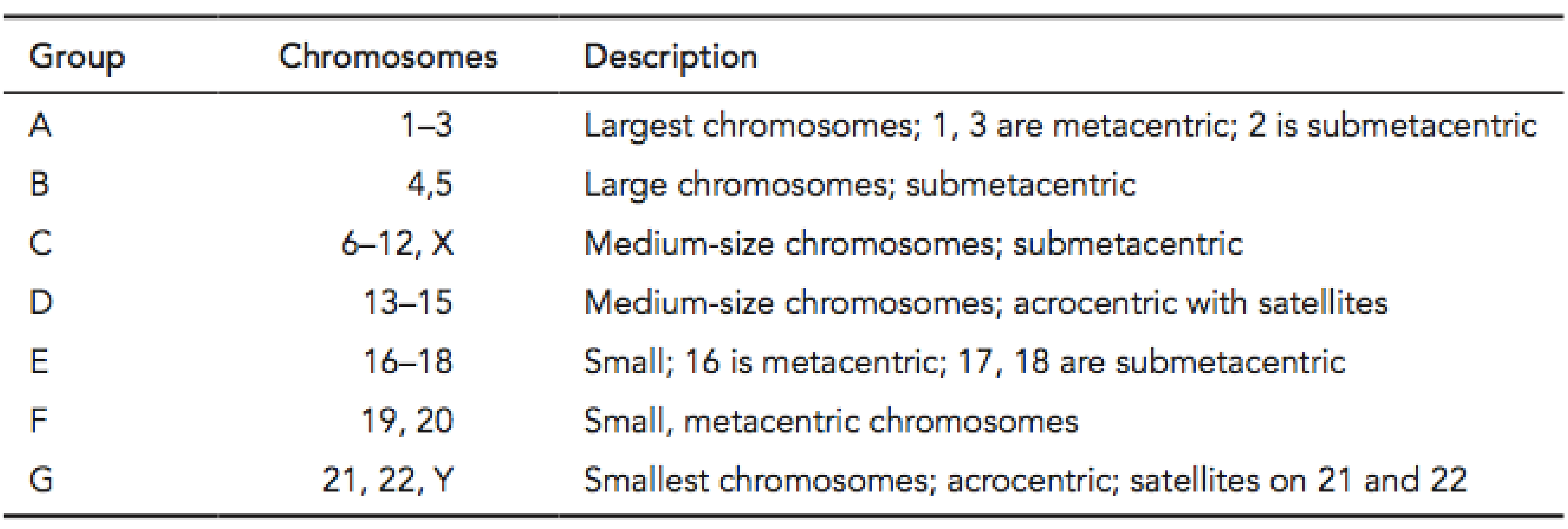
Karyotype

Whole genome help

We'll examine the 25 human chromosomes: 1-22, X, Y, and mitochondrial. Ensembl offers a good starting point.

B&FG 3e
Page 979

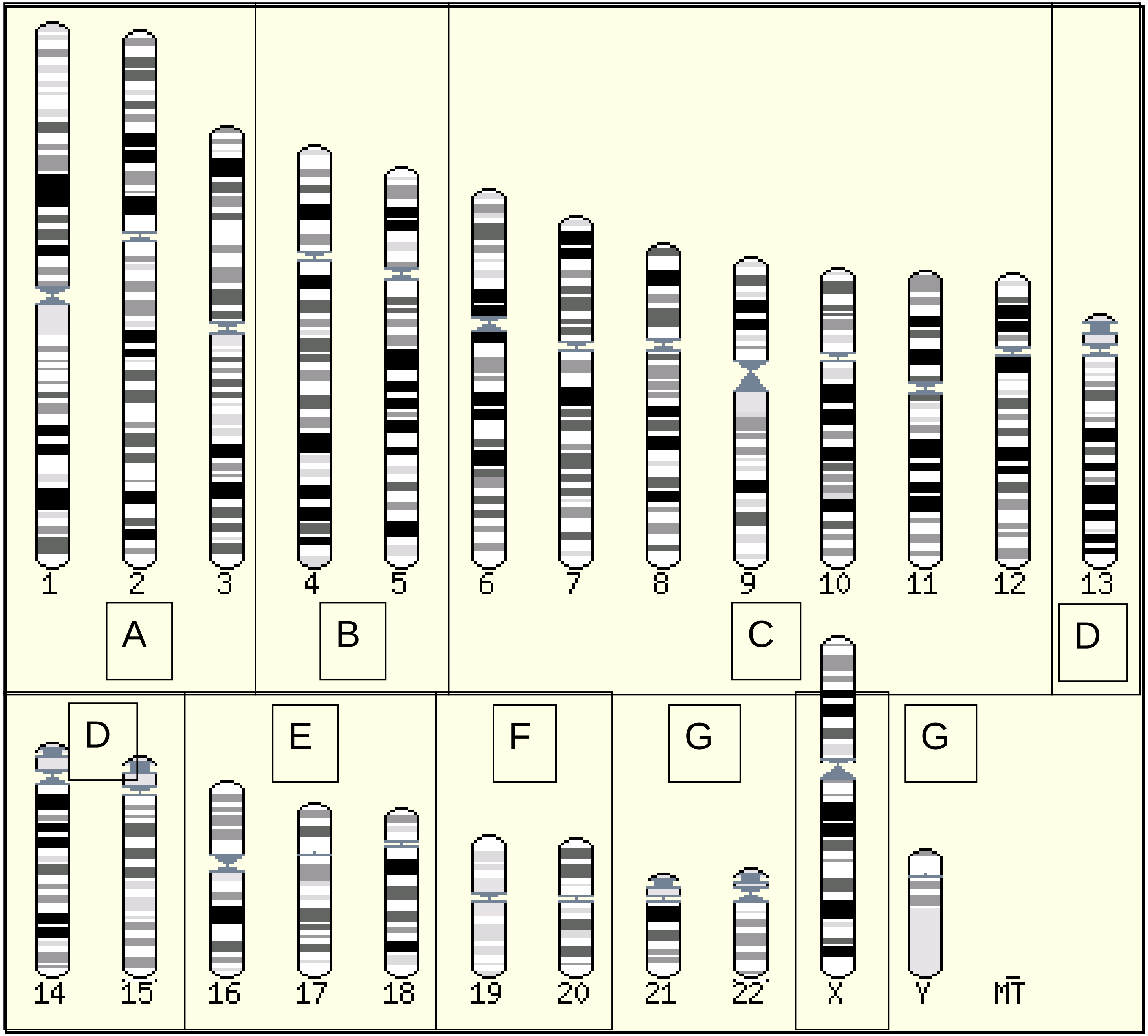
Human chromosome groups

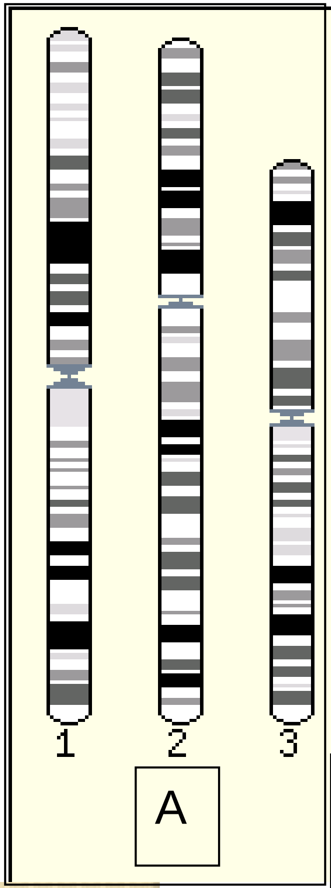


Group	Chromosomes	Description
A	1–3	Largest chromosomes; 1, 3 are metacentric; 2 is submetacentric
B	4,5	Large chromosomes; submetacentric
C	6–12, X	Medium-size chromosomes; submetacentric
D	13–15	Medium-size chromosomes; acrocentric with satellites
E	16–18	Small; 16 is metacentric; 17, 18 are submetacentric
F	19, 20	Small, metacentric chromosomes
G	21, 22, Y	Smallest chromosomes; acrocentric; satellites on 21 and 22

There are seven traditional cytogenetic groups A–G which categorize the chromosomes (other than the mitochondrial genome) according to size and morphological properties.

B&FG 3e
Tab. 20.11
Page 980


1
2
3
4
5
6
7
8
9
10
11
12
13
A
B
C
D
D
E
F
G
G
14
15
16
17
18
19
20
21
22
X
Y
MT

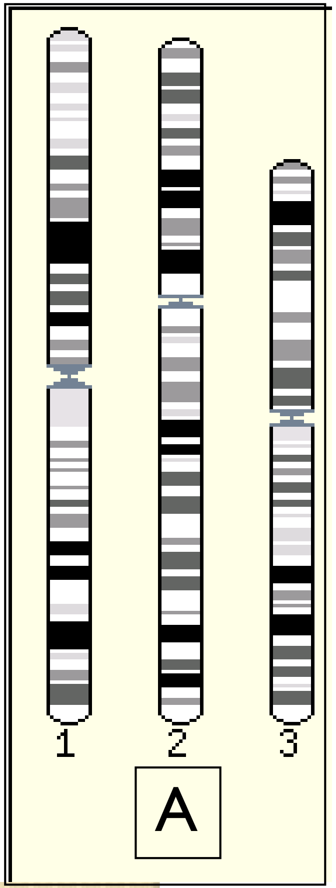


Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
1	249 Mb	3,141	991	24.0	NC_000001.10
2	243 Mb	1,346	1,239	5.0	NC_000002.11
3	198 Mb	1,463	122	3.2	NC_000003.11

Chromosome 1 gene density (14.2 genes per megabase) is nearly twice the genome-wide average (7.8 genes per megabase).

Typical for essentially all the chromosome finishing projects, sequence integrity and completeness were assessed three ways: (1) by determining whether all RefSeq genes assigned to the chromosome were accounted for, (2) by comparing the order of hundreds of chromosome markers to the DeCode genetic map to search for discrepancies, and (3) by aligning over 32,000 pairs of fosmid end sequences to unique positions in the sequence. This resulted in the identification of several misassemblies caused by low-copy repeats. Naturally occurring polymorphisms can confound the analysis; for example, 50% of individuals lack the *GSTM1* gene.

B&FG 3e
Tab. 20.12
Page 980

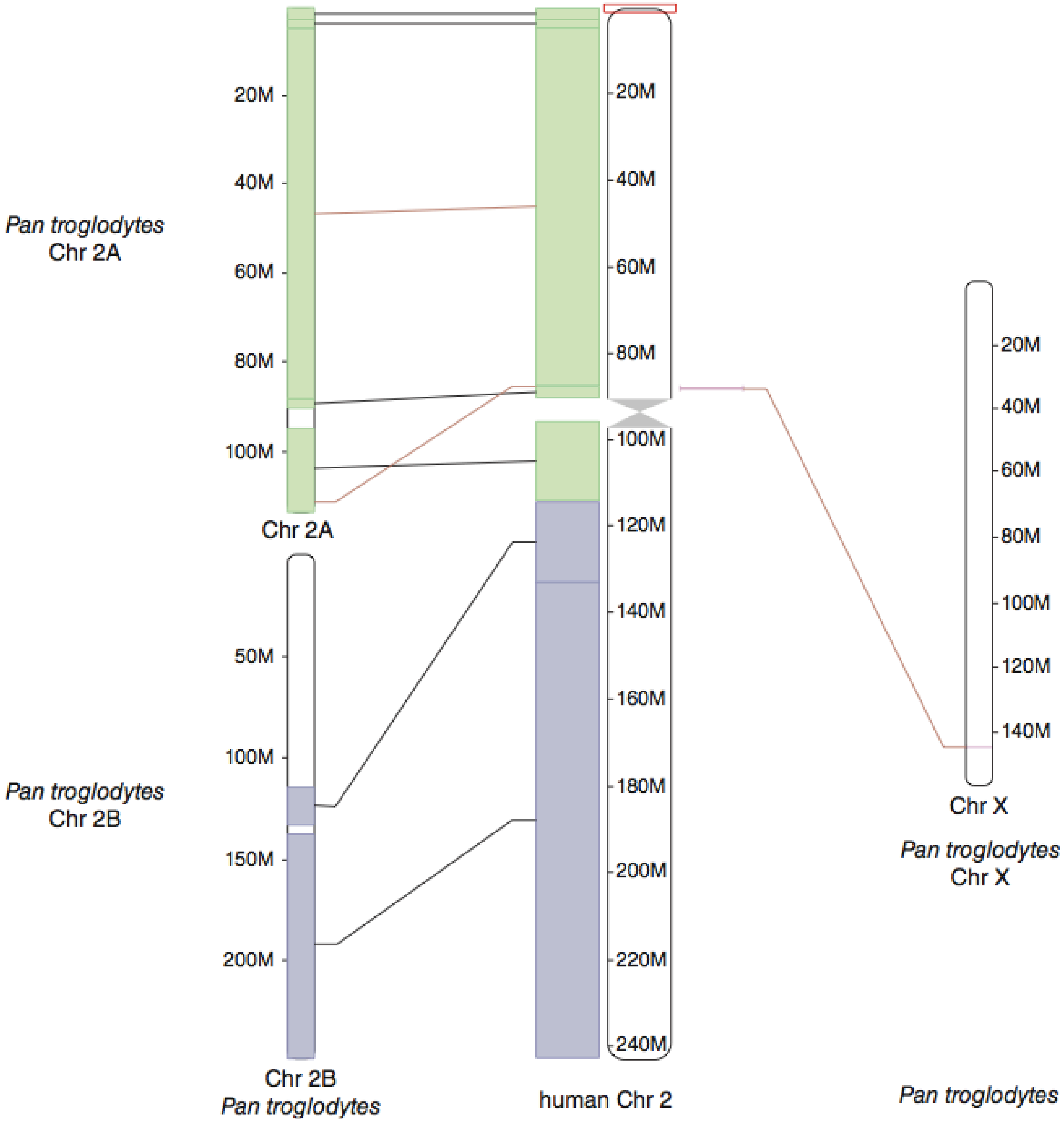


Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
1	249 Mb	3,141	991	24.0	NC_000001.10
2	243 Mb	1,346	1,239	5.0	NC_000002.11
3	198 Mb	1,463	122	3.2	NC_000003.11

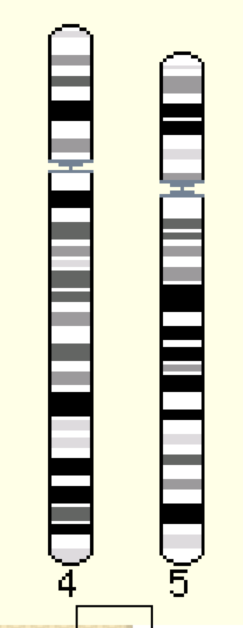
Chromosome 2 corresponds to two intermediate-sized ancestral, acrocentric chromosomes that fused head-to-head. In other primates these chromosomes remained separate, as in the case of chimpanzee chromosomes 2A and 2B. The fusion site has been localized to 2q13–2q14.1. One of the two centromeres (at 2q21) became inactivated, and contains α-satellite remnants.

Chromosome 3 contains the lowest rate of segmental duplications in the genome (1.7% compared to a genome-wide average of 5.3% of nucleotides segmentally duplicated). Chromosomes 3 and 21 derive from a larger ancestral chromosome that split.

Conserved synteny between human chromosome 2 and two smaller chimpanzee chromosomes provides evidence that two ancestral human acrocentric chromosomes fused



B&FG 3e
Fig. 20.14
Page 982



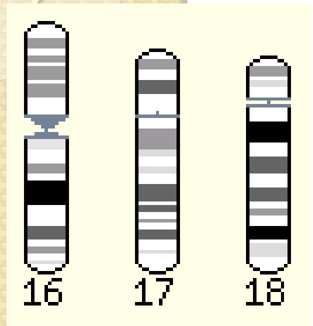
B

Chrom.	Length	# Genes	# Pseudo -genes	Gap size (Mb)	Accession
4	191 Mb	796	778	3.5	NC_000004.11
5	181 Mb	923	577	3.2	NC_000005.9

Chromosome 4 has an unusually low GC content of 38.2%, compared to the genome-wide average of 41%. Try viewing this at the UCSC Genome Browser.

Chromosome 5 has both a very low gene density and a very high rate of intrachromosomal duplications.

B&FG 3e
Tab. 20.13
Page 980



E

Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
16	90 Mb	796	778	11.5	NC_000016.9
17	81 Mb	1,266	274	3.4	NC_000017.10
18	78 Mb	337	171	3.4	NC_000018.9

Chromosome 18 has the lowest gene density of any autosome (4.4 genes per megabase) and encodes only 337 genes (about a quarter of the number of the similar-sized chromosome 17). One region of chromosome 18 has only 3 genes across 4.5 megabases. The sparse number of genes may explain why some individuals with trisomy 18 (Edwards syndrome) survive to birth, while all other autosomal trisomies (except trisomy 13 and trisomy 21) are embryonic lethal.

B&FG 3e
Tab. 20.16
Page 981

F

Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
19	59 Mb	1,461	321	3.3	NC_000019.9
20	63 Mb	727	168	3.5	NC_000020.10

Chromosome 19 has the highest gene density with 26 protein-coding genes per megabase.

B&FG 3e
Tab. 20.17
Page 981

G

Chrom.	Length	# Genes	# Pseudo-genes	Gap size (Mb)	Accession
21	48 Mb	796	778	13.0	NC_000021.8
22	51 Mb	545	134	16.4	NC_000022.10
Y	59 Mb	78	n/a	33.7	NC_000024.9

The Y chromosome was the most technically difficult to sequence because of its extraordinarily repetitive nature. It has short pseudoautosomal regions at the ends that recombine with the X chromosome. A large central region, spanning 95% of its length, is termed the male-specific region (MSY). There are 23 megabases of euchromatin including 8 Mb on Yp and 14.5 Mb on Yq. There are three notable heterochromatic regions:
(1) a centromeric region of about 1 Mb,
(2) a block of ~40 Mb on the long arm, and
(3) an island of 400 kilobases comprised of over 3,000 tandem repeats of 125 base pairs.
Of 156 transcription units, about half encode proteins.

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Human genome variation

We conclude our topic by exploring variation in the human genome, including single-nucleotide polymorphisms (SNPs) and the International HapMap project; the 1000 Genomes Project; and the sequencing of individual human genomes.

SNPs, Haplotypes, and HapMap

SNPs represent a fundamental form of variation in the human population. We can document:

- the number of SNPs. Each person has 3-4 million single nucleotide variants (SNVs).
- the sequence (major and minor allele)
- the copy number
- the allele frequencies (some SNPs occur more often in people from particular geographic backgrounds)
- the relationship to neighboring SNPs: linkage disequilibrium (LD) occurs when SNPs are tightly linked to each other and form blocks. The behavior of one SNP (a "tag SNP") can serve as a proxy for the genotypes of neighboring SNPs.

B&FG 3e
Page 987

Viewing and Analyzing SNPs and Haplotypes

SNP data can be visualized with many tools, including:

- browsers at UCSC, NCBI, and Ensembl;
- HaploView;
- Integrative Genomics Viewer (IGV) from the Broad;
- PLINK

Major Conclusions of HapMap Project

The HapMap project contributed basic knowledge of human genetic variation. Conclusions included:

- Most variation occurs in African populations.
- Linkage disequilibrium displays a block-like structure
- Some regions (e.g. centromeres) are characterized by lack of recombination across extended haplotype structures.
- SNPs are useful for genome-wide association studies
- Natural selection can remove deleterious mutations and preserve (fix) advantageous variants.
- The prevalence of structural variation can be measured through SNP analysis.

B&FG 3e
Page 994

Major Conclusions of 1000 Genomes Project

The goal of the 1000 Genomes Project was to create a comprehensive resource on human genetic variation. It is significant as the first publicly available whole-genome sequence dataset on the population scale. One specific aim was to identify most (>95%) of the genetic variants that have at least a 1% frequency in the populations being studied. Major conclusions:

- High rates of variation occur at the HLA and subtelomeric regions. Lowest rates occur in a 5 Mb, gene-dense region around 3p21.
- SNPs were imputed to support GWAS.
- The number of variants has been described for different functional classes.

B&FG 3e
Page 995

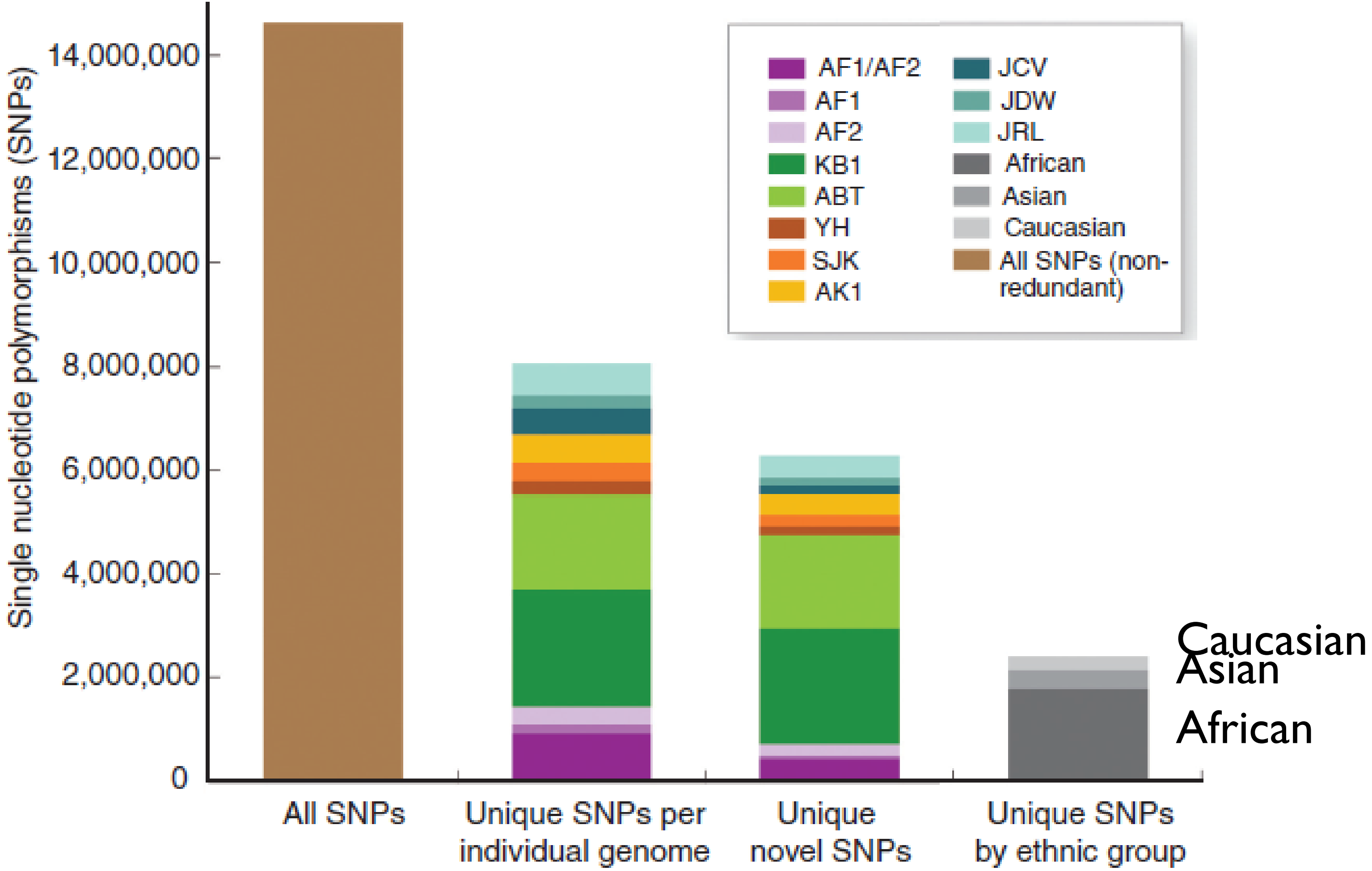
Variation: Sequencing Individual Genomes

- The cost of genome sequencing continues to decline.
- Individual genome sequencing has the potential to facilitate the start of an era of individualized medicine in which DNA changes that are associated with a disease condition are identified.
- In 2007 the first two individual genomes were reported: J. Craig Venter and James Watson.
- Each genome harbors 3-4 million SNVs and perhaps 600,000 structural variants.
- Trio sequencing shows that each individual harbors on the order of 100 de novo variants.

B&FG 3e
Page 998

SNPs identified in 10 personal genomes

All SNPs in each genome were compared with the 9 others



B&FG 3e
Fig. 20-22
Page 1000

Gonzaga-Jauregui et al. (2012) PMID: 22248320

Outline

Introduction

Main conclusions of human genome project

Gateways to access the human genome

Human Genome Project

- Background
- Strategic issues
- Human genome assemblies
- Broad genomic landscape
- Repeat content of human genome
- Gene content of human genome

25 human chromosomes

Human genome variation

Perspective

Perspective

Two major technological advances enabled the human genome to be sequenced: (1) the invention of automated DNA sequencing machines in the 1980s allowed nucleotide data to be collected on a large scale; and (2) the computational biology tools necessary to analyze those sequence data were created by biologists and computer scientists.

The human genome project was completed c.2003. The first individual genome sequence was reported in 2007 (that of J. Craig Venter). A decade later over 100,000 genomes have been sequenced, and over 1 million exomes. We can expect to continue to catalog variation and to further relate genotype to phenotype.

B&FG 3e
Page 999