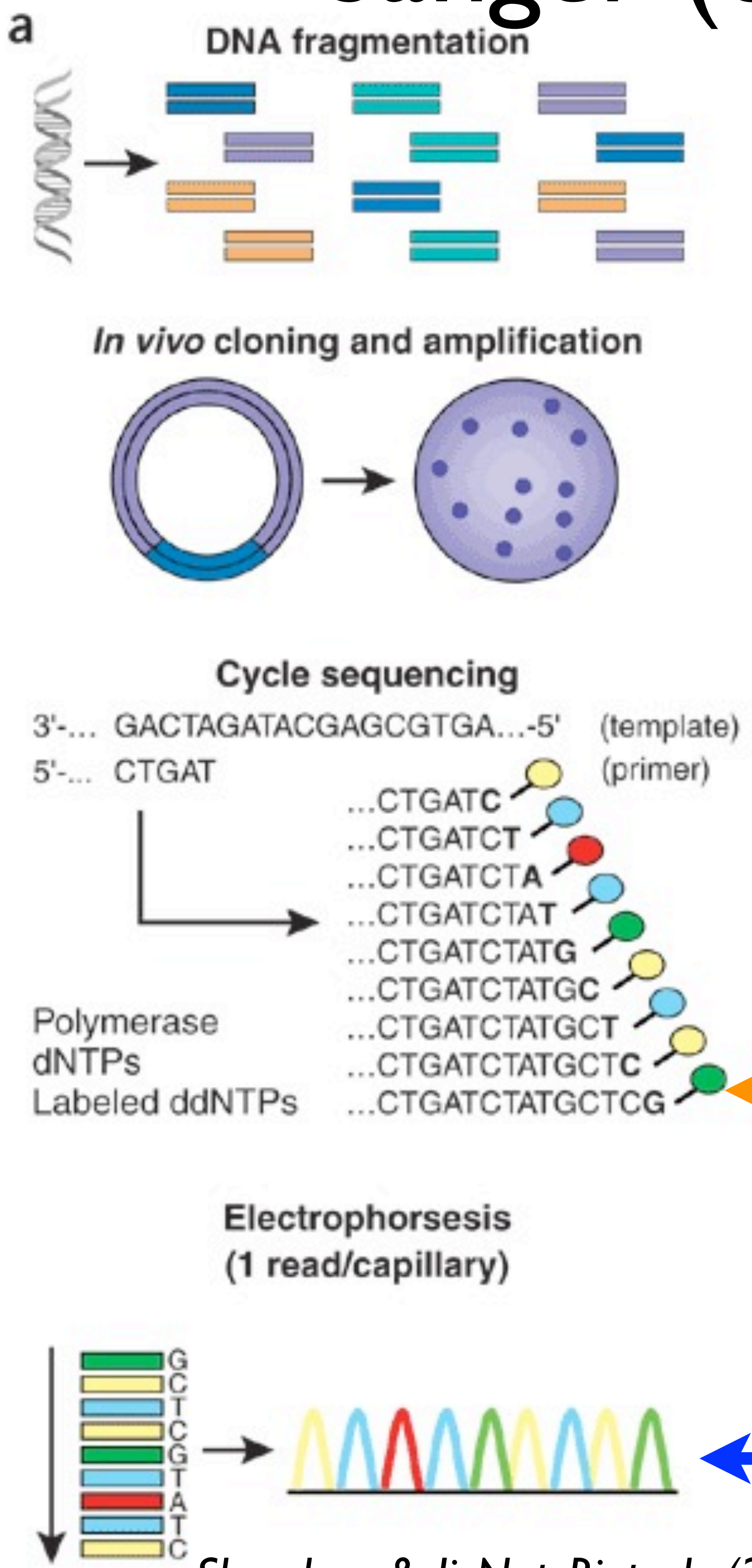
Sequencing technologies: Sanger (capillary sequencing)



1 read at a time,
~1kb

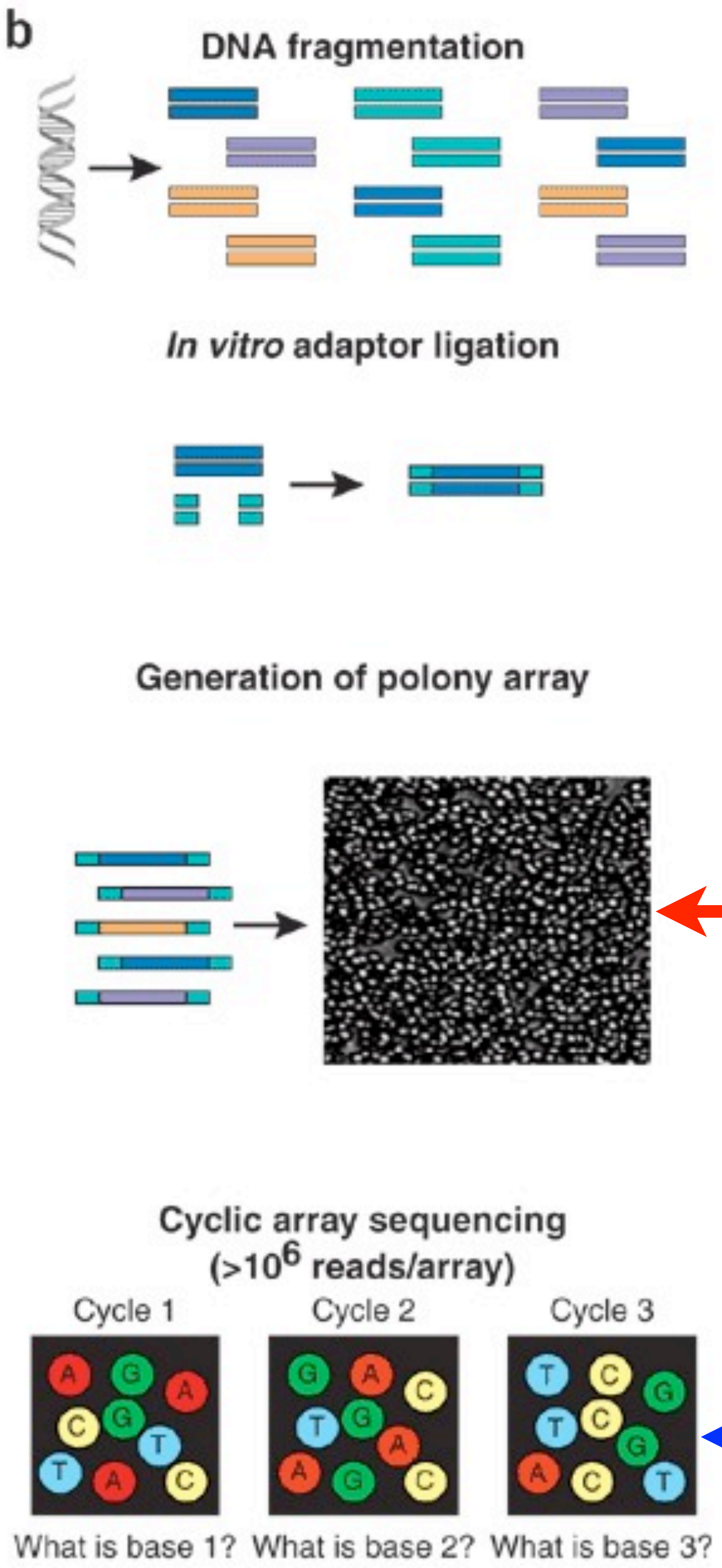
Synthesize complementary strand

Visualize nucleotide incorporations

Computer-assisted image analysis

Shendure & Ji, Nat. Biotech. (2008).

Sequencing technologies: High-Throughput



10^7-10^8 reads in parallel,
limited to ~100bp

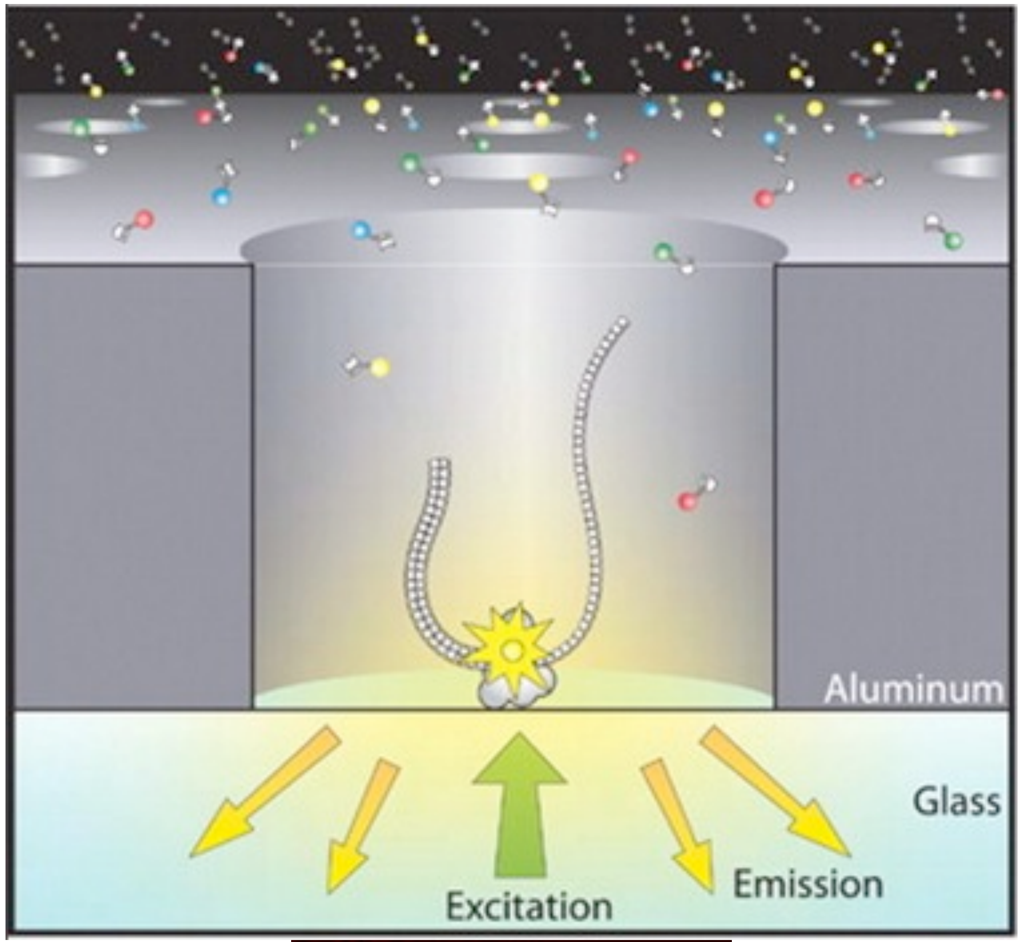
On-chip clonal colonies of original fragment

Optical reading

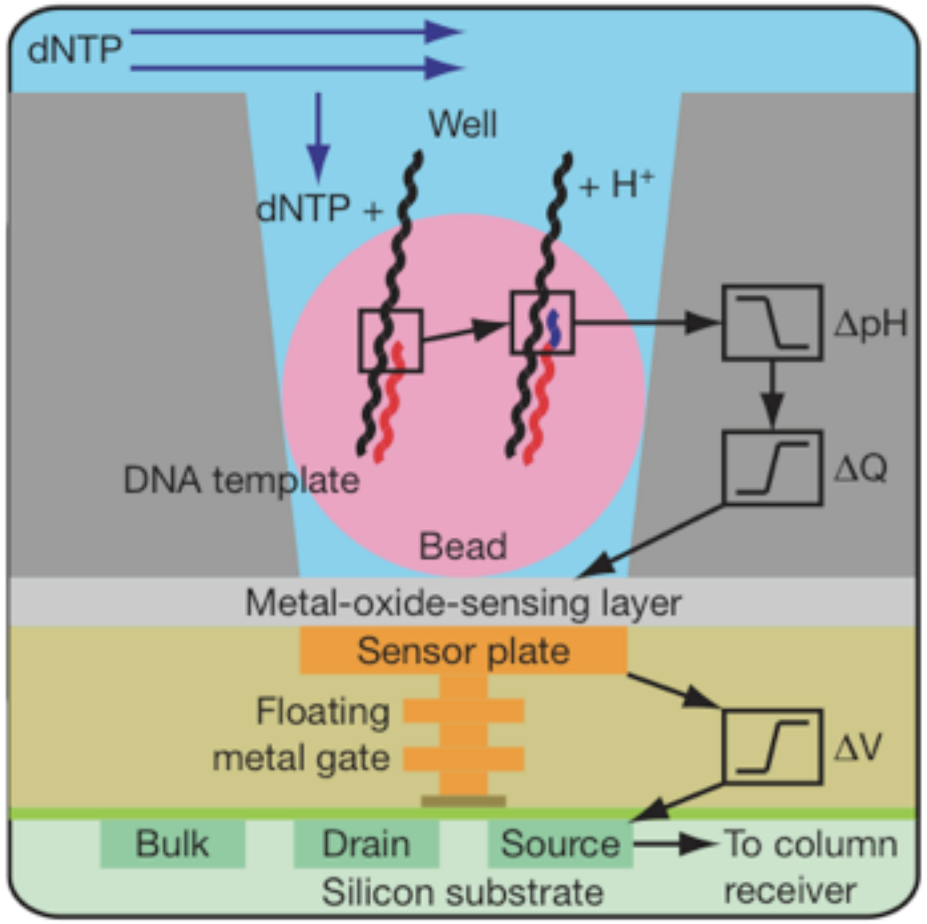
Sequencing technologies: High-Throughput

Current trends:

- increase read length (sequence single molecule)
- use non-optical detection (image analysis is error-prone and laborious)
- decrease cost (smaller volumes, higher throughput)



PACIFIC BIOSCIENCES™



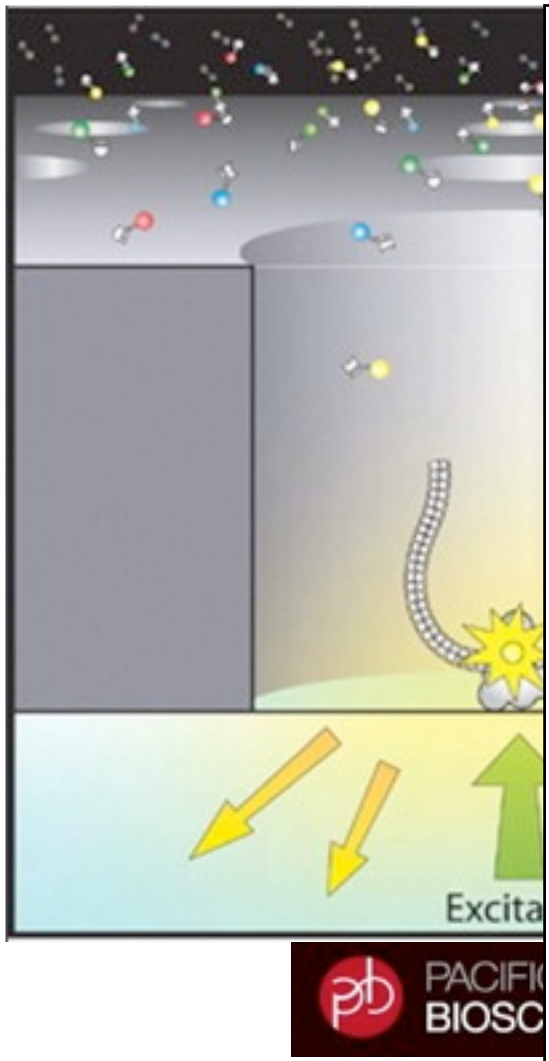
ion torrent

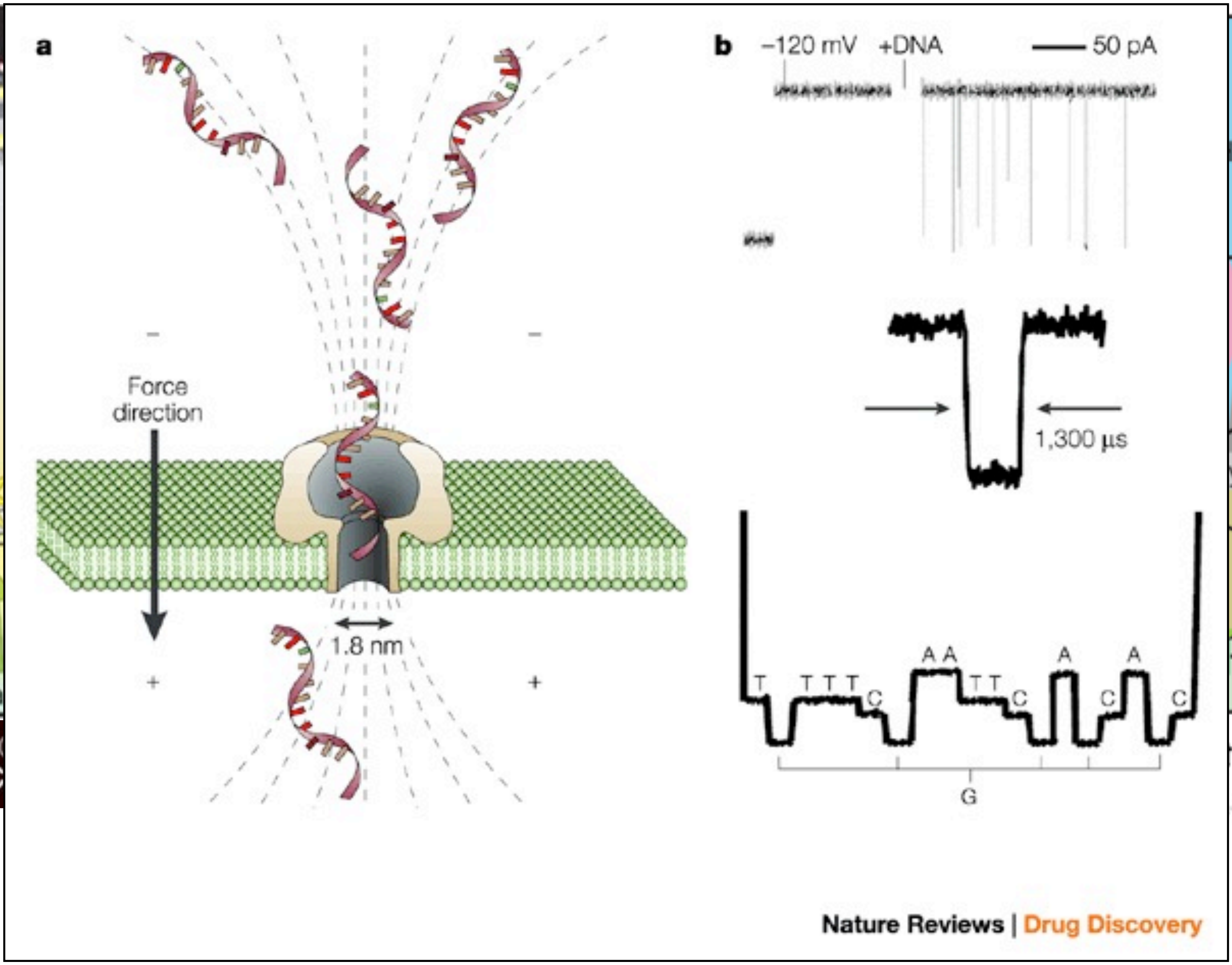
by Life technologies

Sequencing technologies: High-Throughput

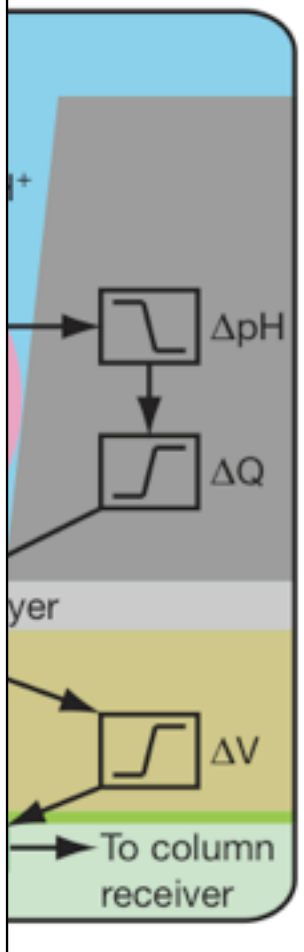
Current trends:

- increase read length (sequence single molecule)
- use non-optical detection (image analysis is error-prone and laborious)
- decrease cost (smaller volumes, higher throughput)

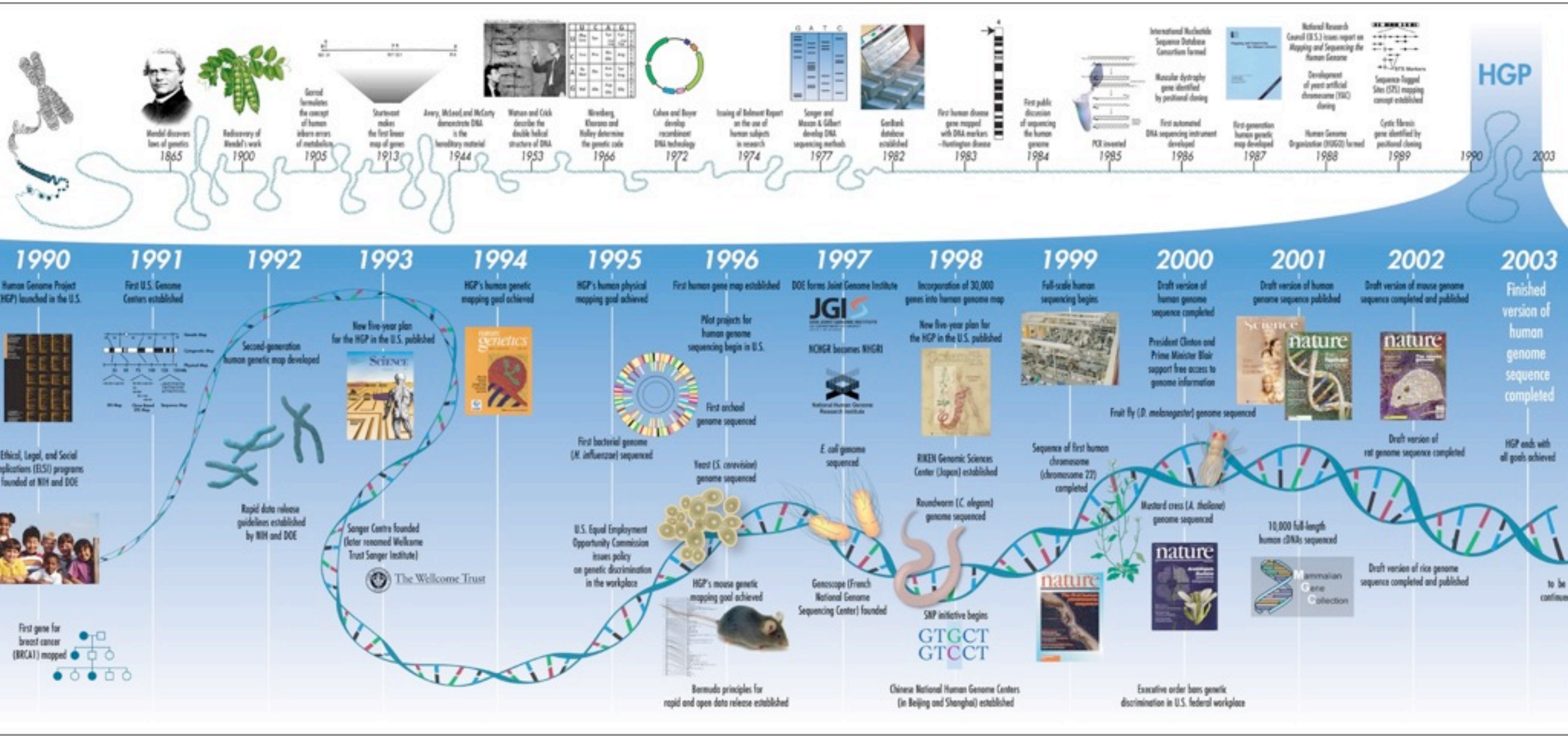




Nature Reviews | Drug Discovery



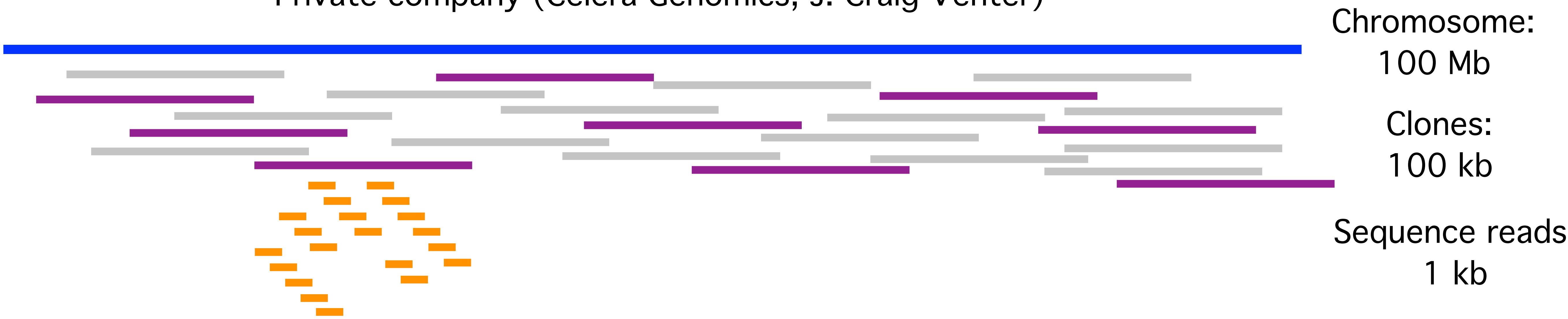
History of genome sequencing projects


Mendel discovers laws of genetics
1865
Rediscovery of Mendel's work
1900
Garrod formulates the concept of human inborn errors of metabolism
1905
Sturtevant makes the first linear map of genes
1913
Avery, McLeod, and McCarty demonstrate DNA is the hereditary material
1944
Watson and Crick describe the double helical structure of DNA
1953
Nirenberg, Khorana and Holley determine the genetic code
1966
Cohen and Boyer develop recombinant DNA technology
1972
Issuing of Belmont Report on the use of human subjects in research
1974
G A T C
Sanger and Maxam & Gilbert develop DNA sequencing methods
1977
GenBank database established
1982
First human disease gene mapped with DNA markers —Huntington disease
1983
First public discussion of sequencing the human genome
1984
PCR invented
1985
International Nucleotide Sequence Database Consortium formed
Muscular dystrophy gene identified by positional cloning
First automated DNA sequencing instrument developed
1986
First-generation human genetic map developed
1987
National Research Council (U.S.) issues report on Mapping and Sequencing the Human Genome
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed
1988
STS Markers
Sequence-Tagged Sites (STS) mapping concept established
Cystic fibrosis gene identified by positional cloning
1989
HGP
1990
2003
1990
Human Genome Project
HGP) launched in the U.S.
plications (ELSI) programs
Ethical, Legal, and Social
founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
Science
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997
DOE forms Joint Genome Institute
JGI
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
nature
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
nature
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continue

http://www.genome.gov/

Human Genome Project: 1990-2000

- Two competing initiatives with different strategies:
  - Public consortium
  - Private company (Celera Genomics, J. Craig Venter)

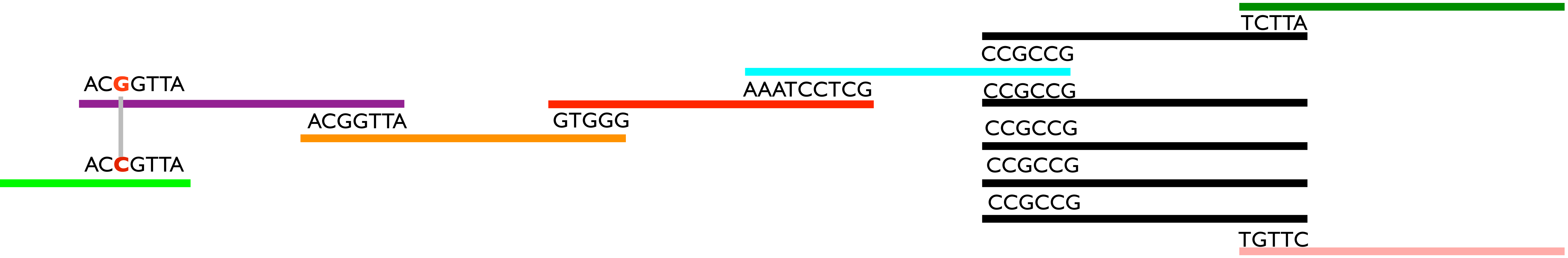


- Hierarchical sequencing:
  - Create library of ordered clones
  - Fragment and sequence them
  - Assemble fragments

- Whole genome shotgun sequencing:
  - Directly fragment the whole genome
  - Use paired-end sequencing to resolve repeats

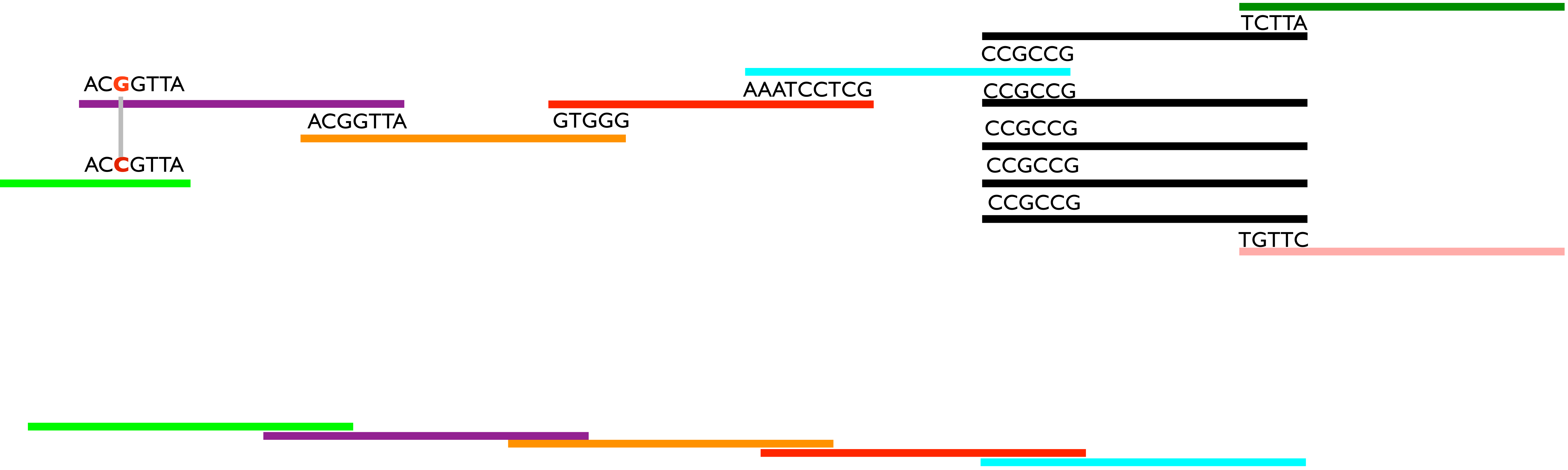
Fragments assembly

- General Procedure
  - Overlap → Layout → Consensus
- Difficulties:
  - Computing overlap with sequencing errors (1-3%) and unknown orientation



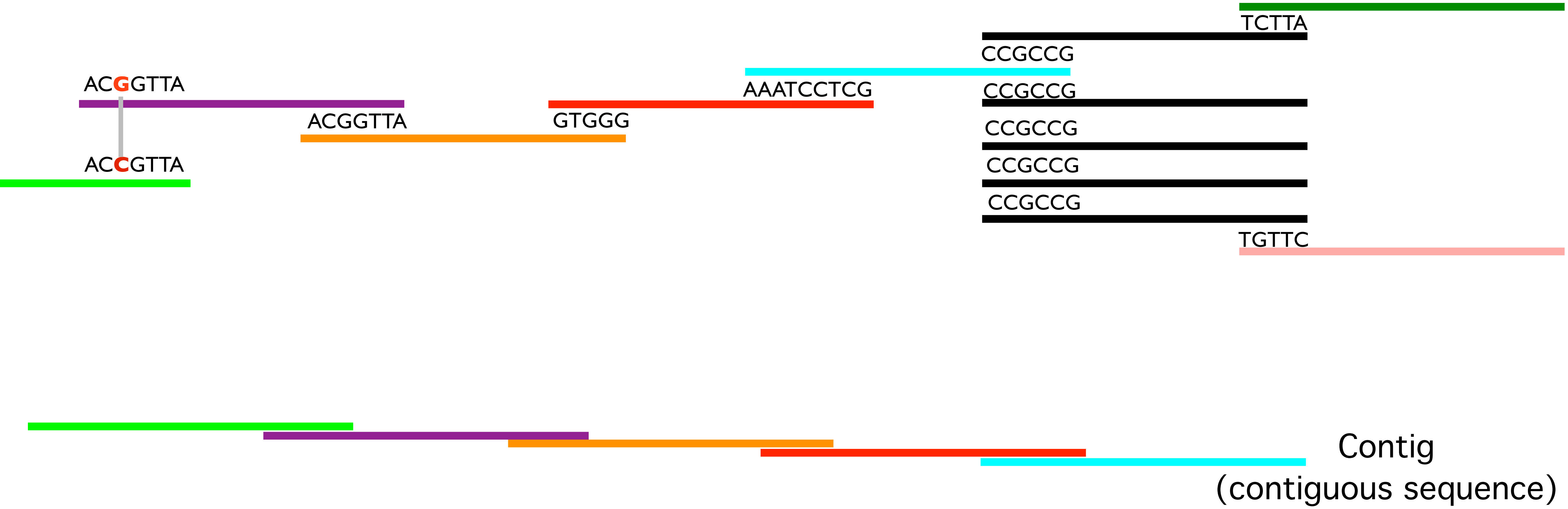
Fragments assembly

- General Procedure
  - Overlap → Layout → Consensus
- Difficulties:
  - Computing overlap with sequencing errors (1-3%) and unknown orientation



Fragments assembly

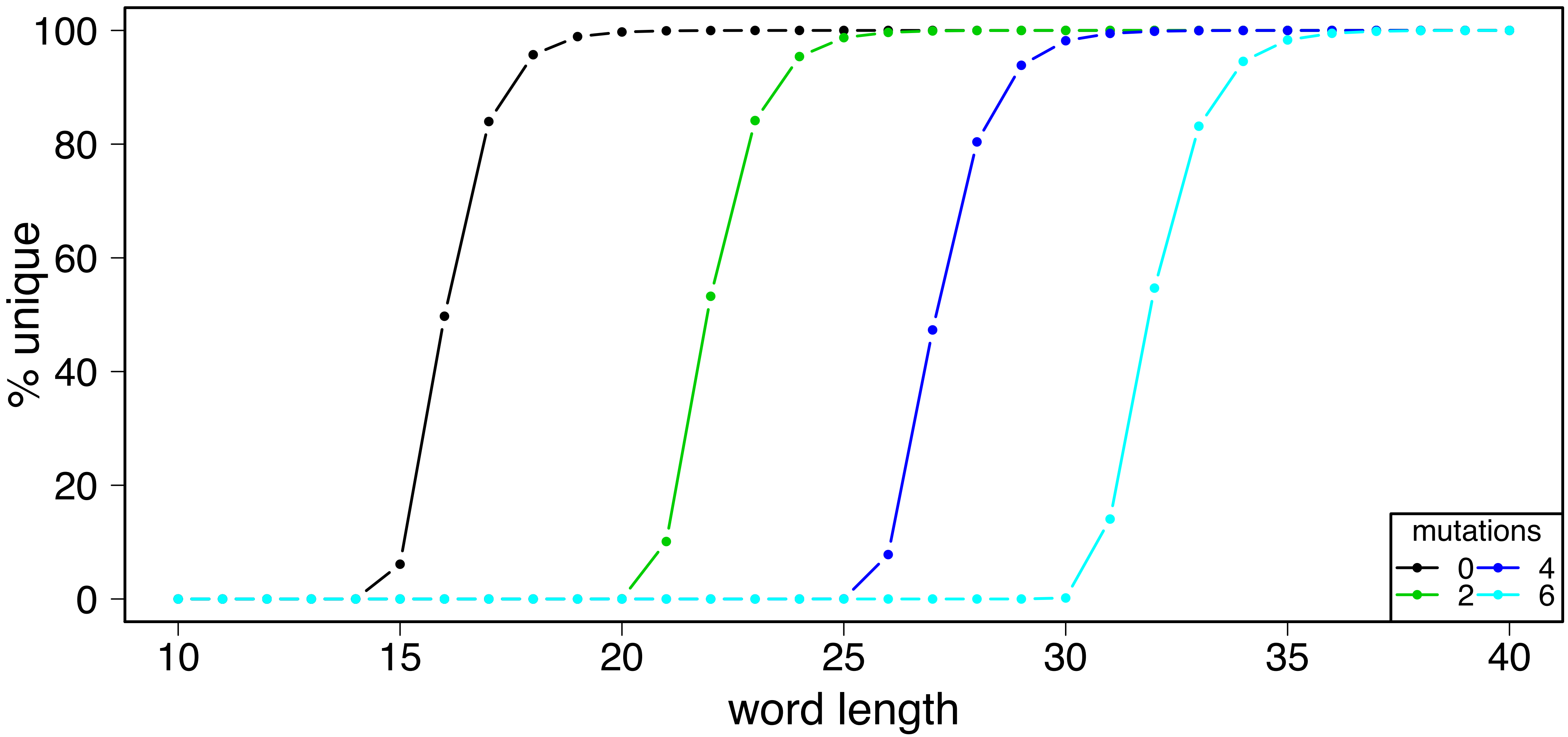
- General Procedure
  - Overlap → Layout → Consensus
- Difficulties:
  - Computing overlap with sequencing errors (1-3%) and unknown orientation



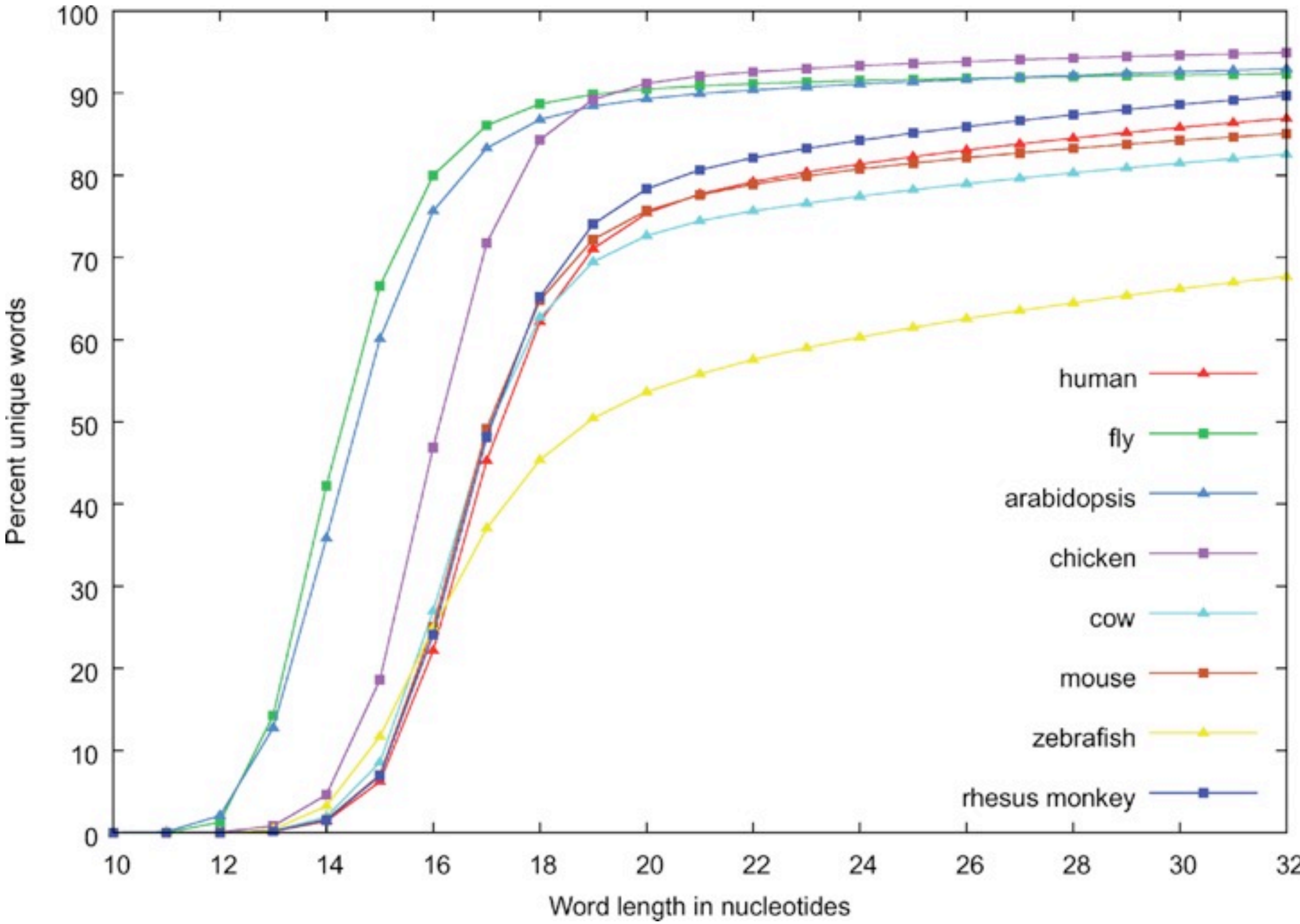
Contig
(contiguous sequence)

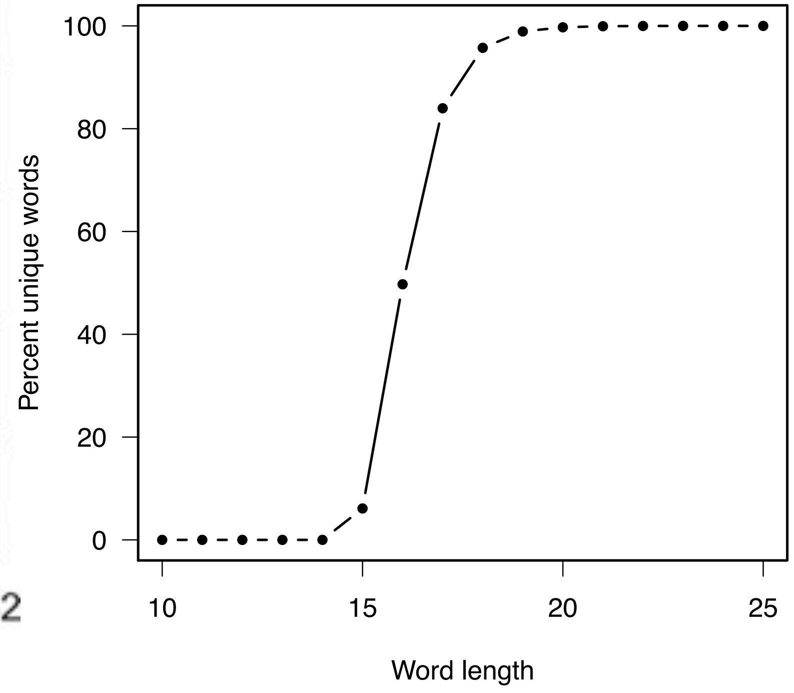
Overlap size

- Human genome is 3Gb, $\log_4(3\times10^9)=15$
- We are assuming up to 3% errors, so two words with a few differences can be considered the same
- If we use 35mers with up to 6 "mistakes", this is still "unique" in the genome
- PHRED score: $-10\times \log_{10}$ (Prob of wrong base call)



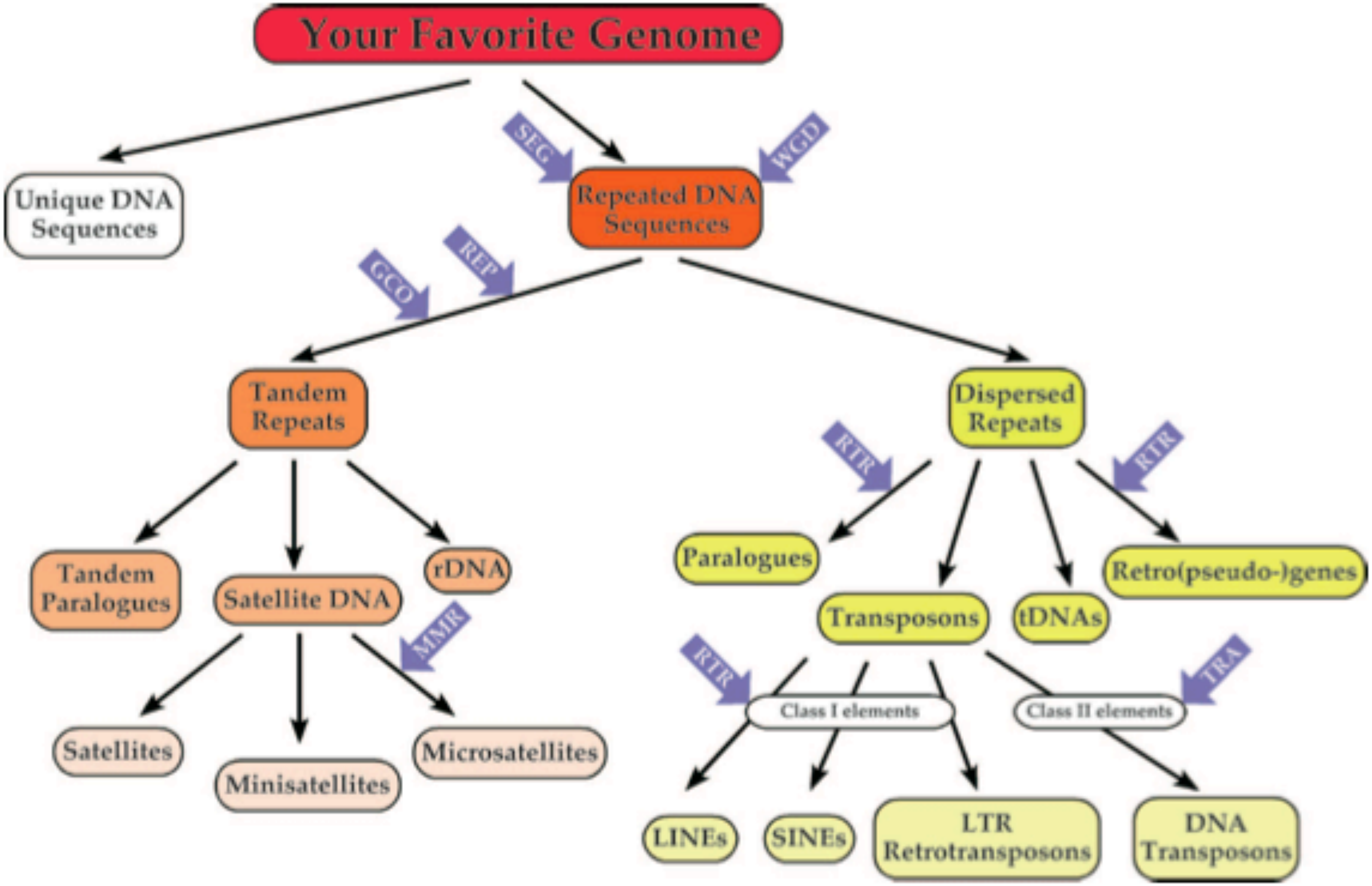
Digression: sequence repeats





Iseli et al. PLoS ONE (2007)

Digression: sequence repeats



Richard, G.-F., et al. MMBR (2008).

Digression: sequence repeats

	WGD	tDNA	LINEs/ SINEs	LTRs	DNA
Yeast	1	274	-	52 elem.	-
Drosophila	0	292	0.7%	1.5%	0.7%
Mouse	2	335	27%	10%	1%
Human	2	345	34%	8%	3%

- WGD: Whole Genome Duplications
- tDNA: genes encoding for tRNA
- LINE: 6-8 Kb, contains 2 ORFs
- SINE: 100-300 bp (Human Alu, Mouse B1/B2)
- LTR: up to 80% of plant genomes

Digression: sequence repeats

- RepBase: a database of consensus transposable elements
- RepeatMasker: a tool to identify sequences similar to these elements in other sequences (genomes)
- Common strategy in genome assembly is to mask repeats before computing read overlaps

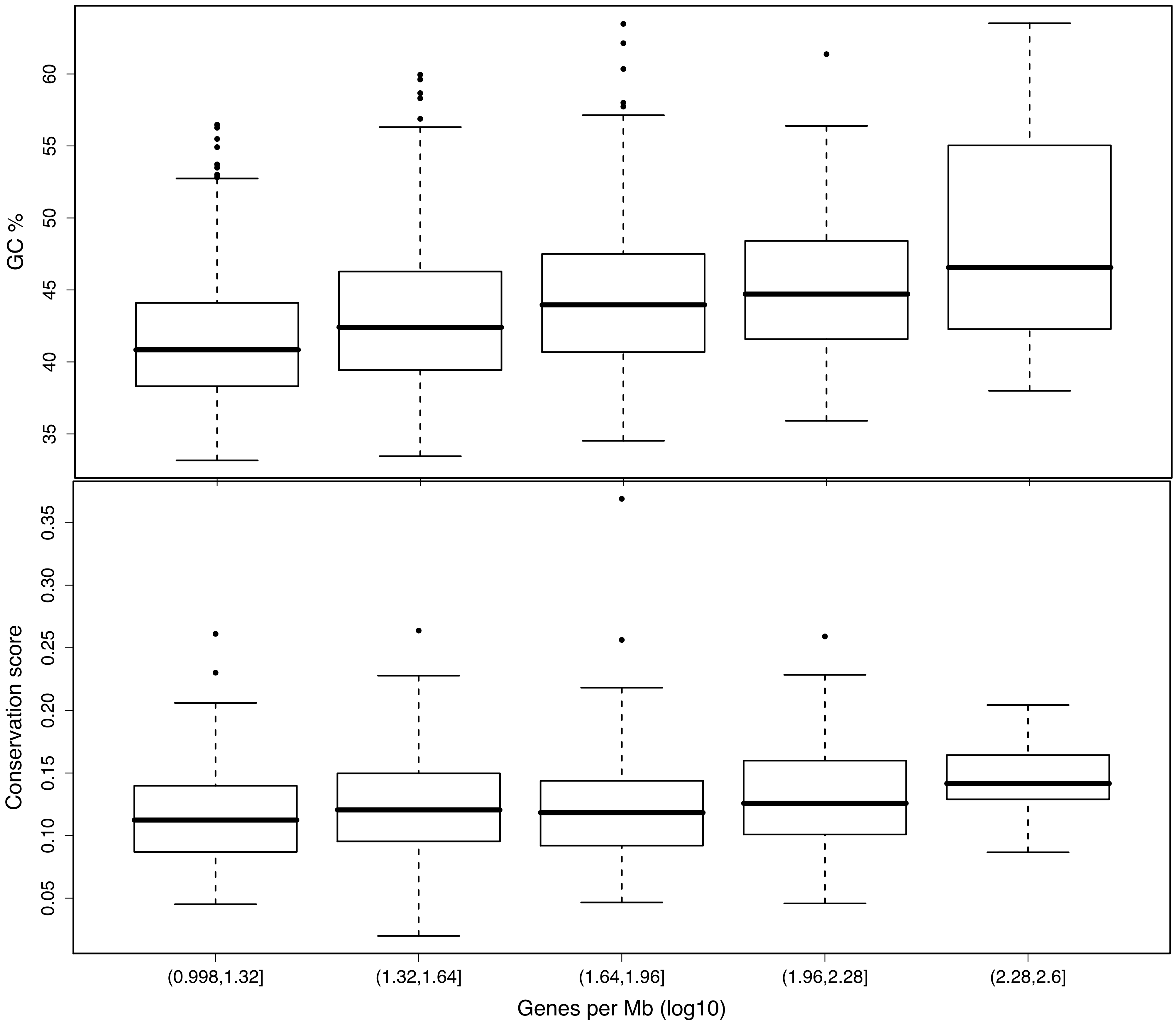
http://girinst.org/

Outcome: Human genome

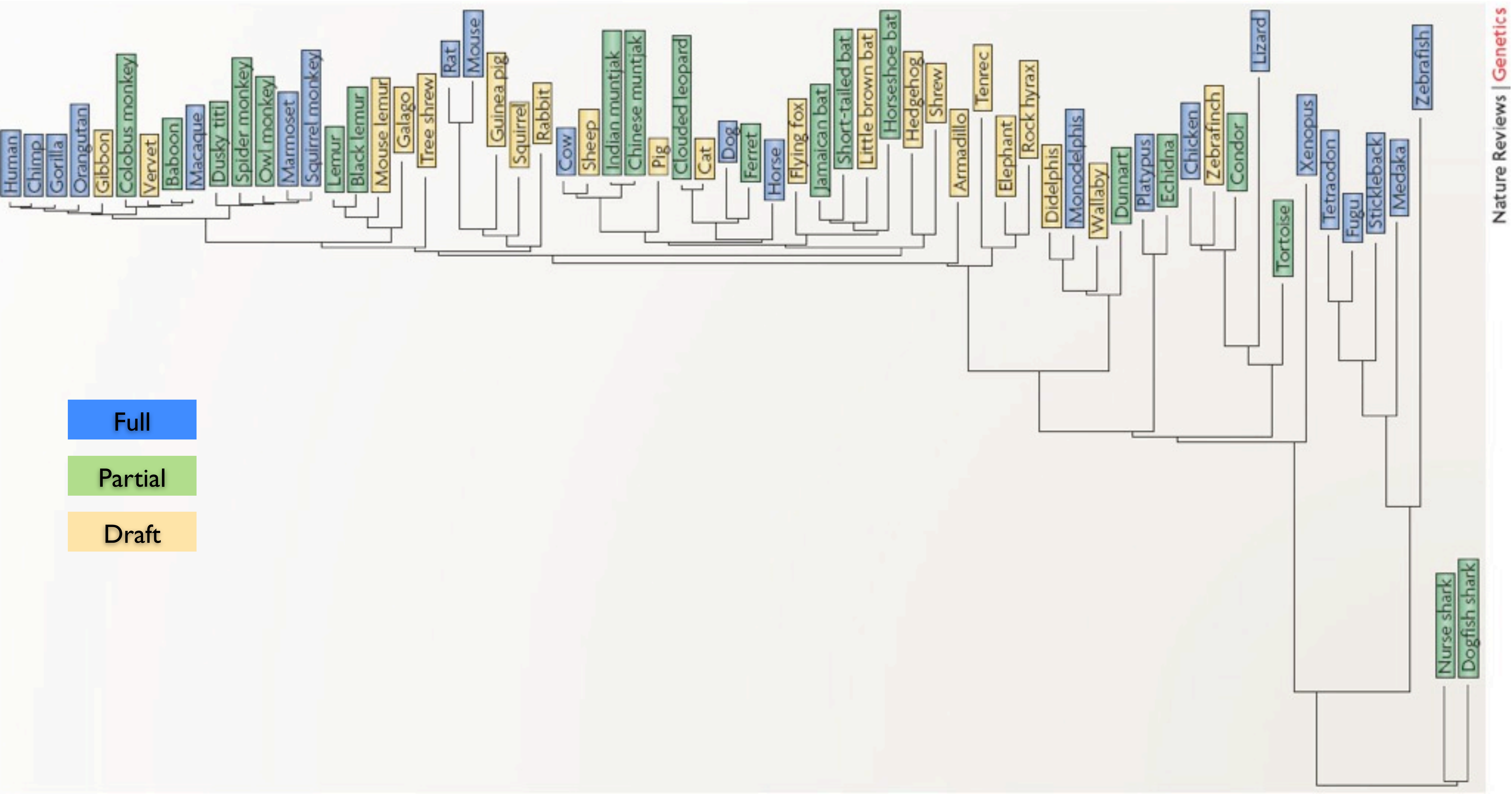
	Size (Mb)	GC content	Nb genes	N50
Yeast	12	38%	6,696	
Drosophila	169	42%	13,781	
Mouse	2,717	42%	21,879	39Mb
Human	3,102	40%	20,469	46Mb

- N50: size of smallest contig such that 50% genome is covered
- Mycobacterium Tuberculosis GC: 66%

Outcome: Human genome

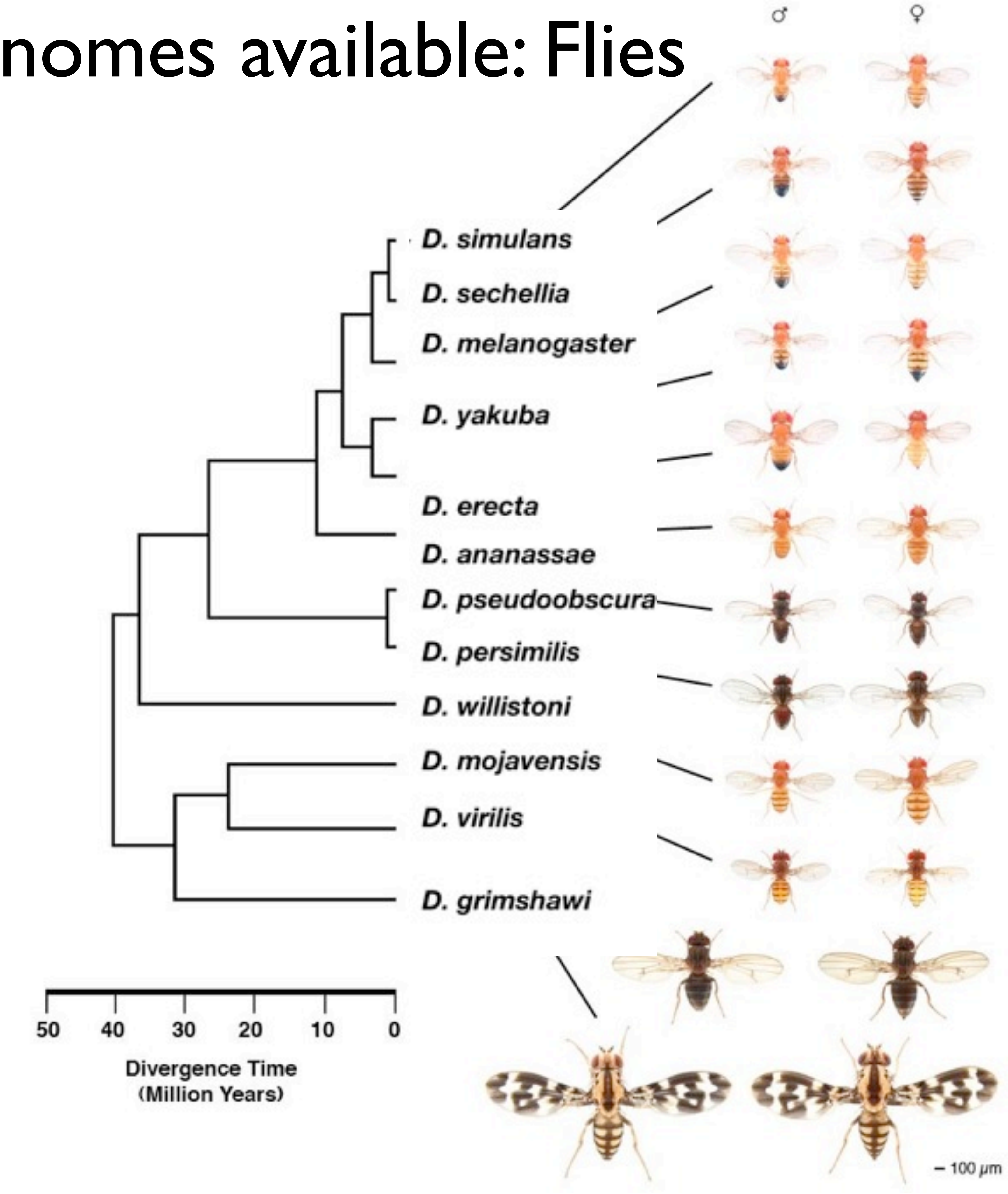


Other genomes available: Vertebrates



Margulies, E. H., & Birney, E. Nat. Rev. Genet. (2008).

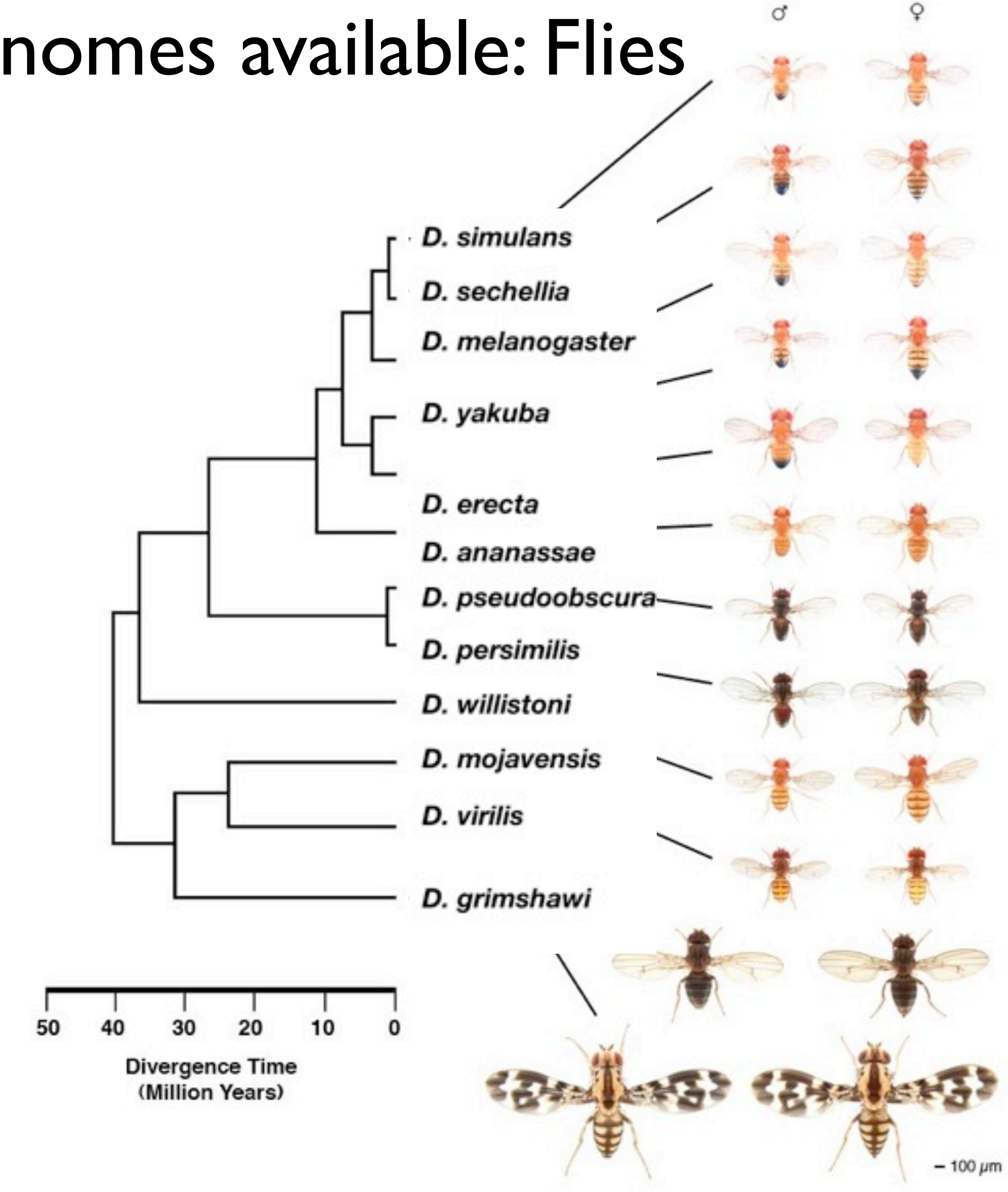
Other genomes available: Flies



http://flybase.org/

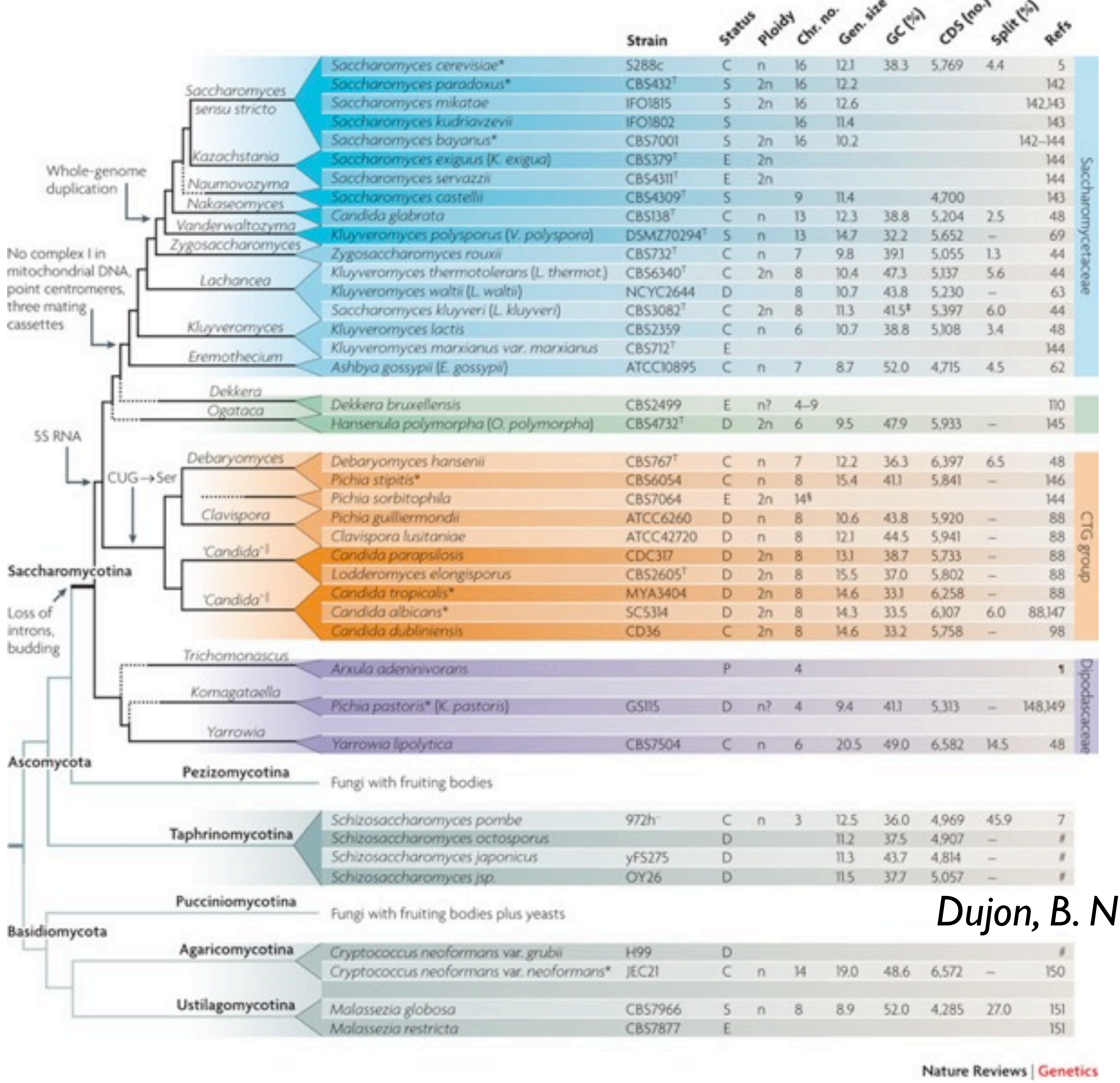
Other genomes available: Flies

12 Drosophila genomes



http://flybase.org/

Other genomes available: yeasts

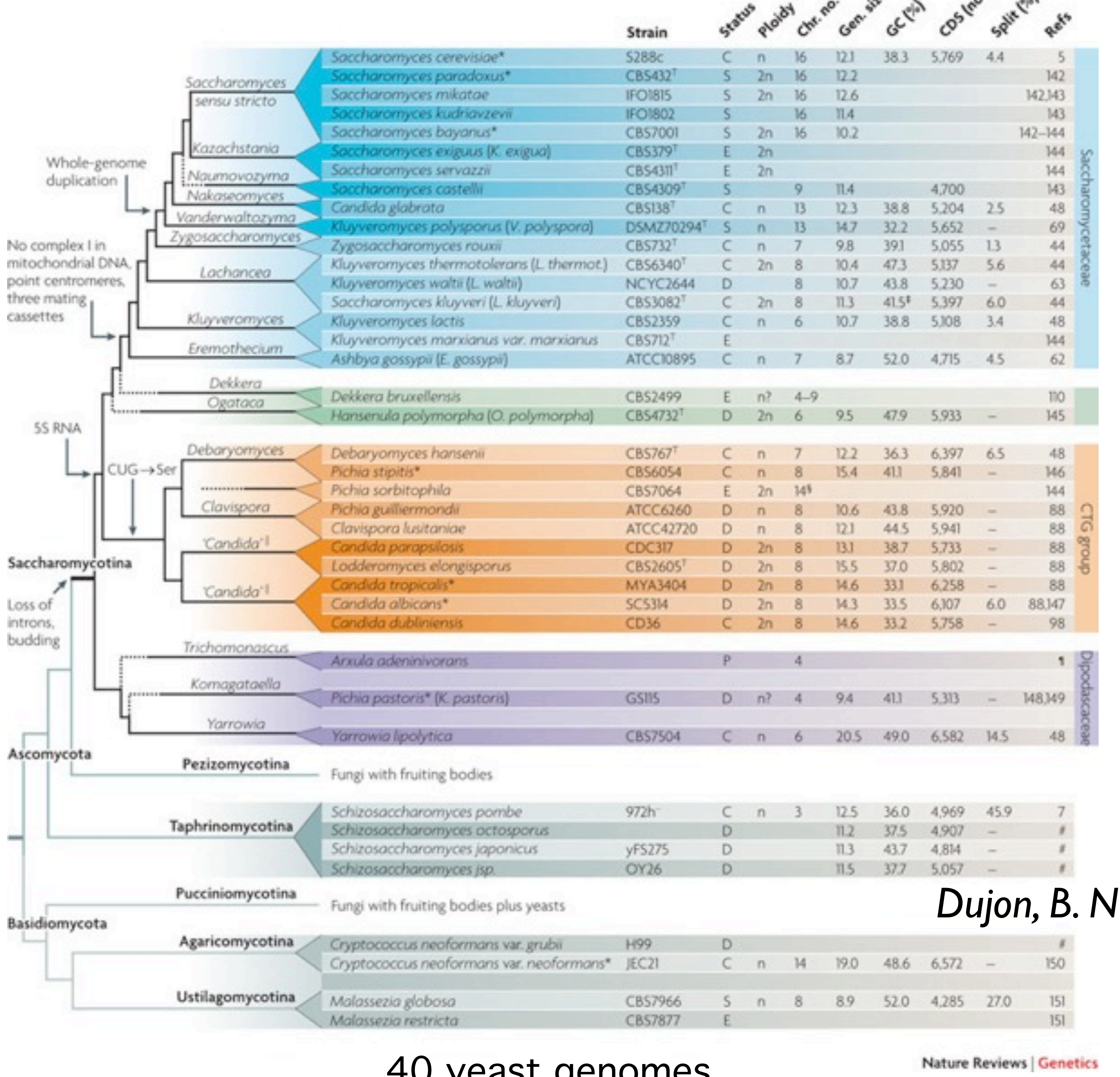


	Strain	Status	Ploidy	Chr. no.	Gen. size	GC (%)	CDS (no.)	Split (%)	Refs
*Saccharomyces cerevisiae**	S288c	C	n	16	12.1	38.3	5,769	4.4	5
*Saccharomyces paradoxus**	CBS432^T	S	2n	16	12.2				142
Saccharomyces mikatae	IFO1815	S	2n	16	12.6				142,143
Saccharomyces kudriavzevii	IFO1802	S		16	11.4				143
*Saccharomyces bayanus**	CBS7001	S	2n	16	10.2				142–144
Saccharomyces exiguus (K. exigua)	CBS379^T	E	2n						144
Saccharomyces servazzii	CBS4311^T	E	2n						144
Saccharomyces castellii	CBS4309^T	S		9	11.4		4,700		143
Candida glabrata	CBS138^T	C	n	13	12.3	38.8	5,204	2.5	48
Kluyveromyces polysporus (V. polyspora)	DSMZ70294^T	S	n	13	14.7	32.2	5,652	–	69
Zygosaccharomyces rouxii	CBS732^T	C	n	7	9.8	39.1	5,055	1.3	44
Kluyveromyces thermotolerans (L. thermot.)	CBS6340^T	C	2n	8	10.4	47.3	5,137	5.6	44
Kluyveromyces waltii (L. waltii)	NCYC2644	D		8	10.7	43.8	5,230	–	63
Saccharomyces kluyveri (L. kluyveri)	CBS3082^T	C	2n	8	11.3	41.5‡	5,397	6.0	44
Kluyveromyces lactis	CBS2359	C	n	6	10.7	38.8	5,108	3.4	48
Kluyveromyces marxianus var. marxianus	CBS712^T	E							144
Ashbya gossypii (E. gossypii)	ATCC10895	C	n	7	8.7	52.0	4,715	4.5	62
Dekkera bruxellensis	CBS2499	E	n?	4–9					110
Hansenula polymorpha (O. polymorpha)	CBS4732^T	D	2n	6	9.5	47.9	5,933	–	145
Debaryomyces hansenii	CBS767^T	C	n	7	12.2	36.3	6,397	6.5	48
*Pichia stipitis**	CBS6054	C	n	8	15.4	41.1	5,841	–	146
Pichia sorbitophila	CBS7064	E	2n	14§					144
Pichia guilliermondii	ATCC6260	D	n	8	10.6	43.8	5,920	–	88
Clavispora lusitaniae	ATCC42720	D	n	8	12.1	44.5	5,941	–	88
Candida parapsilosis	CDC317	D	2n	8	13.1	38.7	5,733	–	88
Lodderomyces elongisporus	CBS2605^T	D	2n	8	15.5	37.0	5,802	–	88
*Candida tropicalis**	MYA3404	D	2n	8	14.6	33.1	6,258	–	88
*Candida albicans**	SC5314	D	2n	8	14.3	33.5	6,107	6.0	88,147
Candida dubliniensis	CD36	C	2n	8	14.6	33.2	5,758	–	98
Arxula adeninivorans		P		4					¶
Pichia pastoris (K. pastoris)*	GS115	D	n?	4	9.4	41.1	5,313	–	148,149
Yarrowia lipolytica	CBS7504	C	n	6	20.5	49.0	6,582	14.5	48
Fungi with fruiting bodies									
Schizosaccharomyces pombe	972h$^-$	C	n	3	12.5	36.0	4,969	45.9	7
Schizosaccharomyces octosporus		D			11.2	37.5	4,907	–	#
Schizosaccharomyces japonicus	yFS275	D			11.3	43.7	4,814	–	#
Schizosaccharomyces jsp.	OY26	D			11.5	37.7	5,057	–	#
Fungi with fruiting bodies plus yeasts									
Cryptococcus neoformans var. grubii	H99	D							#
*Cryptococcus neoformans var. neoformans**	JEC21	C	n	14	19.0	48.6	6,572	–	150
Malassezia globosa	CBS7966	S	n	8	8.9	52.0	4,285	27.0	151
Malassezia restricta	CBS7877	E							151

Nature Reviews | Genetics

Dujon, B. Nat. Rev. Genet. (2010)

Other genomes available: yeasts



	Strain	Status	Ploidy	Chr. no.	Gen. size	GC (%)	CDS (no.)	Split (%)	Refs
*Saccharomyces cerevisiae**	S288c	C	n	16	12.1	38.3	5,769	4.4	5
*Saccharomyces paradoxus**	CBS432^{T}	S	2n	16	12.2				142
Saccharomyces mikatae	IFO1815	S	2n	16	12.6				142,143
Saccharomyces kudriavzevii	IFO1802	S		16	11.4				143
*Saccharomyces bayanus**	CBS7001	S	2n	16	10.2				142–144
Saccharomyces exiguus (K. exigua)	CBS379^{T}	E	2n						144
Saccharomyces servazzii	CBS4311^{T}	E	2n						144
Saccharomyces castellii	CBS4309^{T}	S		9	11.4		4,700		143
Candida glabrata	CBS138^{T}	C	n	13	12.3	38.8	5,204	2.5	48
Kluyveromyces polysporus (V. polyspora)	DSMZ70294^{T}	S	n	13	14.7	32.2	5,652	–	69
Zygosaccharomyces rouxii	CBS732^{T}	C	n	7	9.8	39.1	5,055	1.3	44
Kluyveromyces thermotolerans (L. thermot.)	CBS6340^{T}	C	2n	8	10.4	47.3	5,137	5.6	44
Kluyveromyces waltii (L. waltii)	NCYC2644	D		8	10.7	43.8	5,230	–	63
Saccharomyces kluyveri (L. kluyveri)	CBS3082^{T}	C	2n	8	11.3	41.5‡	5,397	6.0	44
Kluyveromyces lactis	CBS2359	C	n	6	10.7	38.8	5,108	3.4	48
Kluyveromyces marxianus var. marxianus	CBS712^{T}	E							144
Ashbya gossypii (E. gossypii)	ATCC10895	C	n	7	8.7	52.0	4,715	4.5	62
Dekkera bruxellensis	CBS2499	E	n?	4–9					110
Hansenula polymorpha (O. polymorpha)	CBS4732^{T}	D	2n	6	9.5	47.9	5,933	–	145
Debaryomyces hansenii	CBS767^{T}	C	n	7	12.2	36.3	6,397	6.5	48
*Pichia stipitis**	CBS6054	C	n	8	15.4	41.1	5,841	–	146
Pichia sorbitophila	CBS7064	E	2n	14§					144
Pichia guilliermondii	ATCC6260	D	n	8	10.6	43.8	5,920	–	88
Clavispora lusitaniae	ATCC42720	D	n	8	12.1	44.5	5,941	–	88
Candida parapsilosis	CDC317	D	2n	8	13.1	38.7	5,733	–	88
Lodderomyces elongisporus	CBS2605^{T}	D	2n	8	15.5	37.0	5,802	–	88
*Candida tropicalis**	MYA3404	D	2n	8	14.6	33.1	6,258	–	88
*Candida albicans**	SC5314	D	2n	8	14.3	33.5	6,107	6.0	88,147
Candida dubliniensis	CD36	C	2n	8	14.6	33.2	5,758	–	98
Arxula adeninivorans		P		4					¶
Pichia pastoris (K. pastoris)*	GS115	D	n?	4	9.4	41.1	5,313	–	148,149
Yarrowia lipolytica	CBS7504	C	n	6	20.5	49.0	6,582	14.5	48
Fungi with fruiting bodies									
Schizosaccharomyces pombe	972h^{-}	C	n	3	12.5	36.0	4,969	45.9	7
Schizosaccharomyces octosporus		D			11.2	37.5	4,907	–	#
Schizosaccharomyces japonicus	yFS275	D			11.3	43.7	4,814	–	#
Schizosaccharomyces jsp.	OY26	D			11.5	37.7	5,057	–	#
Fungi with fruiting bodies plus yeasts									
Cryptococcus neoformans var. grubii	H99	D							#
*Cryptococcus neoformans var. neoformans**	JEC21	C	n	14	19.0	48.6	6,572	–	150
Malassezia globosa	CBS7966	S	n	8	8.9	52.0	4,285	27.0	151
Malassezia restricta	CBS7877	E							151

Nature Reviews | Genetics

Dujon, B. Nat. Rev. Genet. (2010)

40 yeast genomes,
1744 bacterial genomes,
2695 virus genomes.