



BIO-463

Genomics and bioinformatics

Lecture 1: Introduction, DNA sequencing

Professors: Jacques Rougemont, Anne-Florence Bitbol, Raphaëlle Luisier

EPFL

Objective of the course

**Overview of tools and methods that define
bioinformatics in the context of the
development of genomics**

Genomics technology



Biological data

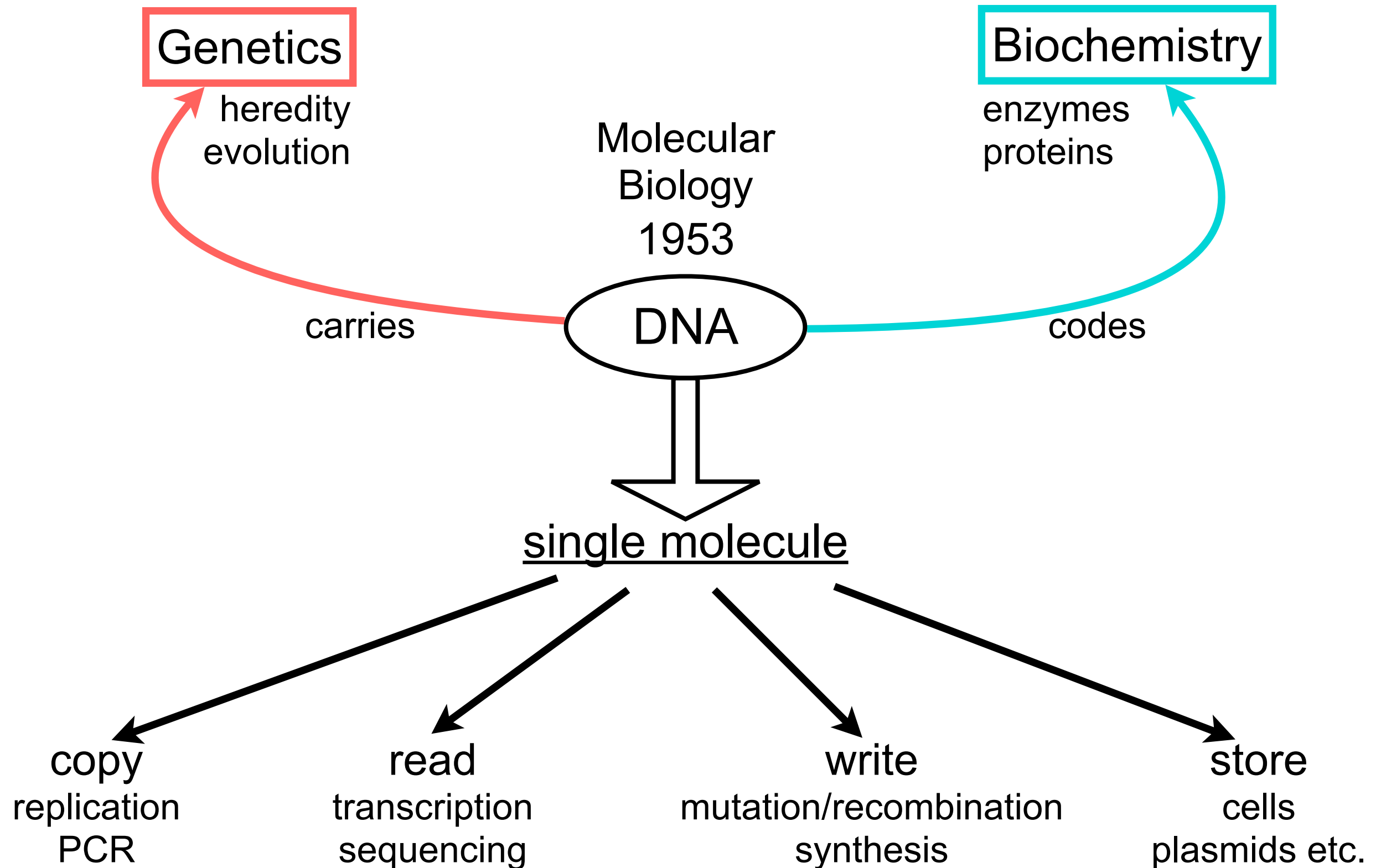


Interpretation problem



Algorithmic solution

Historical perspective



“biological information science”

Definitions

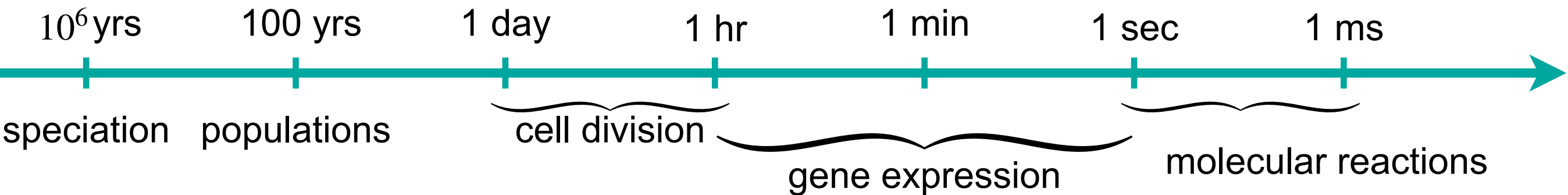
Bioinformatics: The science and technology of biological information (in particular its processing by the cell)

- theoretical science
- uses mathematics, statistics and algorithms to interpret large-scale biological data

(Structural) Genomics: The study of genomes as static objects: their nucleotide sequence, the location and order of features such as genes, bands, polymorphisms, regulatory elements, etc.

Functional Genomics: The extension of genomics to quantitative and dynamical phenomena, in particular transcription and translation.
It includes mechanisms regulating gene expression, protein-DNA and protein-protein interactions, etc.

Time scales



more static

more dynamic

structural
genomics

functional
genomics

sequencing programs

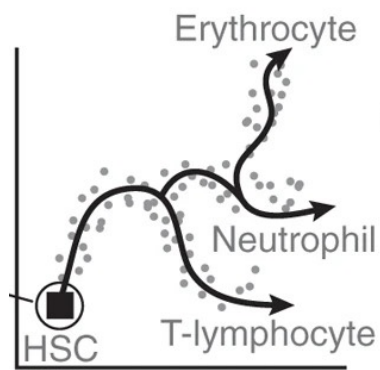
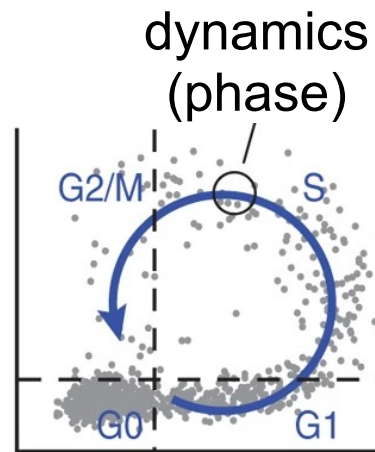
history of science

systems biology

course

Cell state / cell identity

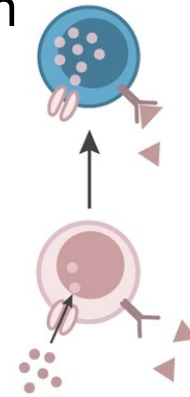
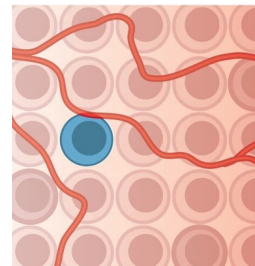
cell state



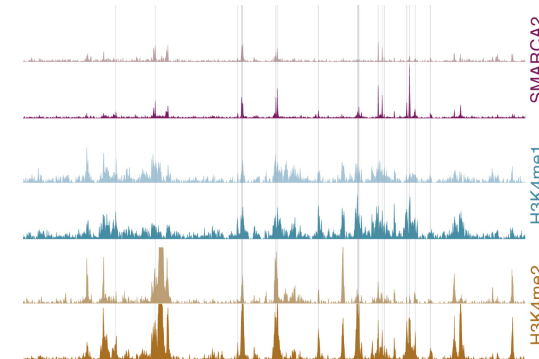
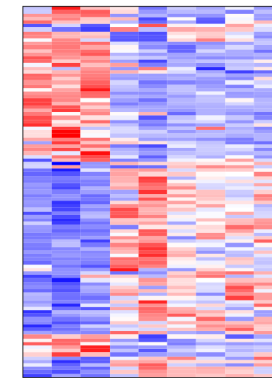
developmental trajectory
(history)

cell identity

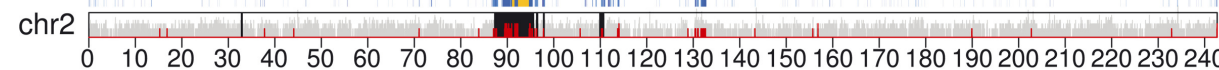
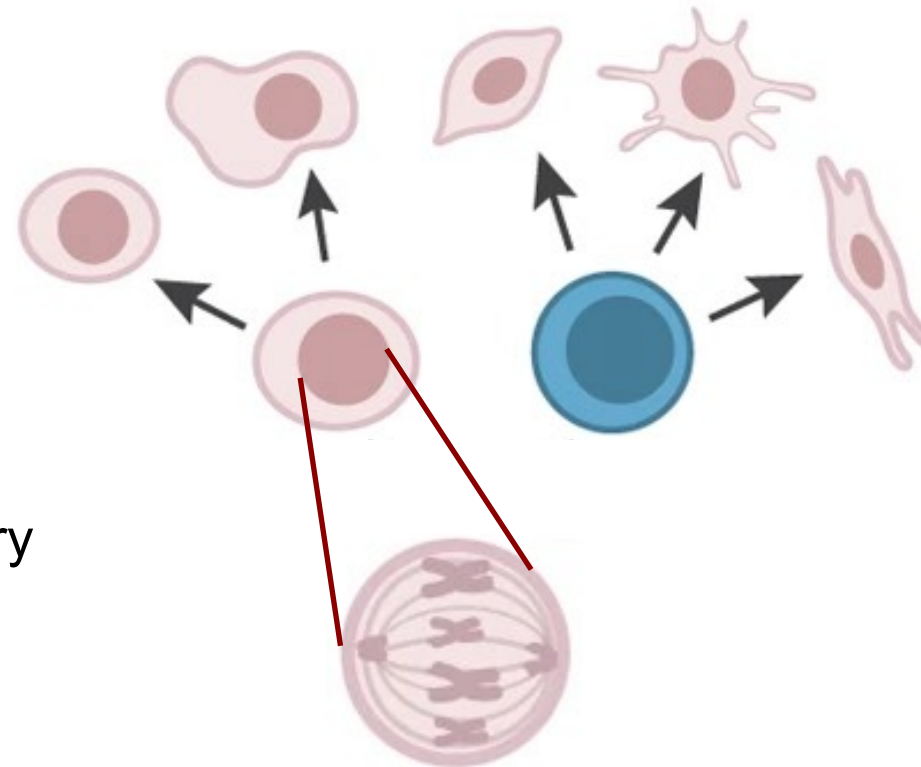
microenvironment,
communication



gene expression portrait



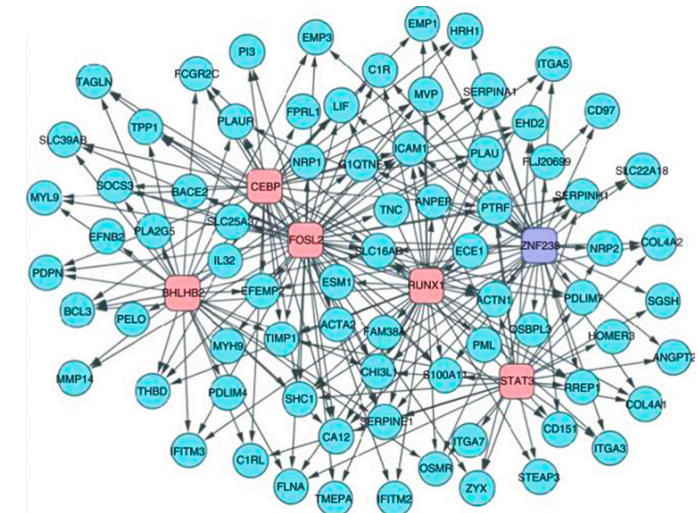
chromatin state



genome (species)

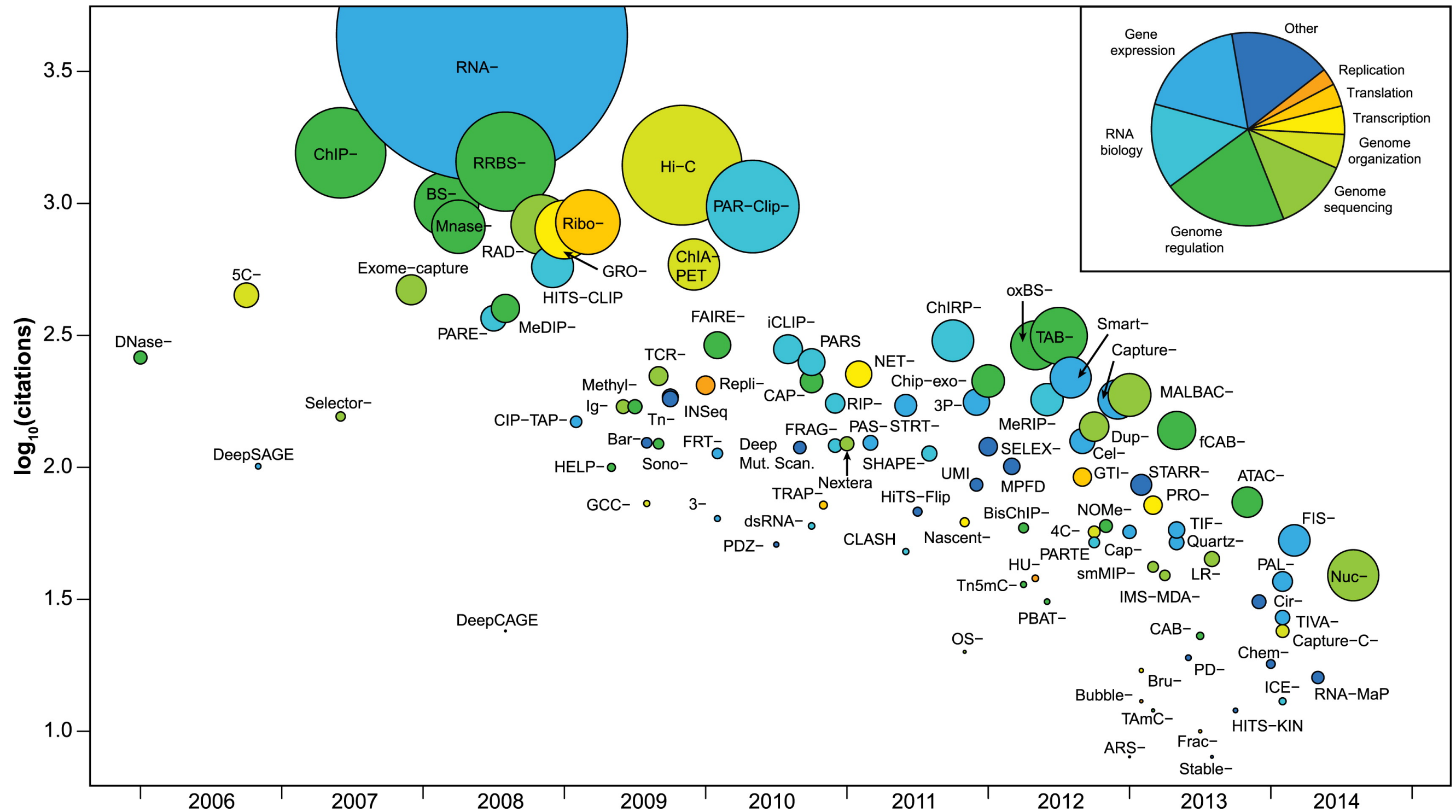
genotype (heredity)

somatic mutations (e.g. oncogenes)



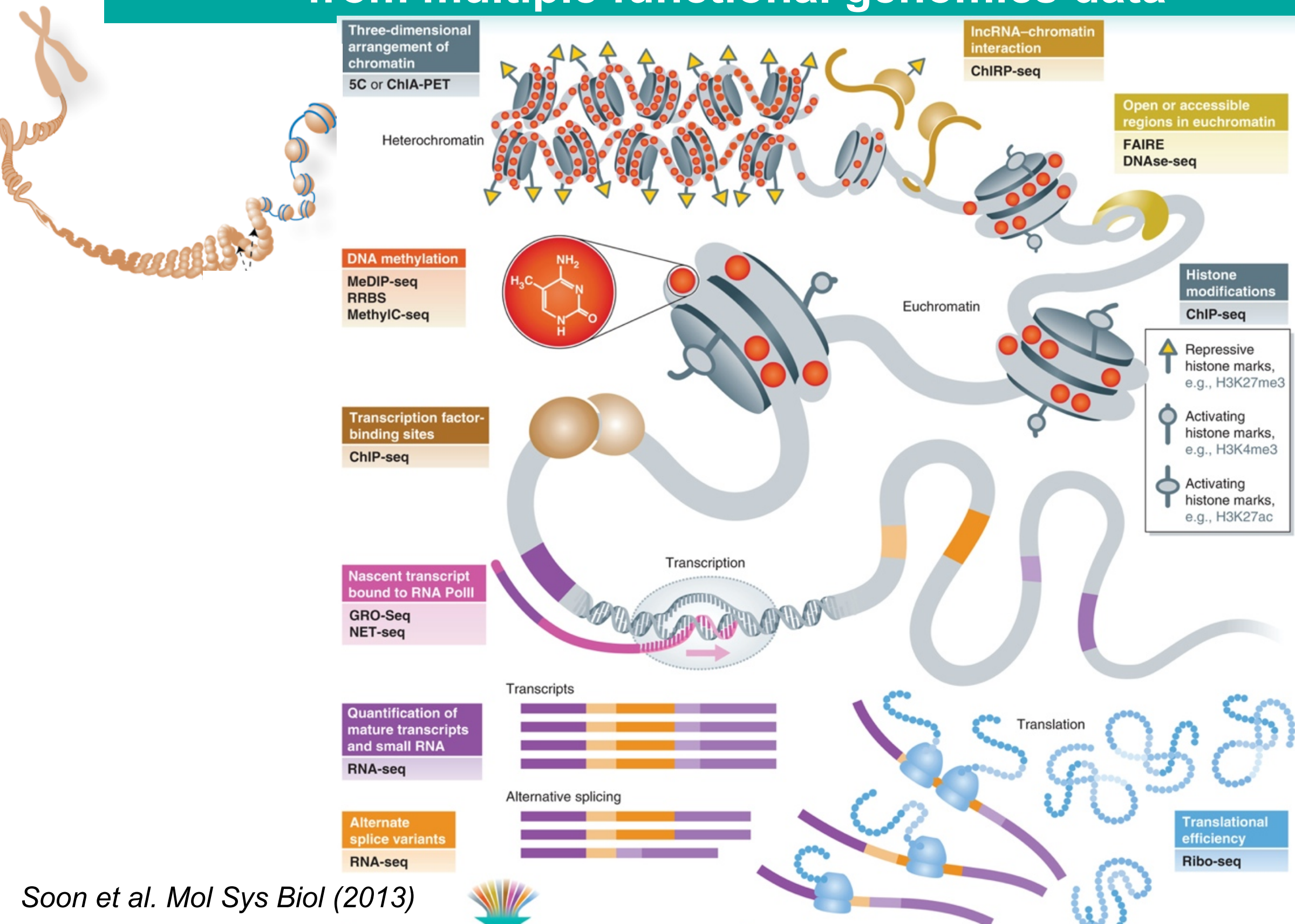
regulatory network,
pathways

Sequencing DNA for....



Sequencing DNA for....

A dynamic picture of the cell state from multiple functional genomics data



The UCSC genome browser and database

genome-euro.ucsc.edu

← → ↻ 🏠 genome.ucsc.edu/cgi-bin/hgGateway?hgsid=512469365_oRLXGI9hiQSJf8TqsJyl5mAeyEW7 ☆ 📶 🍪 ⚙️ 🔍 🌐 🚫 🛑



Genome Browser Gateway

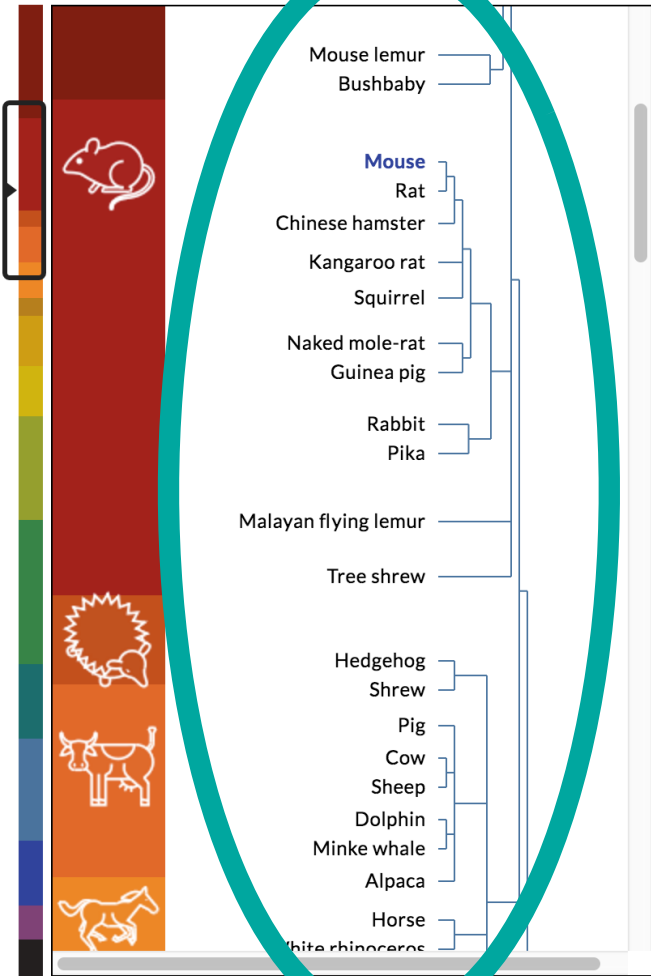
🏠 Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Browse/Select Species

POPULAR SPECIES

 Human  Mouse  Rat  Fruitfly  Worm  Yeast

REPRESENTED SPECIES



Find Position

Mouse Assembly
Dec. 2011 (GRCm38/mm10) ▾

Position/Search Term
 

Current position: chr12:56,694,976-56,714,605 🔗

position →

Mouse Genome Browser - mm10 assembly

[view sequences](#)

The Dec. 2011 *Mus musculus* assembly (Genome Reference Consortium Mouse Build 38 (GCA_000001635.2)) was produced by the [Mouse Genome Reference Consortium](#). For more information about this assembly, see [GRCm38](#) in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic region, an mRNA or EST, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the Mouse genome. See the [User's Guide](#) for more information.

Request:	Genome Browser Response:
chr16	Displays all of chromosome 16
chr16:1-5000000	Displays first 5 million bases of chr 16
chr16:1000000+2000	Displays a region of chr 16 that spans 2000 bases, starting with position 1000000
GL456161.2	Displays region of contig GL456161.2
BC012683	Displays alignment location of mRNA with GenBank accession BC012683
AW045217	Displays alignment location of EST with GenBank accession AW045217
Ncam2	Displays region of genome with official MGI mouse genetic nomenclature Ncam2
pseudogene mRNA	Lists transcribed pseudogenes but not cDNAs
zinc finger	Lists many zinc finger mRNAs
kruppel zinc finger	Lists only kruppel-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
Smith,D.	Lists mRNAs deposited by co-author D. Smith

Use this last format for author queries. Although GenBank requires the search format *Smith D*, internally it uses the format *Smith,D*.

Assembly details

Genome Reference Consortium GRCm38, which includes approximately 2.6 Gb of sequence, is considered to be "essentially complete". The assembly includes chromosomes 1-19, X, Y, M (mitochondrial DNA) and chr*_random (unlocalized) and chrUn_* (unplaced clone contigs). For information about the process used to assemble this version, see the [GRC website](#). For a glossary of assembly-related terms, please see the GRC [Assembly Terminology](#) page. Note that the UCSC mm10 database contains only the reference strain C57BL/6J.

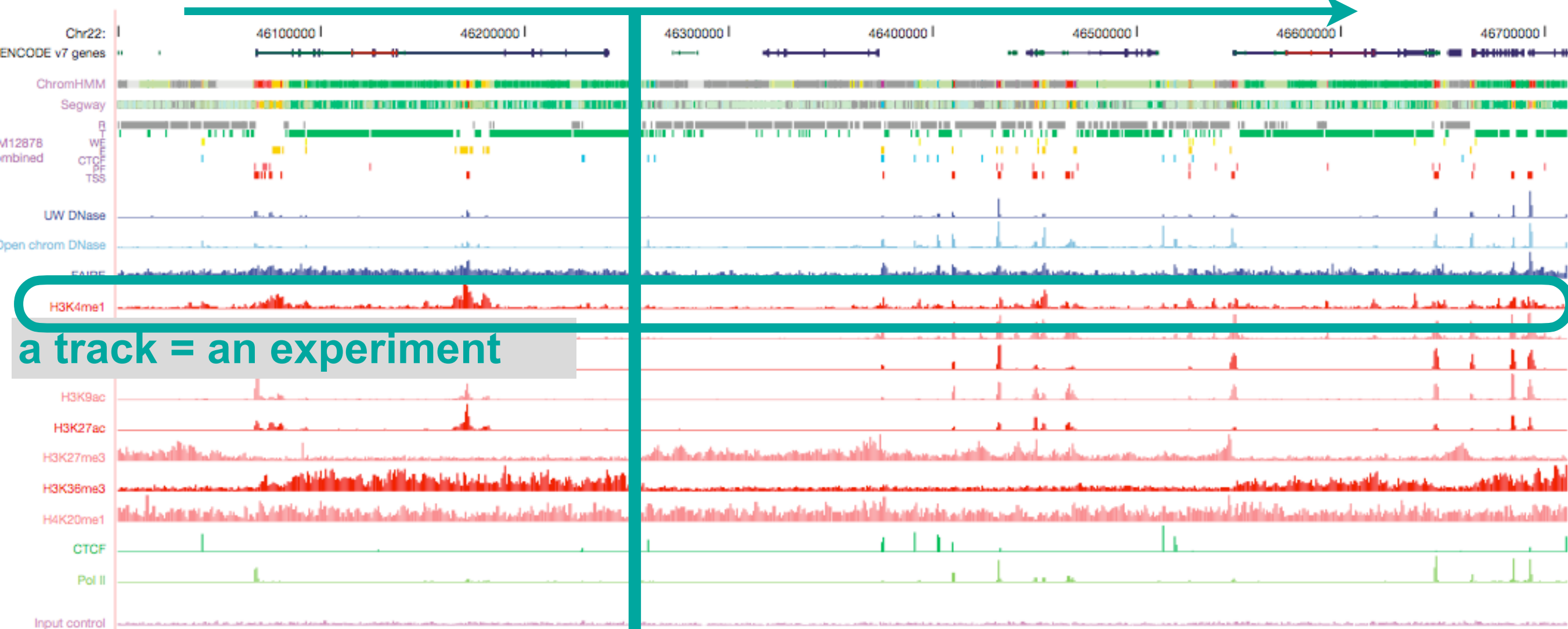


Mus musculus
(Photo courtesy of [The Jackson Laboratory](#))

Genome collection

Chromatin landscape

Genomic coordinates

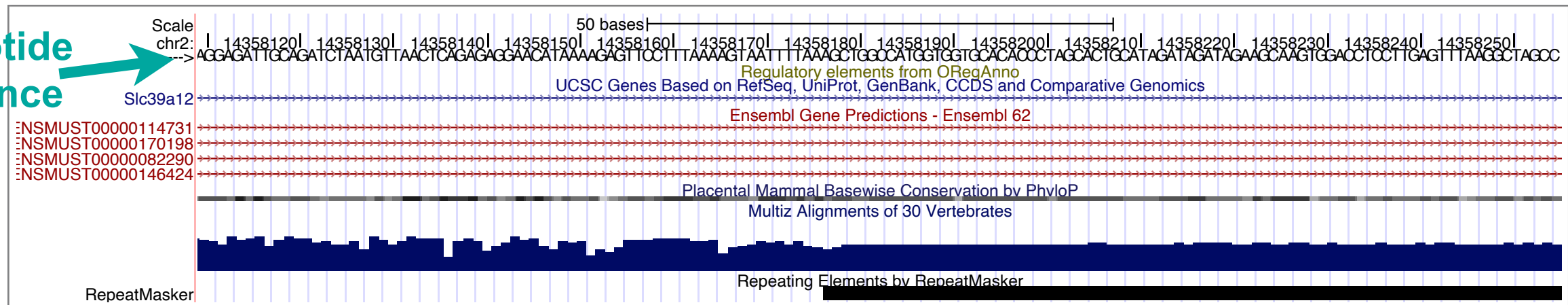


data at a genomic position

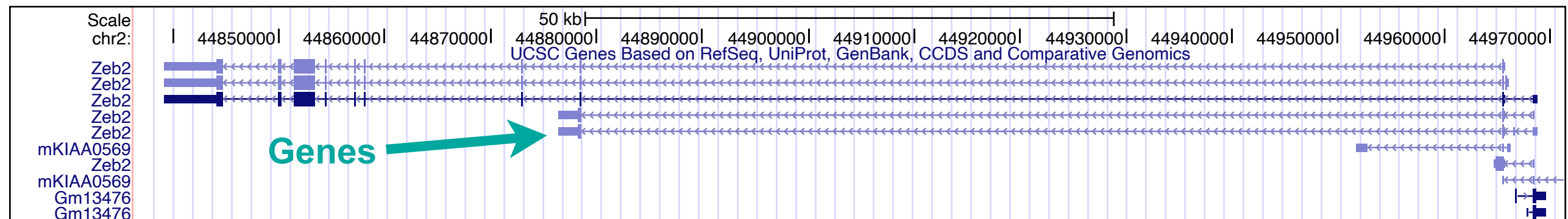
Zooming out

chr2 (qA2) 2aA1 2qA3 2qB 2aC1.1 1aC1.3C2 2aC3 2qD 2aE1 E2 2qE3 2aE5 2qF1 2qF3aG1 aG3 2qH1 H2 2qH3 2qH4

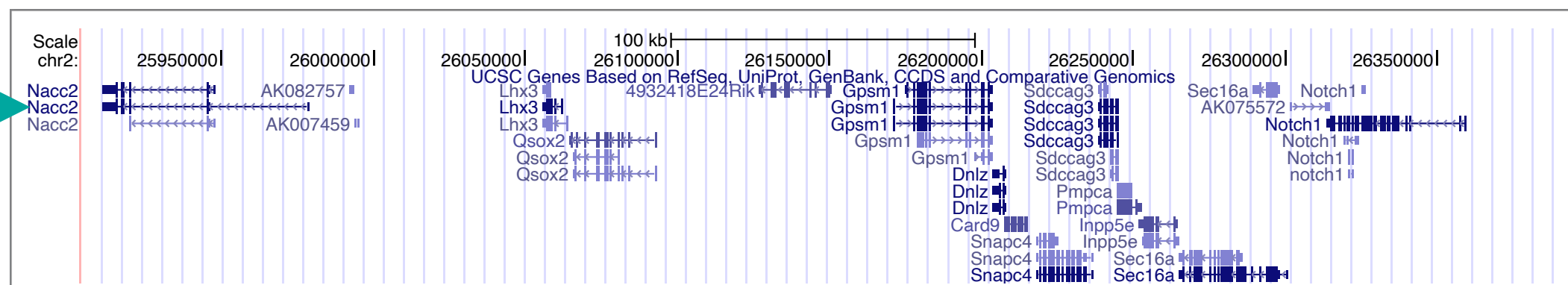
Nucleotide
sequence



chr2 (qA3) 2aA1 2qA3 2qB 2aC1.1 1aC1.3C2 2aC3 2qD 2aE1 E2 2qE3 2aE5 2qF1 2qF3aG1 aG3 2qH1 H2 2qH3 2qH4



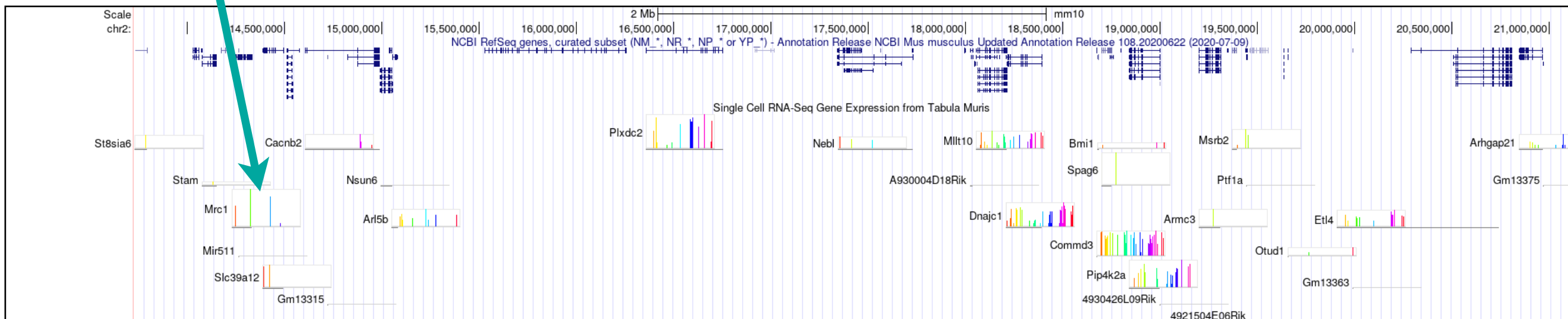
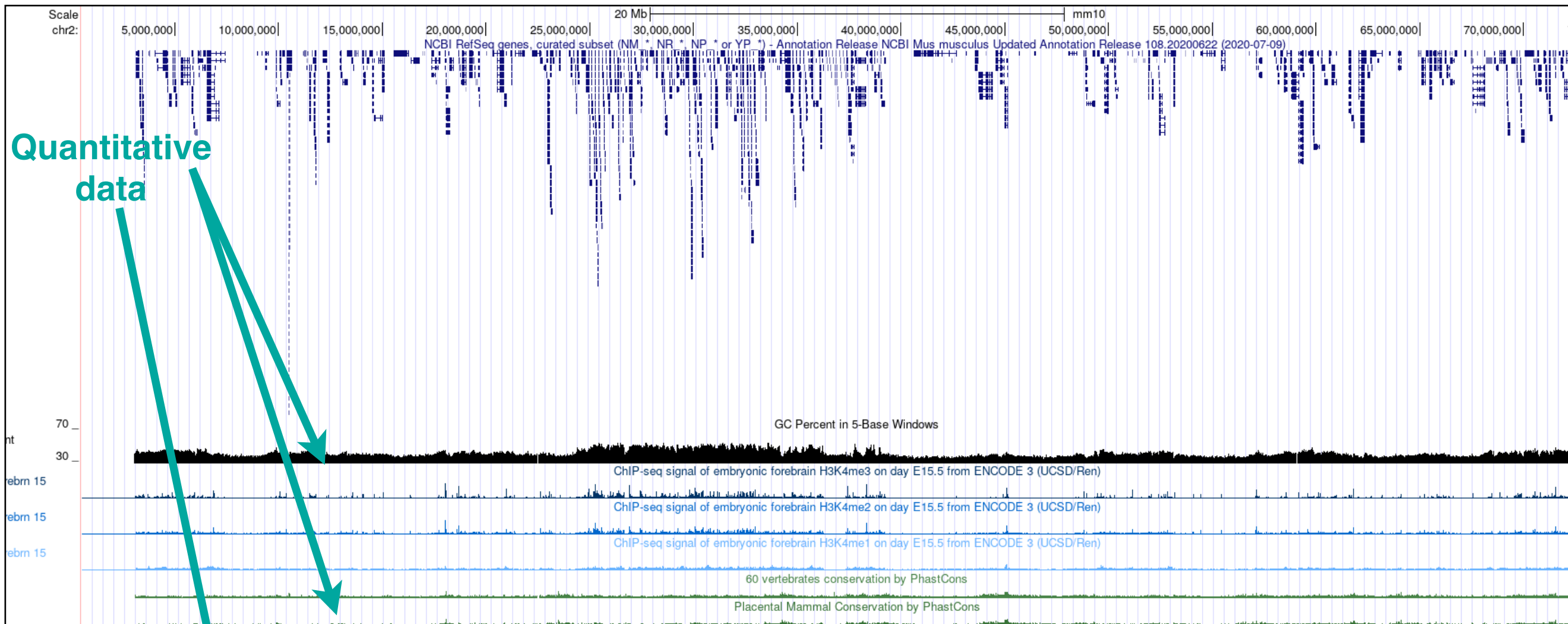
Alternative
transcripts
("isoforms")



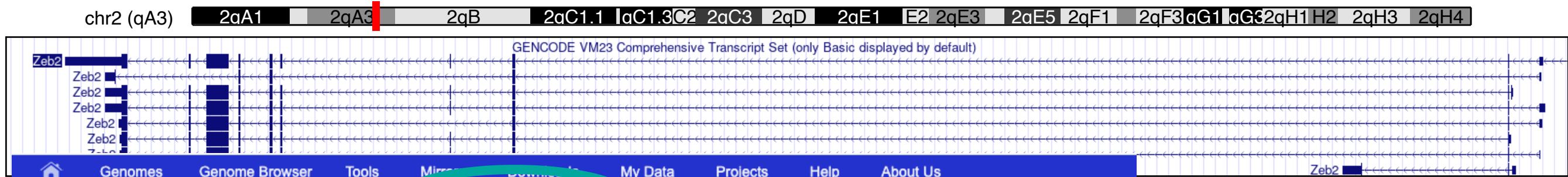
Combining data

chr2 (qA1-qC3)

2aA1 2aA3 2aB 2aC1.1 C1.3 C2 2aC3 2aD 2aE1 E2 2aE3 2aE5 2aF1 2aF3aG1 aG3 2aH1H2 2aH3 2aH4



Linked annotations



Mouse Gene Zeb2 (ENSMUST00000176438.1) from GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)

Description: Transcriptional inhibitor that binds to DNA sequence 5'-CACCT-3' in different promoters. Represses transcription of E-cadherin. (from UniProt)

Gencode Transcript: ENSMUST00000176438.8

Gencode Gene: ENSMUSG00000026872.18

Transcript (Including UTRs)

Position: mm10 chr2:44,983,632-45,117,395 **Size:** 133,764 **Total Exon Count:** 11 **Strand:** -

Coding Region

Position: mm10 chr2:44,988,573-45,110,088 **Size:** 121,516 **Coding Exon Count:** 9

Page Index	Sequence and Links	UniProtKB Comments	CTD	RNA Structure	Protein Structure
Other Species	GO Annotations	mRNA Descriptions	Other Names	Model Information	Methods

Data last updated at UCSC: 2019-09-20

Sequence and Links to Tools and Databases

Genomic Sequence (chr2:44,983,632-45,117,395)			mRNA (may differ from genome)		Protein (1215 aa)
Gene Sorter	Genome Browser	Other Species FASTA	VisiGene	Table Schema	AlphaFold
BioGPS	CGAP	Ensembl	ExonPrimer	GeneCards	MGI
PubMed	Stanford SOURCE	UniProtKB	Wikipedia		

Comments and Description Text from UniProtKB

ID: [ZEB2_MOUSE](#)

DESCRIPTION: RecName: Full=Zinc finger E-box-binding homeobox 2; AltName: Full=Smad-interacting protein 1; AltName:

FUNCTION: Transcriptional inhibitor that binds to DNA sequence 5'- CACCT-3' in different promoters. Represses transcr

SUBUNIT: Interacts with CBX4 and CTCF.
SUBCELLULAR LOCATION: Nucleus.

PTM: Sumoylation on Lys-391 and Lys-866 is promoted by the E3 SUMO-protein ligase CBX4, and impairs interaction with

SIMILARITY: Belongs to the delta-EF1/ZFH-1 C2H2-type zinc-finger family.

SIMILARITY: Contains 7 C2H2-type zinc fingers.

SIMILARITY: Contains 1 homeobox DNA-binding domain.

- **Comparative Toxicogenomics Database (CTD)**

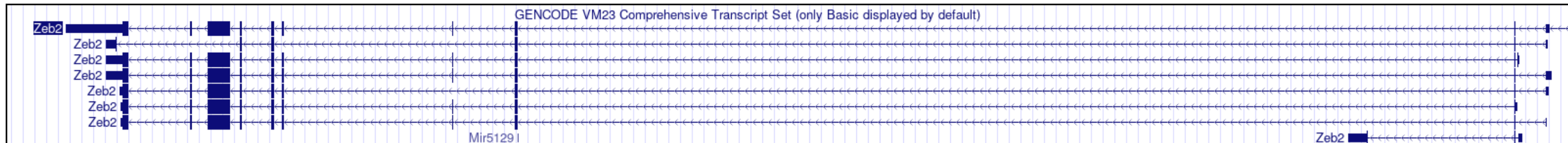
The following chemicals interact with this gene

- [D001151](#) Arsenic
- [D002737](#) Chloroprene
- [D015073](#) 2-Acetylaminofluorene
- [C009505](#) 4,4'-diaminodiphenylmethane
- [D015122](#) 6-Mercaptopurine
- [D000082](#) Acetaminophen

SENSMUST00000176438.8 (Zeb2) length=1215
MKQPI MADGPRCKRRKQANPRRKNVVDNVVDAGSETDEEDKLHIAEDDSL
ANP SPASMPNHESSPHMSQGLLPREEEEEELRESVVEHSWHSGEILQASVAGPEEMKE
GPEATIQT TINNGTVKNANCTSDFEFYFAKRKLEERDGHAVSIEEYLQ RSDTAI
EELSRLGTPEANGQEENDLPPGTPDAFAQLLTCPYCDRGYKRLTSLKEHIKYRHE
FSCPLCSYTFAYRTQLERHVMVTHKPGTDQHQM LTOGAGNRKFKCTECGKAFKYKH
LRIHSGEKPYECPNCKKRFSHSGSYSSHIS SKKCI GLISVNGRMRNNIKTGSSPN
PTNSAITQLRNKLENGKPLSMSEQTGLLKI KTEPLDFNDYKVL MATHGFSGSSPF
GATSPLGVHPSAQSPMQHLGVGMEAPLLGFPTMNSNLSEVQKVLQIVDNTVSRQK
EDISKLKGYHMKDPCSQP EEQGV TSPNIPPVGLPVVSHNGATKSIIDY TLEK VNE
QSLTTDSRRQISNIKKEKLRTLIDLVTDDKMIENHSISTPFSCQFCKESFPGPPI
RYLCKMNEEIKAVLQPHENIVPNKAGVFVDNKALLSSVLSEKGLTSPINPYKDH
AYYAMNMEPNSDELLKISIAVGLPQEFVK EWFEQRKVYQYSNSRSPSLERTSKPL
PTTKDSL LPRSPVKPMDSITSPSIAELHNSVTSCDPPLRLTKSSHFTNIKA VDKL
NTPSPLNLSSSTSSKNSHSSSYTPNSFSSEELQAEPLDLSLPKQ MREP KGIATKN
SINLDHNSVSSSSSENSDEPLNLTFIKKEFSNSNNDLNKSNNPVFGMNPFS AKPLY
QSAFPPATFMPPVQTSIPGLRPYPGLDQMSFLPHMAYTYPTGAATFADMQQRKY
FQGDLLDGAQDYMSGLDDMTDSDSCLSRKKIKKTESGMYACDLCDKTFQKSSSL
HTGKRPHQCQICKKAFKHKHHLIEHSRLHSGEKPYQCDKCGKRFSHSGSYSQHMN
CKREAEEREAAEREAREKGHLEPTELLMNRAYLQ SITPQGYSDSEERESMPRDGE
EKEGEEGYGKLRRRDGDEEEEEEEEEEESENKSMDTDPETIRDEEETGDHSMDDSS
TKSDHEEDNMEDGME

Querying tools

chr2 (qA3) 2aA1 2aA3 2aB 2aC1.1 2aC1.3C2 2aC3 2aD 2aE1 E2 2aE3 2aE5 2aF1 2aF3aG1 aG32aH1 H2 2aH3 2aH4



Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About

D. melanogaster BLAT Search

BLAT Search Genome

Genome: ☐ Search all
D. melanogaster

Query type: Release 6 + ISO1 MT/dm6) BLAT's guess

Tools

- Blat
- In-Silico PCR
- Table Browser
- LiftOver
- Gene Sorter
- Variant Annotation Integrator
- Data Integrator
- Genome Graphs
- Gene Interactions
- Other Tools

☐ All Results (no minimum matches)

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by line sequence name.

File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence: No file chosen

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.

The Ensembl database


[BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#)

ensembl.org

Tools

[All tools](#)

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species

for

Go

e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

All genomes

-- Select a species --



Pig breeds

[Pig reference genome](#) and [12 additional breeds](#)

[View full list of all species](#)

Favourite genomes



Human

GRCh38.p13

[Still using GRCh37?](#)



Mouse

GRCh39

Ensembl Release 108 (Oct 2022)

- Changes in the default tracks in the Location view: cDNAs EST clusters (UniGene) CCDS to be removed when MANE Select is available
- RNASeq tracks including data from GeneSWITCH consortium for c
- Variation data for crab-eating macaque, pike-perch, prairie vole, Ja
- and collared flycatcher
- Retirement of postGAP tool

[More release news](#)

Ensembl Rapid Release

New assemblies with gene and protein annotation every two we

Note: species that already exist on this site will continue to be updated the full range of annotations.

Go

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#)


[BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#)


Zebrafish (GRCz11)

[Location-based displays](#)
[Whole genome](#)
[Chromosome summary](#)
[Region overview](#)
[Region in detail](#)
[Comparative Genomics](#)
[Synteny](#)
[Alignments \(image\)](#)
[Alignments \(text\)](#)
[Region Comparison](#)
[Genetic Variation](#)
[Variant table](#)
[Resequencing](#)
[Linkage Data](#)
[Markers](#)
[Other genome browsers](#)
[UCSC](#)
[NCBI](#)

[Configure this page](#)
[Custom tracks](#)
[Export data](#)
[Share this page](#)
[Bookmark this page](#)

Chromosome 10: 10,138,322-10,349,251

[Add/remove tracks](#) | [Share](#) | [Export image](#) | [Reset configuration](#)

Assembly exceptions
Chr. 10
Assembly exceptions

Region in detail

[Add/remove tracks](#) | [Share](#) | [Resize image](#) | [Export image](#) | [Reset configuration](#) | [Reset track order](#) | [Switch image](#) | Scroll: | Track height: | [Drag/Select](#)

Assembly exceptions
Contigs
Genes
(Comprehensive set)

Gene Legend
 merged Ensembl/Havana
 RNA gene
 processed transcript

[Go](#)
 [Go](#)

[Add/remove tracks](#) | [Custom tracks](#) | [Share](#) | [Resize image](#) | [Export image](#) | [Reset configuration](#) | [Reset track order](#) | [Drag/Select](#)

Assembly exceptions
65 way GERP elements
cDNA alignment
EST_cluster (Unigene)

Constrained elements for 65 fish EPO-Extended
Showing all 8 features - click to show fewer

The Ensembl database

ensembl.org/index.html

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

ensembl.org

Tools

[All tools](#)

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports comparative genomics, evolution, sequence variation and transcription. Ensembl annotate genes, computes multiple alignments, predicts regulatory elements and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 108 (Oct 2019)

Search

All species for

Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

All genomes

-- Select a species --



Pig breeds

Pig reference genome and 12 additional breeds

[View full list of all species](#)

Favourite genomes



Human

GRCh38.p13



Mouse

GRChm39

[Still using GRCh37?](#)

ensembl.org/Danio_rerio/Location/View?r=10:10138322-10349251

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Zebrafish (GRCz11)

Location: 10:10,138,322-10,349,251

- Location-based displays
 - Whole genome
 - Chromosome summary
 - Region overview
 - Region in detail
- Comparative Genomics
 - Synteny
 - Alignments (image)
 - Alignments (text)
 - Region Comparison
- Genetic Variation
 - Variant table
 - Resequencing
 - Linkage Data
- Markers
- Other genome browsers
 - UCSC
 - NCBI

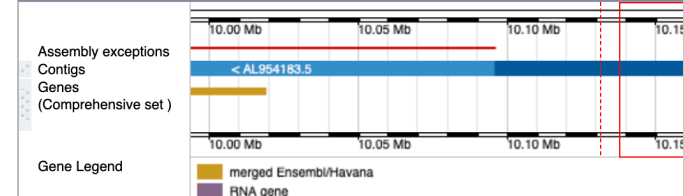
Chromosome 10: 10,138,322-10,349,251

Add/remove tracks | Share | Export image | Reset configuration

Assembly exceptions
Chr. 10
Assembly exceptions

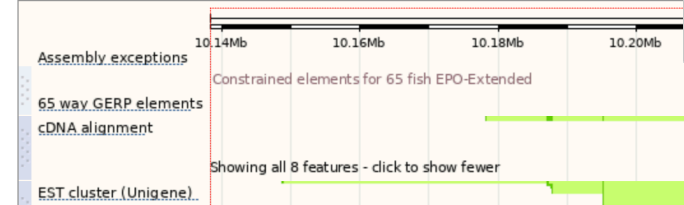
Region in detail

Add/remove tracks | Share | Resize image | Export image | Reset configuration



Location: 10:10138322-10349251 Go Gene:

Add/remove tracks | Custom tracks | Share | Resize image | Export image



ncbi.nlm.nih.gov/gene

An official website of the United States government [Here's how](#)



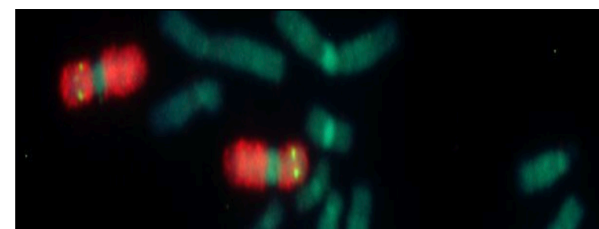
National Library of Medicine
National Center for Biotechnology Information

Gene

Gene

[Advanced](#)

Search



Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequence (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Using Gene

[Gene Quick Start](#)

[FAQ](#)

[Download/FTP](#)

[RefSeq Mailing List](#)

[Gene News](#)

[Factsheet](#)

Gene Tools

[Submit GeneRIFs](#)

[Submit Correction](#)

[Statistics](#)

[BLAST](#)

[Genome Workbench](#)

[Splign](#)

Other Resources

[OMIM](#)

[RefSeq](#)

[RefSeqGene](#)

[Protein Clusters](#)

Representative queries

Find genes by...

free text

chromosome and symbol

partial name and multiple species

associated sequence accession

gene name (symbol)

Search text

[human muscular dystrophy](#)

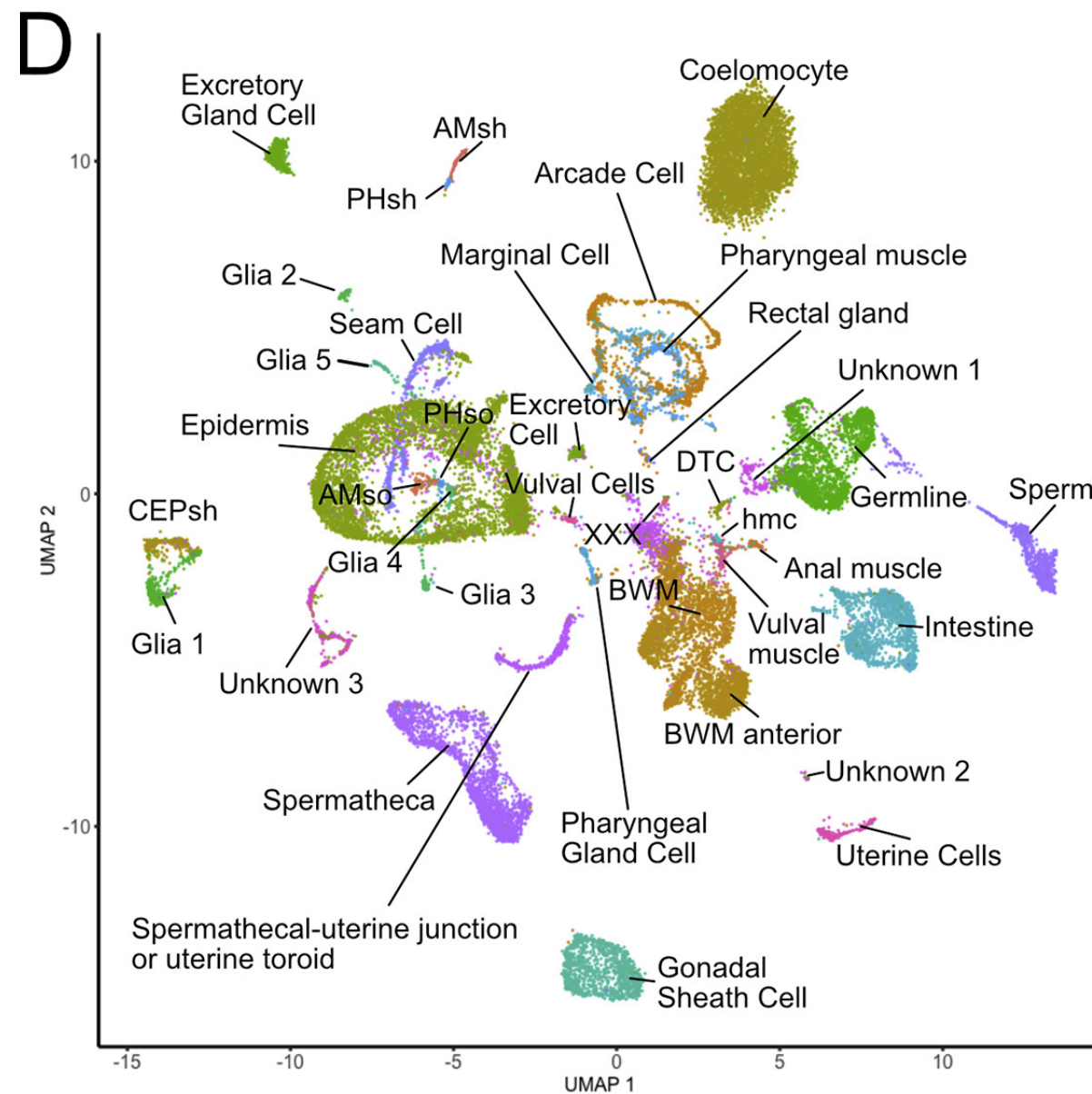
[\(\[chr\] OR 2\[chr\]\) AND adh*\[sym\]](#)

[alive\[prop\] AND transporter\[title\] AND \("Drosophila melanogaster"\[orgn\] OR "Mus musculus"\[orgn\]\)](#)

[M11313\[accn\]](#)

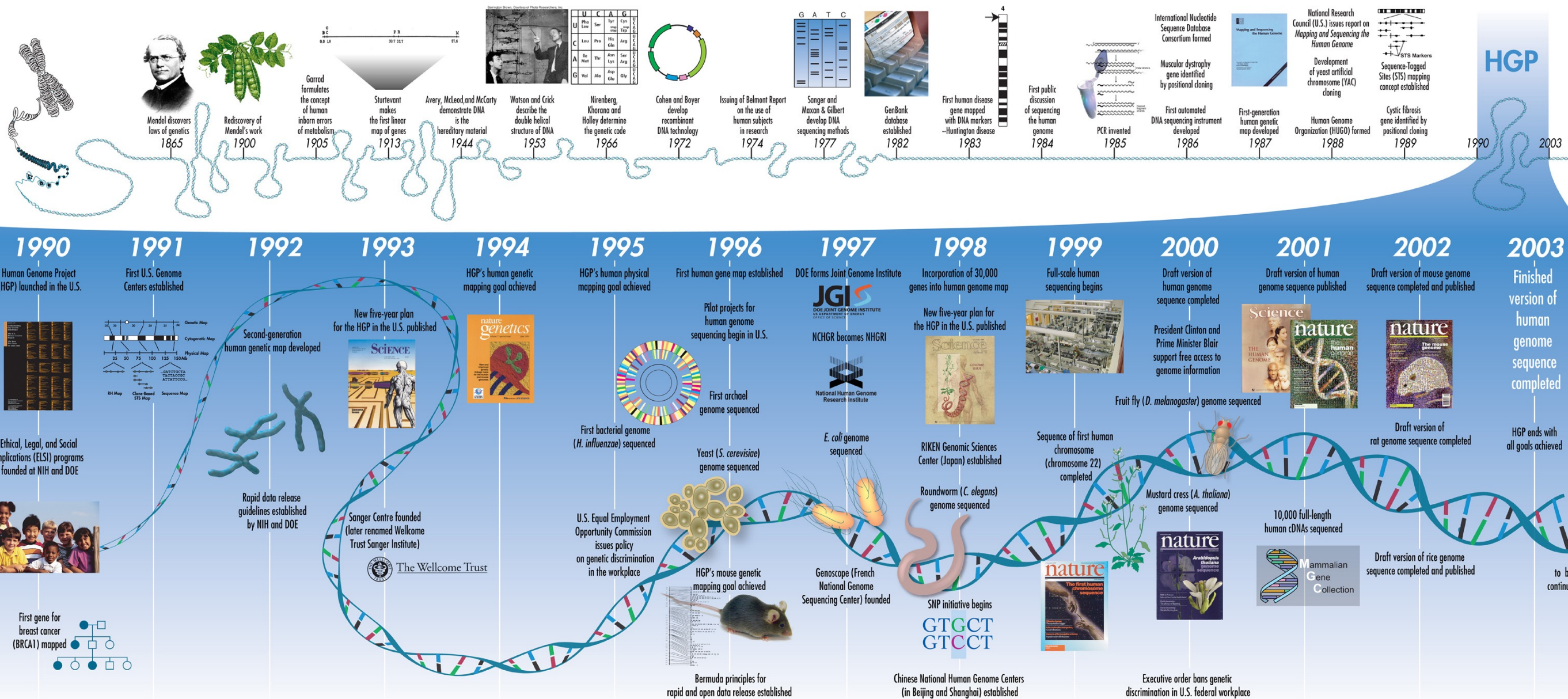
[BRCA1\[sym\]](#)

Maps vs analytics



Taylor SR et al. Molecular topography of an entire nervous system. Cell (2021).

History of genome sequencing projects



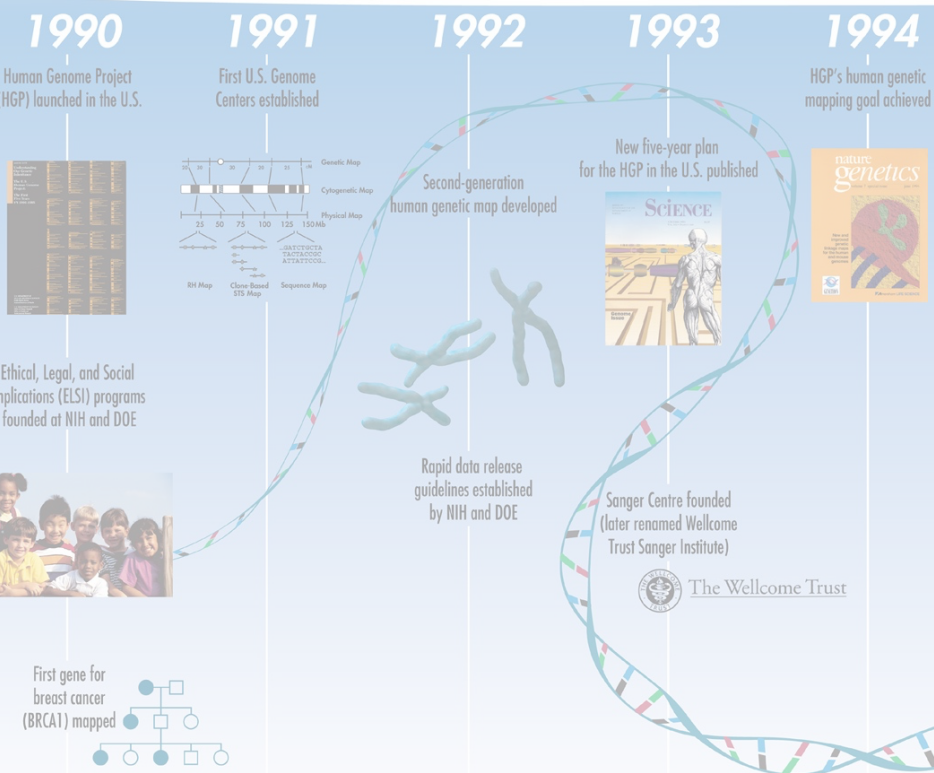
genome.gov

COMMENT • 10 FEBRUARY 2021

A wealth of discovery built on the Human Genome Project – by the numbers

A new analysis traces the story of the draft genome's impact on genomics since 2001, linking its effects on publications, drug approvals and understanding of disease.

Alexander J. Gates, Deisy Morselli Gysi, Manolis Kellis & Albert-László Barabási

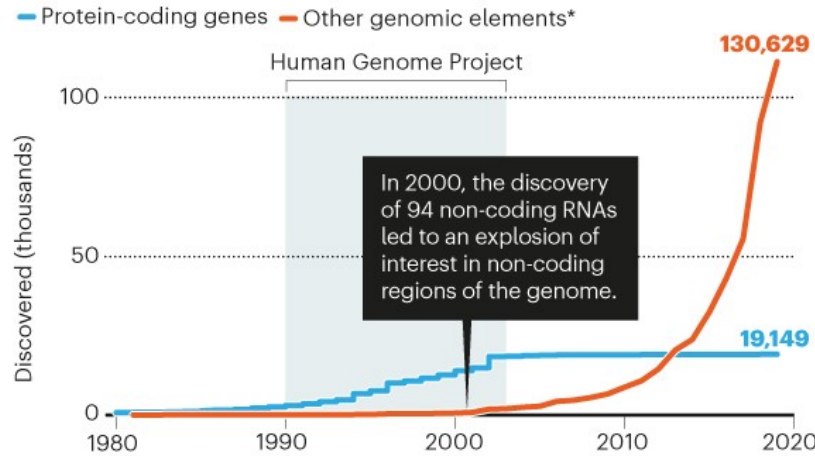


TWENTY YEARS OF JUNK, STARS AND DRUGS

What genomics researchers have studied, when and why — as traced by bibliometric analysis.

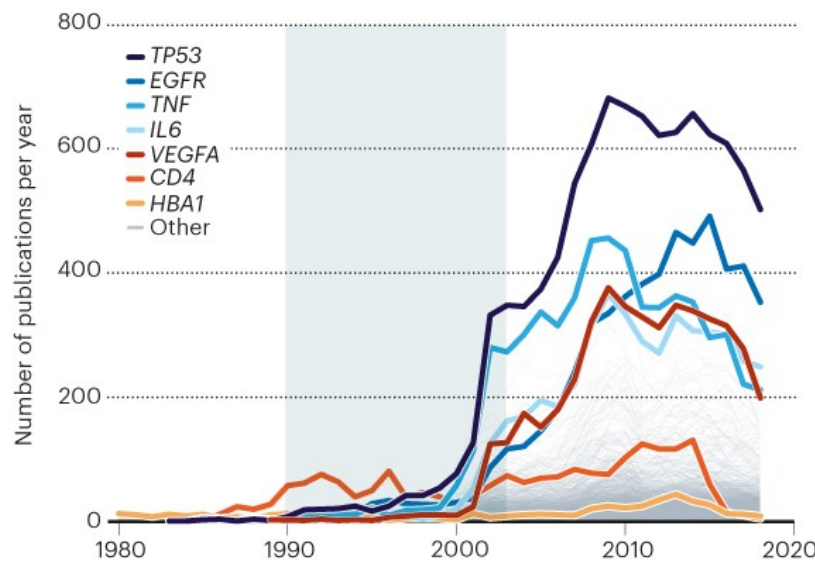
Non-coding elements

Most protein-coding genes were discovered before the first draft of the Human Genome Project (HGP) in 2001. Many other genomic elements, previously called junk DNA, came in for scrutiny after that.



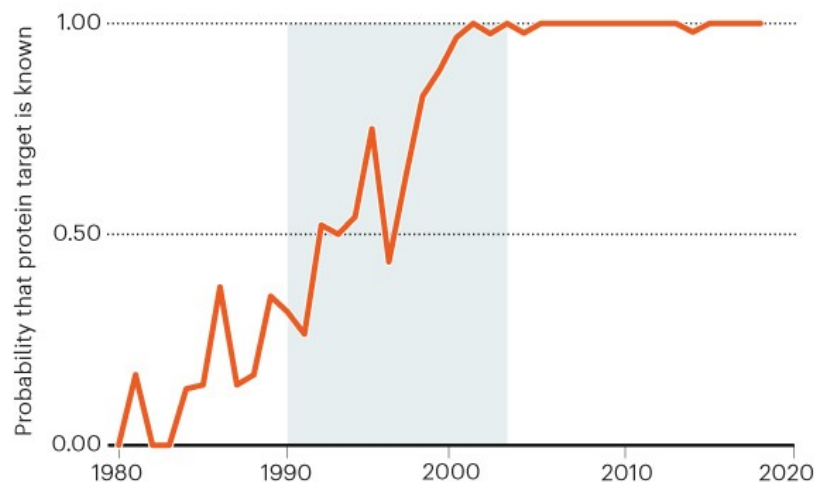
Star genes

The HGP gave rise to an explosion in research concentrated on just a few genes.



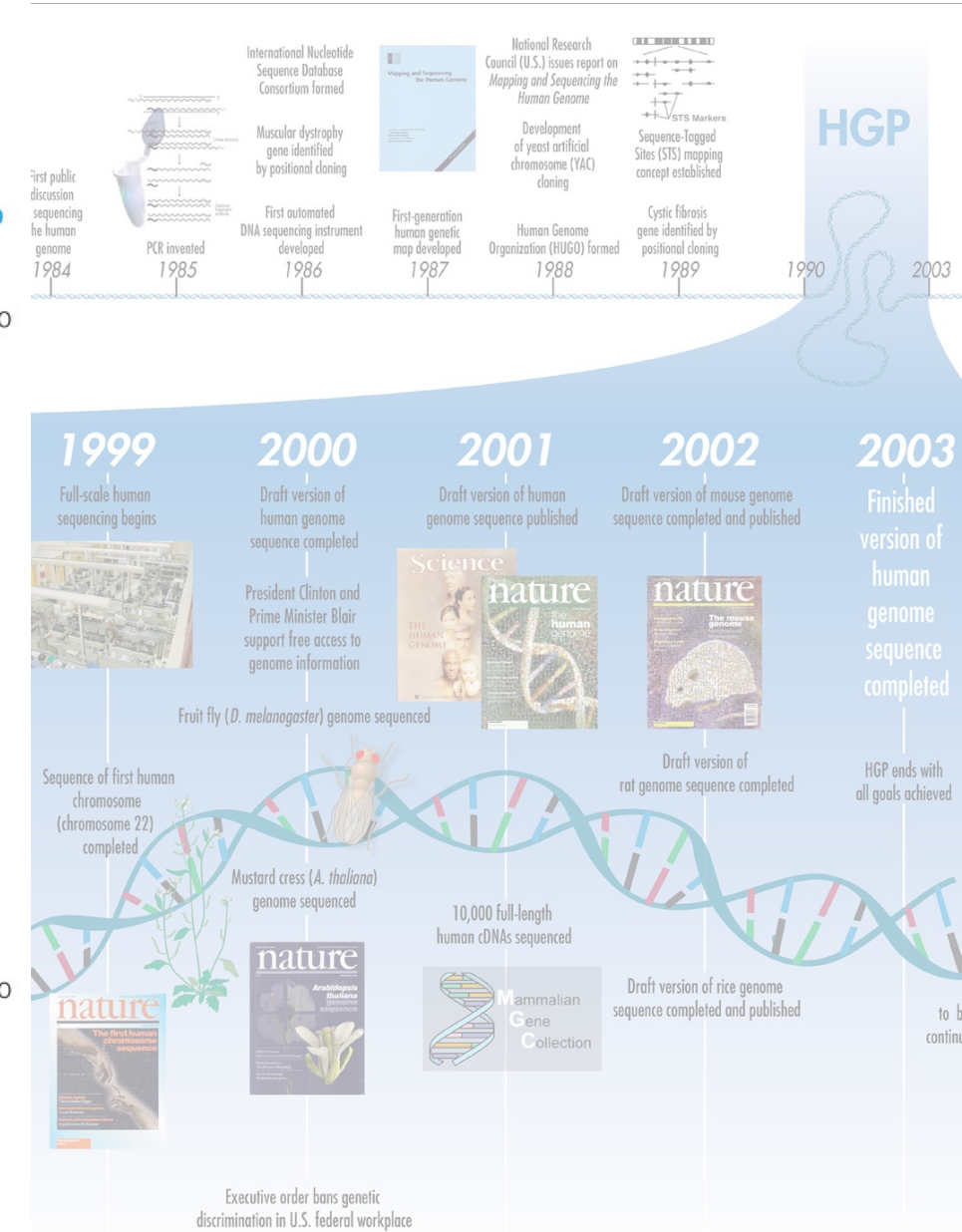
Drug targets

Since 2001, nearly 100% of US drugs licensed in any given year have had all their potential protein targets identified.



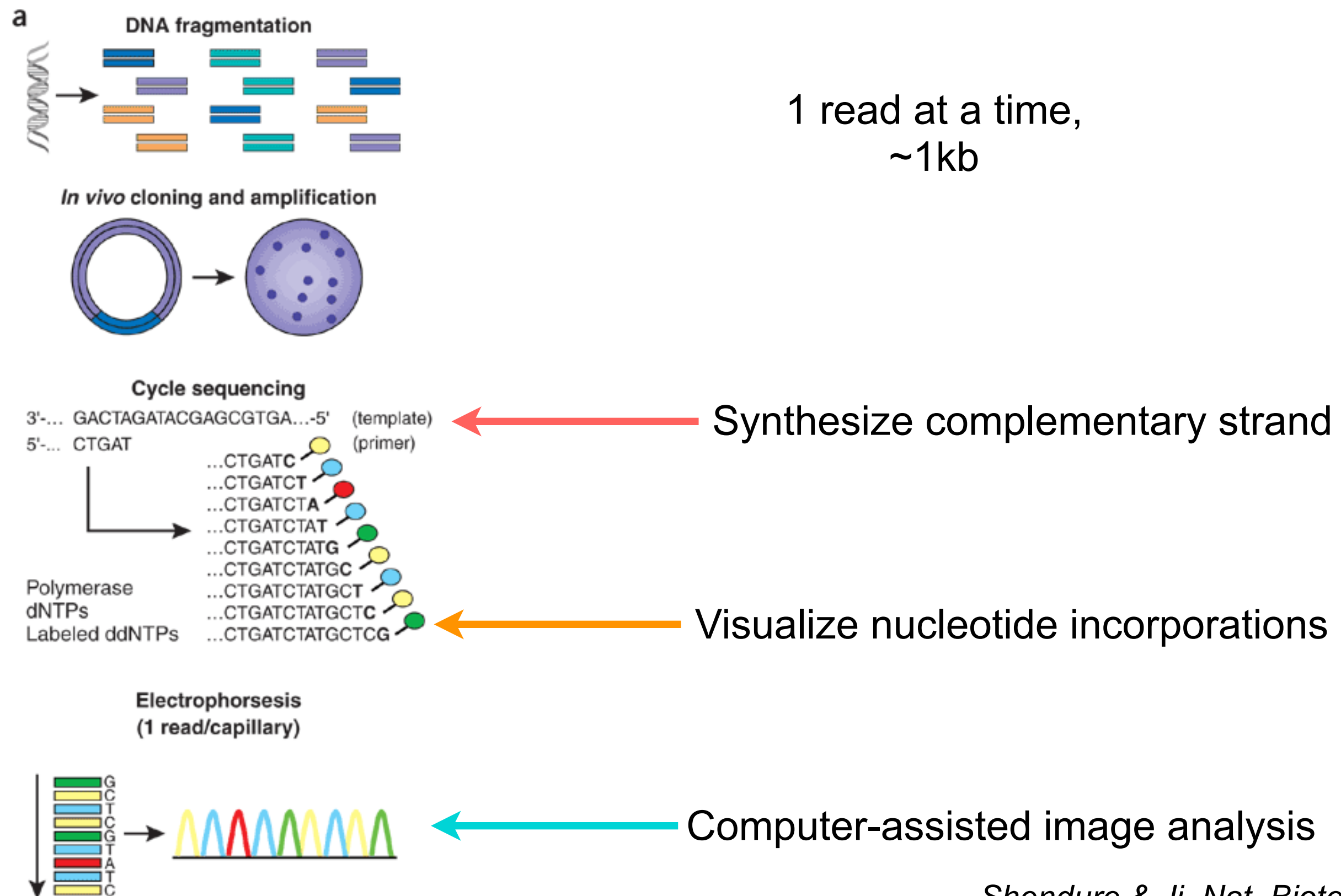
*Including single nucleotide polymorphisms, pseudogenes, non-coding RNAs, promoters and so on.

©nature



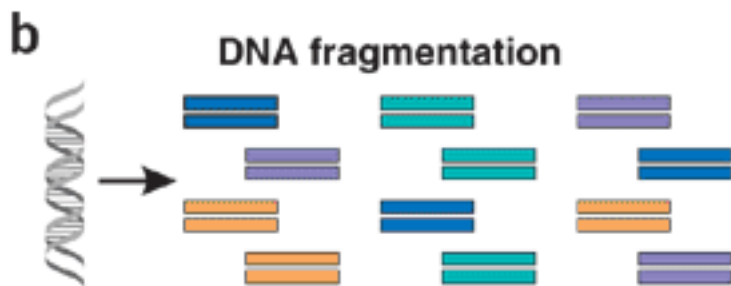
Sequencing technologies

Sanger (capillary sequencing)



Sequencing technologies

High-Throughput sequencing

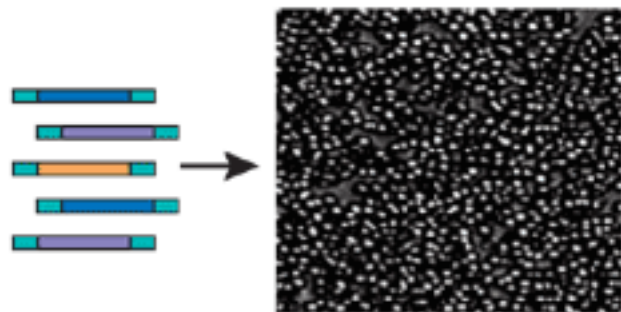


10^7 - 10^9 reads in parallel,
limited to ~200bp

In vitro adaptor ligation



Generation of polony array



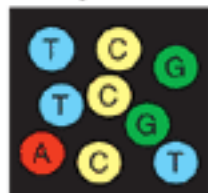
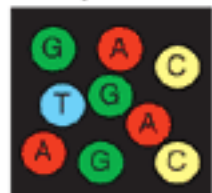
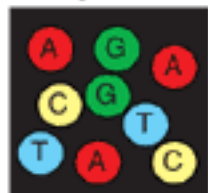
On-chip clonal colonies of original fragment

Cyclic array sequencing
($>10^6$ reads/array)

Cycle 1

Cycle 2

Cycle 3

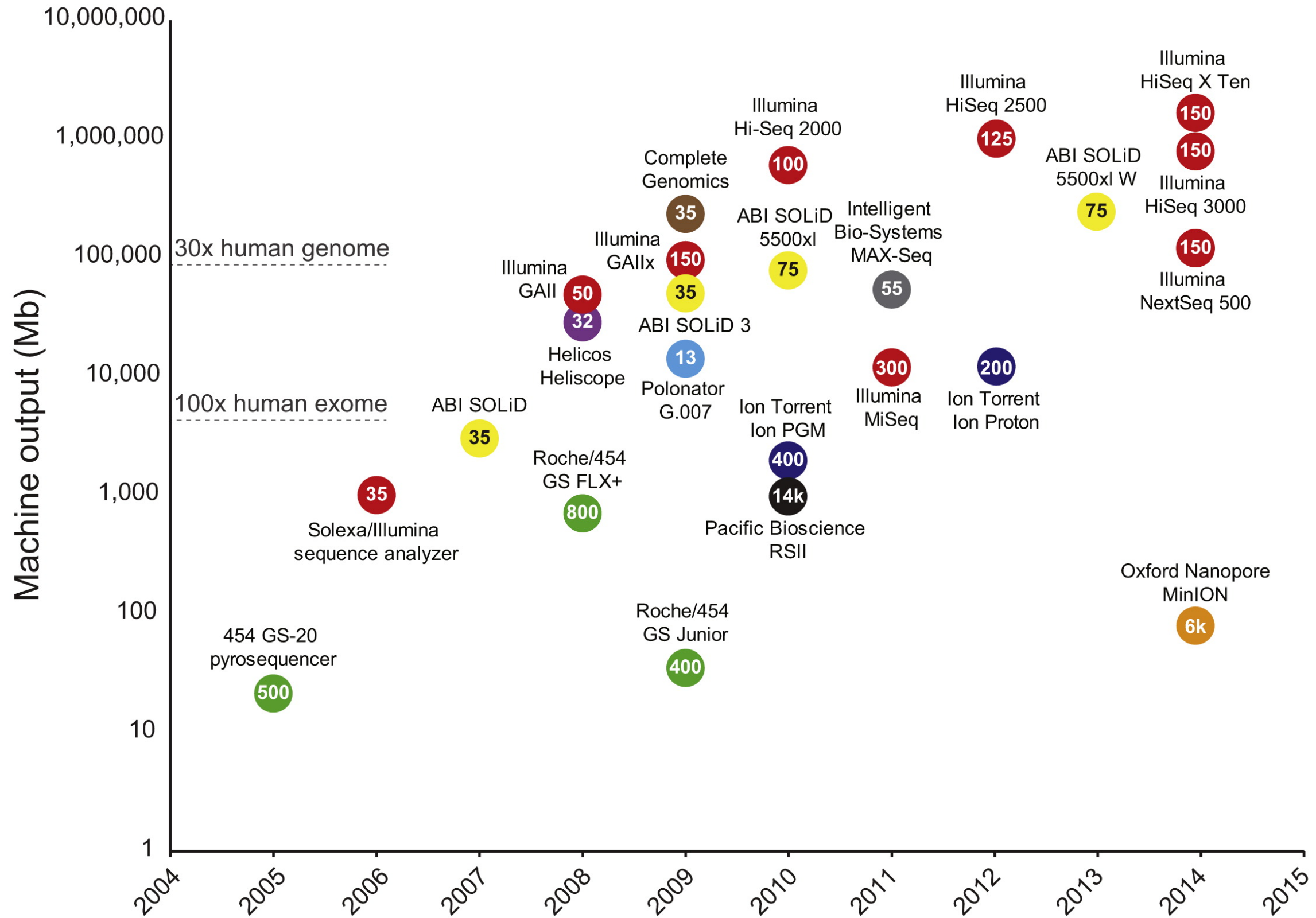


Optical reading

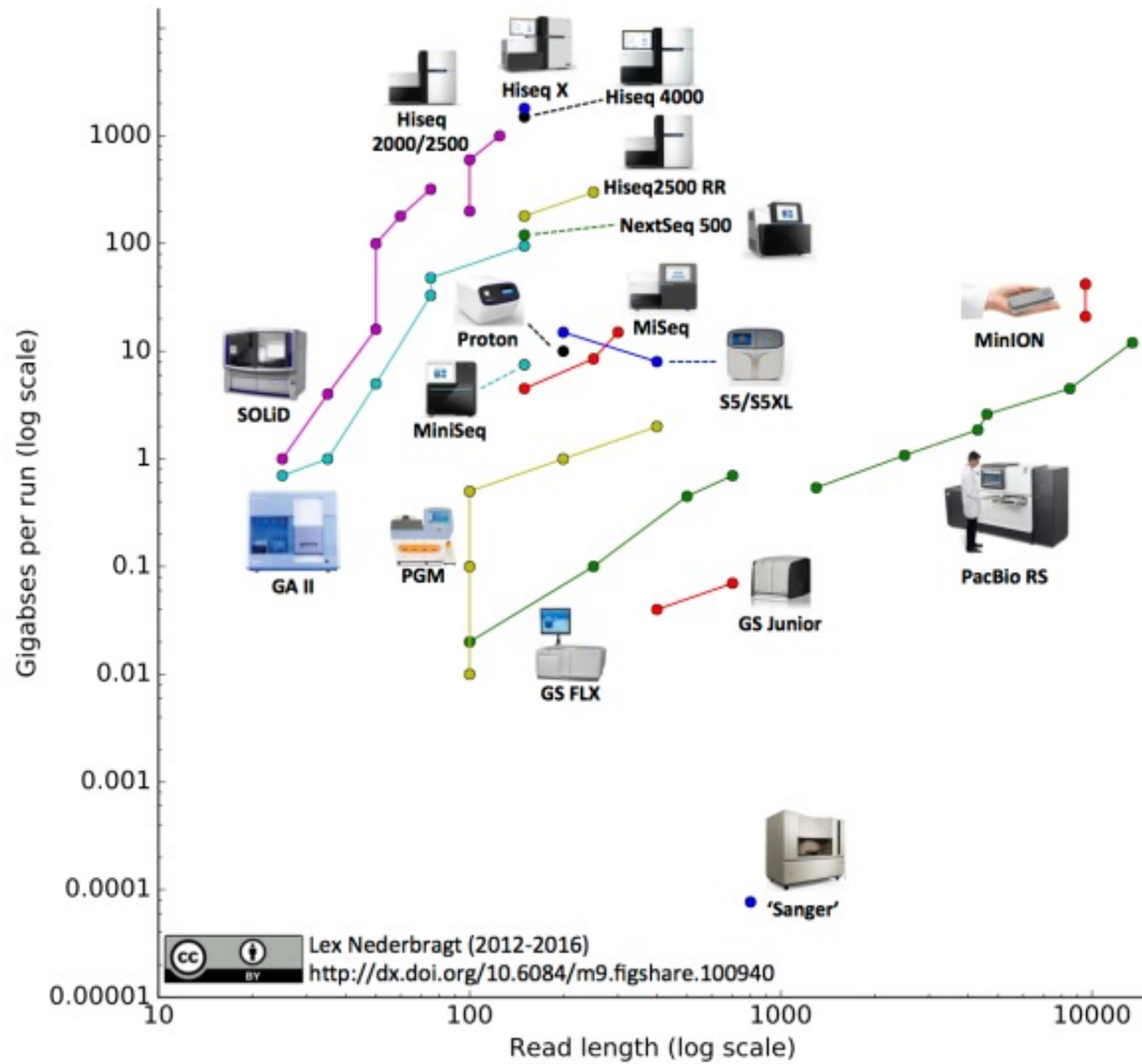
What is base 1? What is base 2? What is base 3?

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

Sequencing technologies



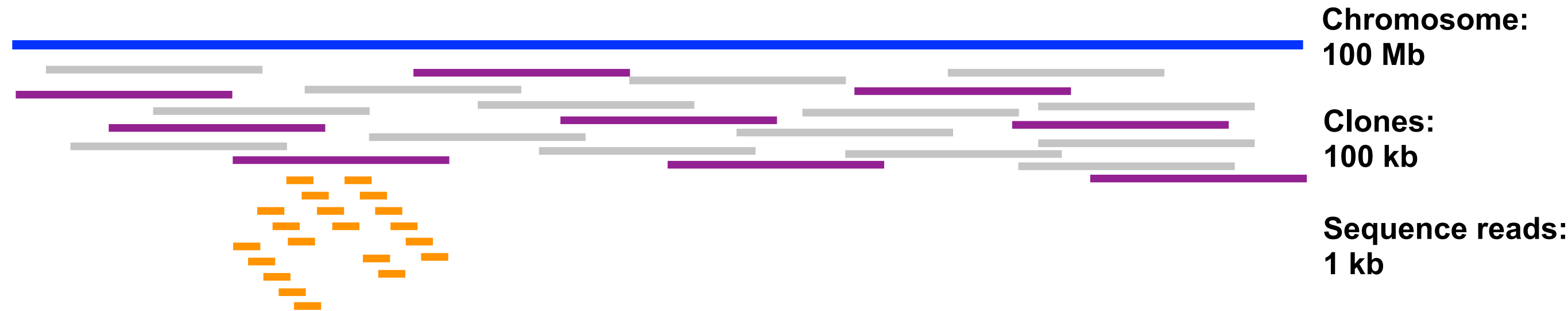
Sequencing technologies



Human Genome Project: 1990-2001

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 chromosome
- Use paired-end sequencing to resolve repeats

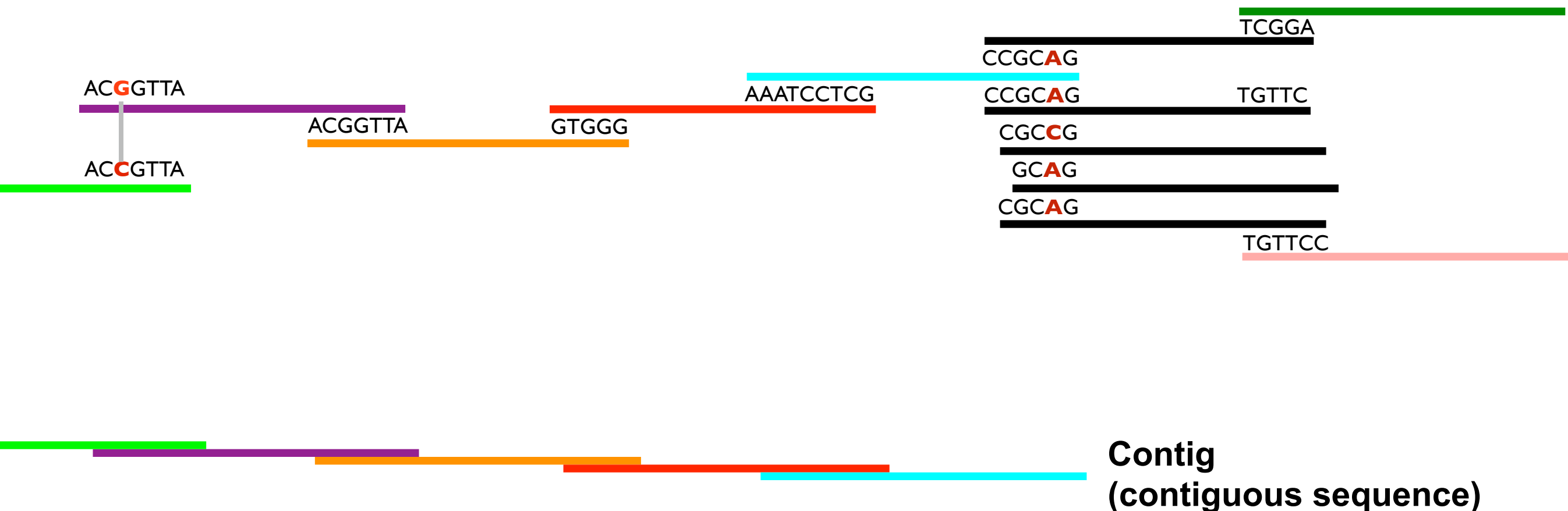
Fragments assembly

General Procedure:

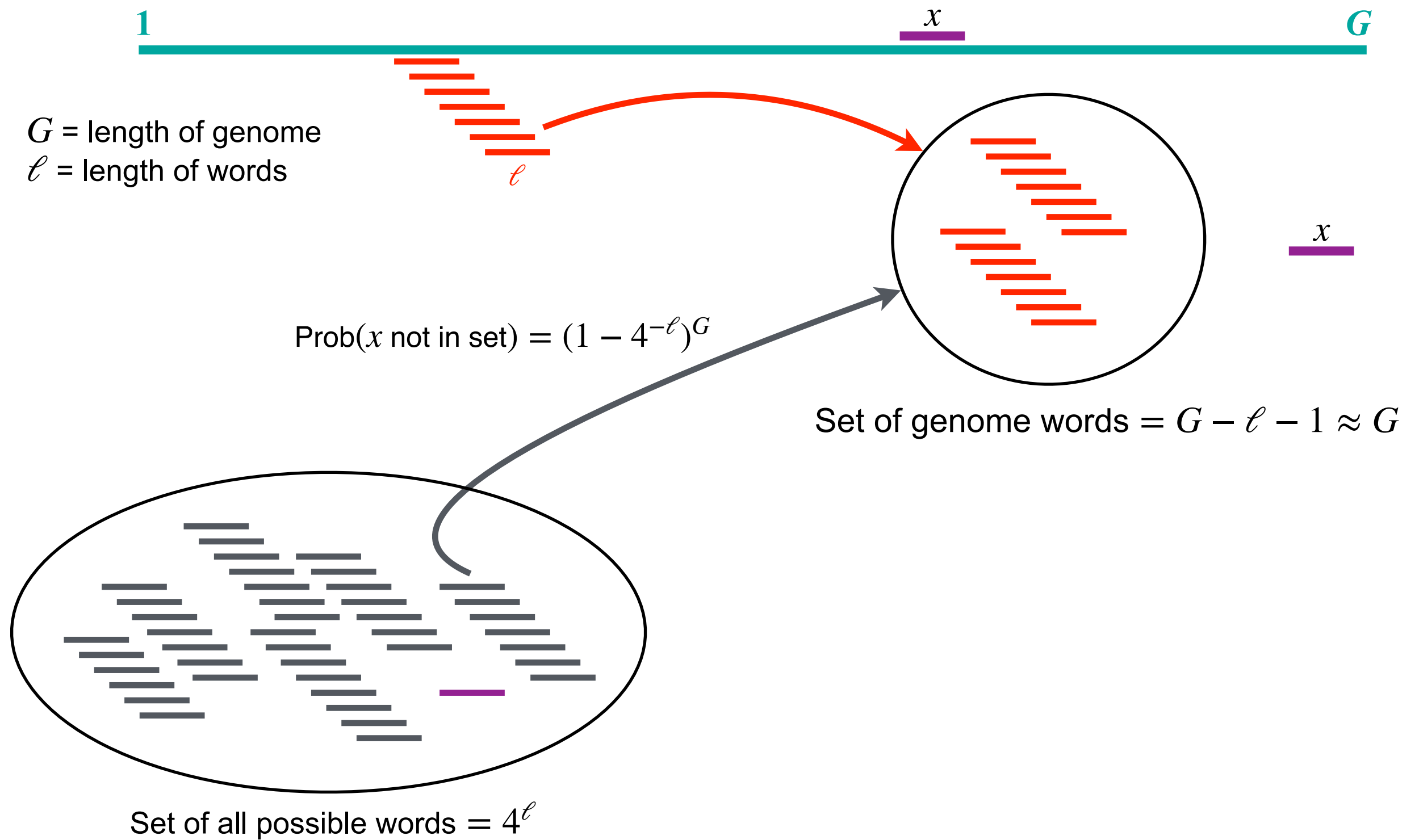
- Overlap → Layout → Consensus

Difficulties:

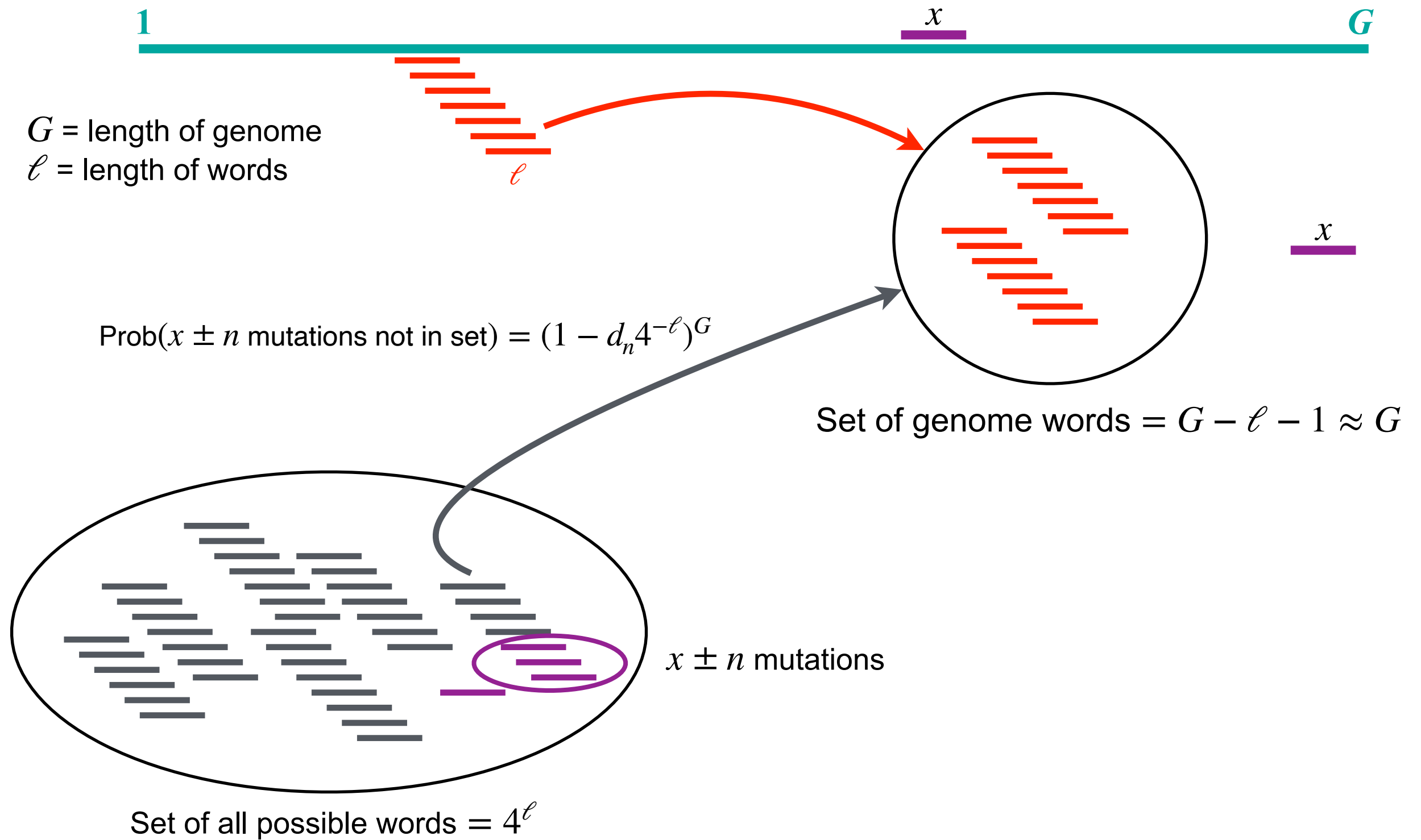
- Computing overlap with sequencing errors (1-3%) and unknown orientation



Overlap size

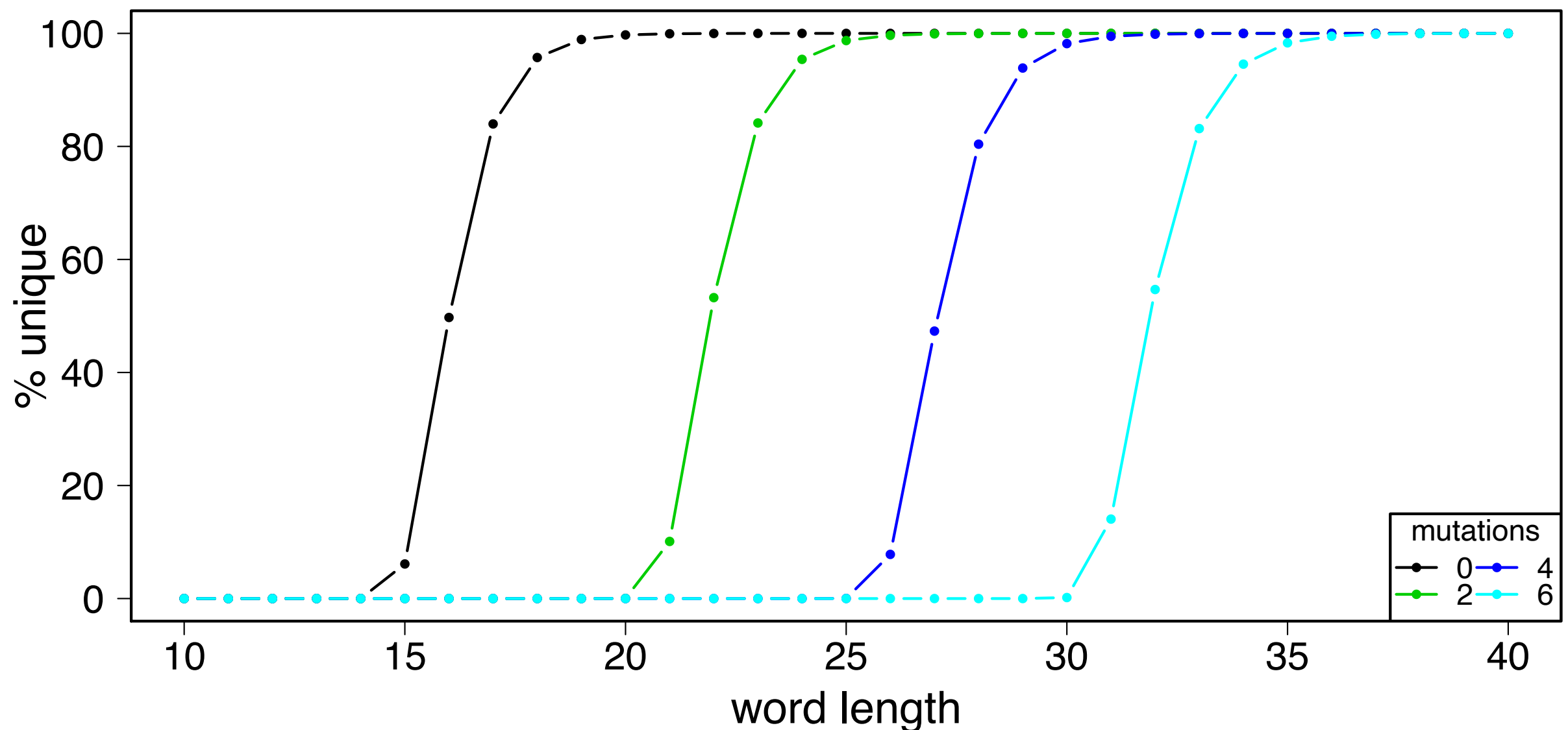


Overlap size

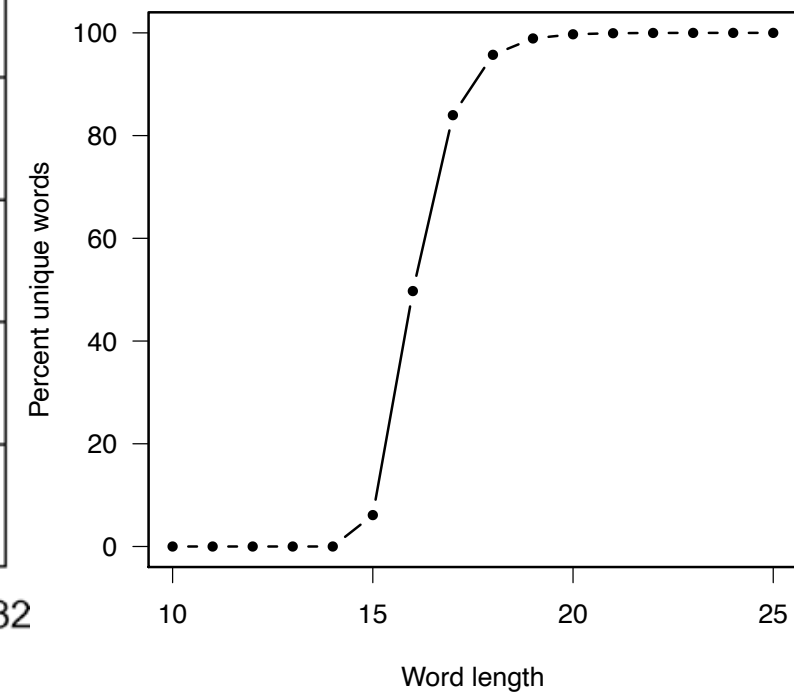
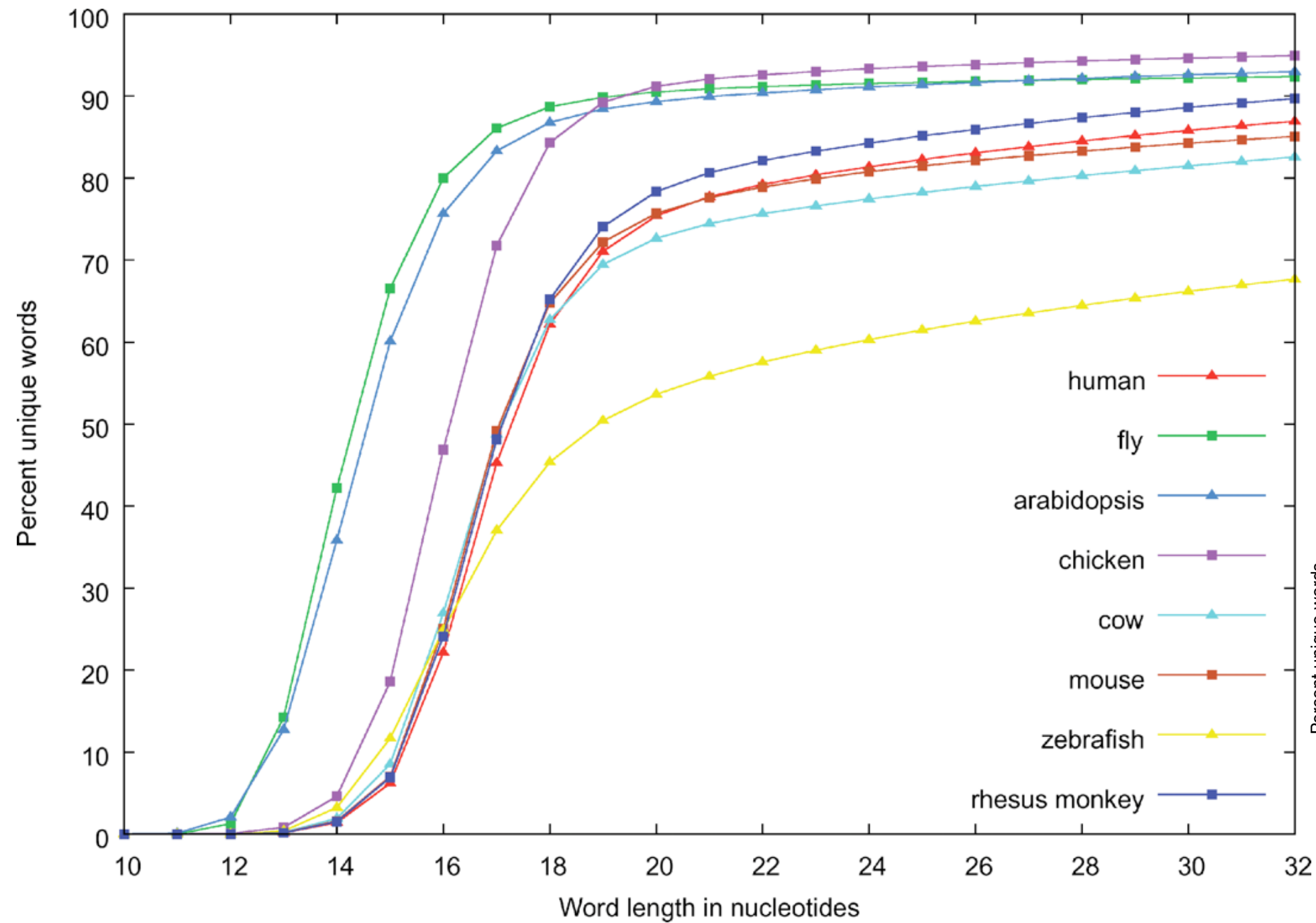


Overlap size

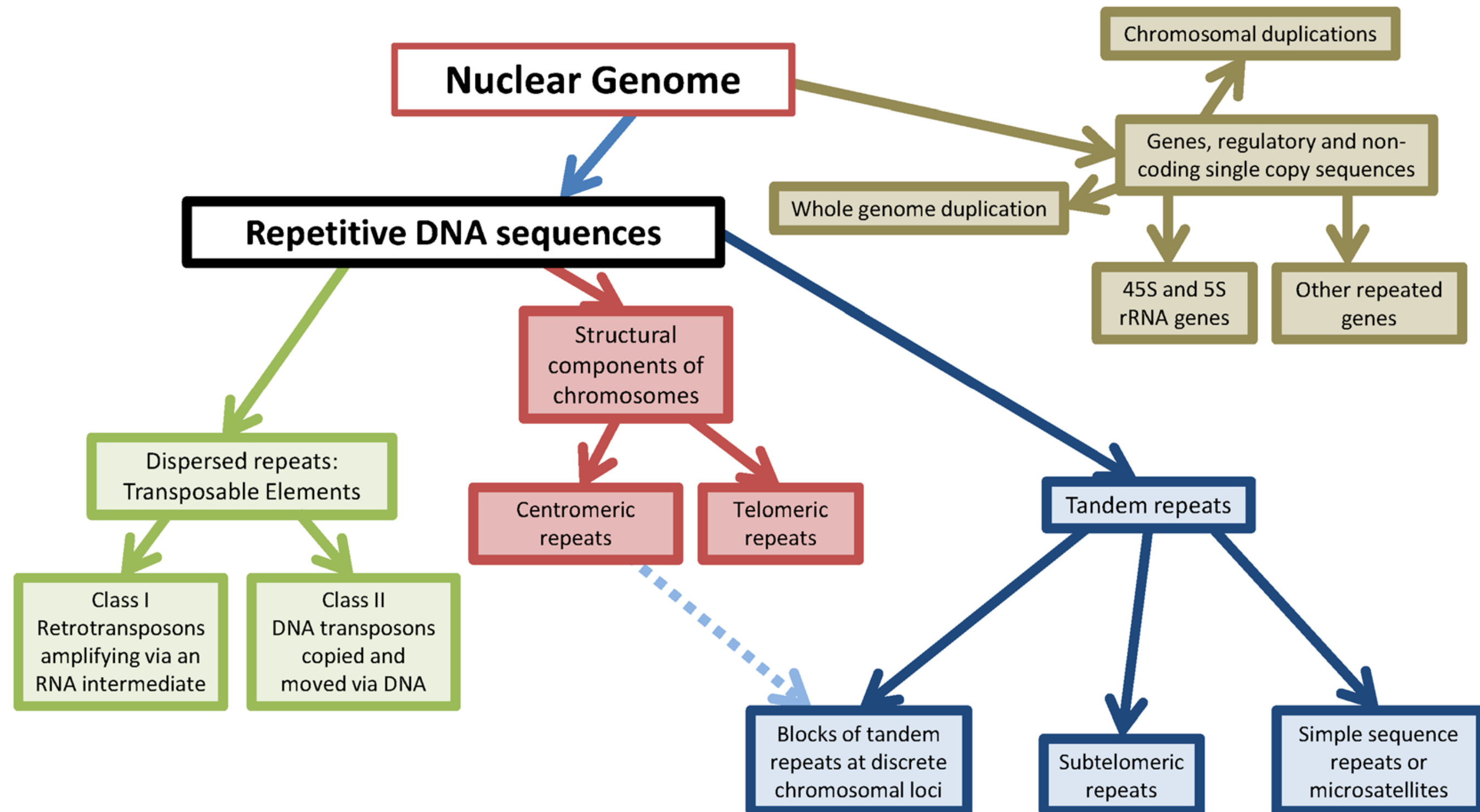
- Human genome is 3Gb, $\log_4(3 \times 10^9) = 15$
- We are assuming up to 3% errors, so two words with a few differences should be considered the same
- If we use 35mers with up to 6 “mistakes”, this is still “unique” in the genome



Digression: sequence repeats



Digression: sequence repeats

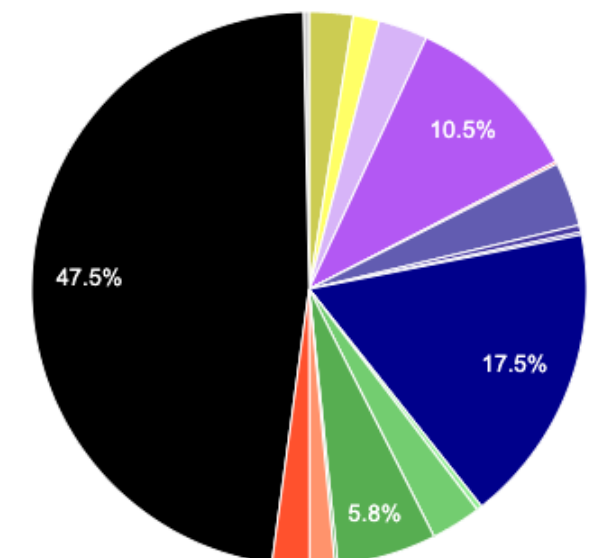


Digression: sequence repeats

	WGD	tDNA	% Masked
Yeast	1	299	-
Drosophila	0	295	20%
Mouse	2	463	42%
Human	2	691	48%

Genome Fraction

- **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

Databases and tools for repetitive sequences

- **RepBase, Dfam:** databases 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

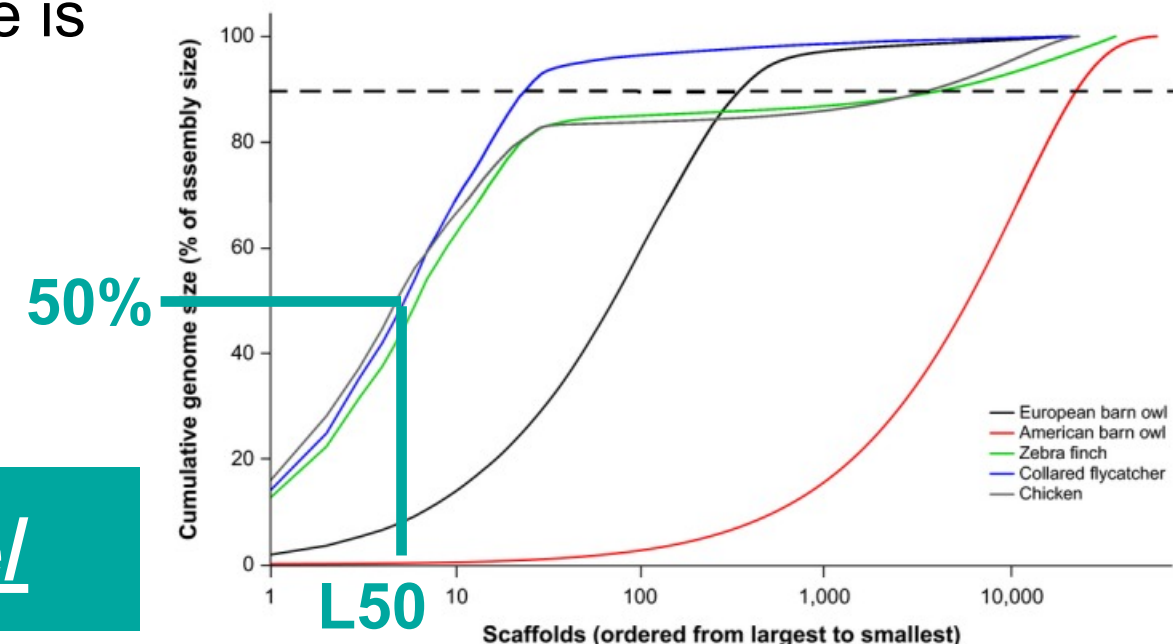
dfam.org

girinst.org

Outcome of genome sequencing programs

	Size (Mb)	GC content	Nb genes	N50 (Mb)/L50
Yeast	12	38%	6'014	0.92 / 6
Drosophila	144	42%	13'962	21 / 3
Mouse	2'700	42%	22'198	59 / 15
Human	3'100	41%	20'080	58 / 18

- N50: size of smallest contig such that 50% genome is covered (L50 = nb of contigs)
- Mycobacterium Tuberculosis GC: 66%



ncbi.nlm.nih.gov/datasets/genome/

Vertebrate genomes project



PROJECT PHASE I

PHASE 1 OF THE VGP WILL GENERATE NEAR ERROR-FREE REFERENCE GENOMES OF 260 SPECIES REPRESENTING ALL VERTEBRATE ORDERS WITH A DIVERGENCE TIME OF ~50 MILLION YEARS AGO (MYA) OR GREATER FROM THEIR MOST RECENT COMMON ORDINAL ANCESTOR, INCLUDING HUMAN AND SOME SPECIES ON THE BRINK OF EXTINCTION. WE WILL SEQUENCE THE HETEROGAMETIC SEX (WHEN IT EXISTS) SO THAT BOTH SEX CHROMOSOMES CAN BE ASSEMBLED FOR EACH SPECIES.



54

BIRDS



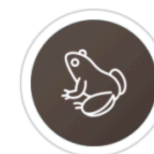
58

MAMMALS



32

REPTILES



30

AMPHIBIANS



90

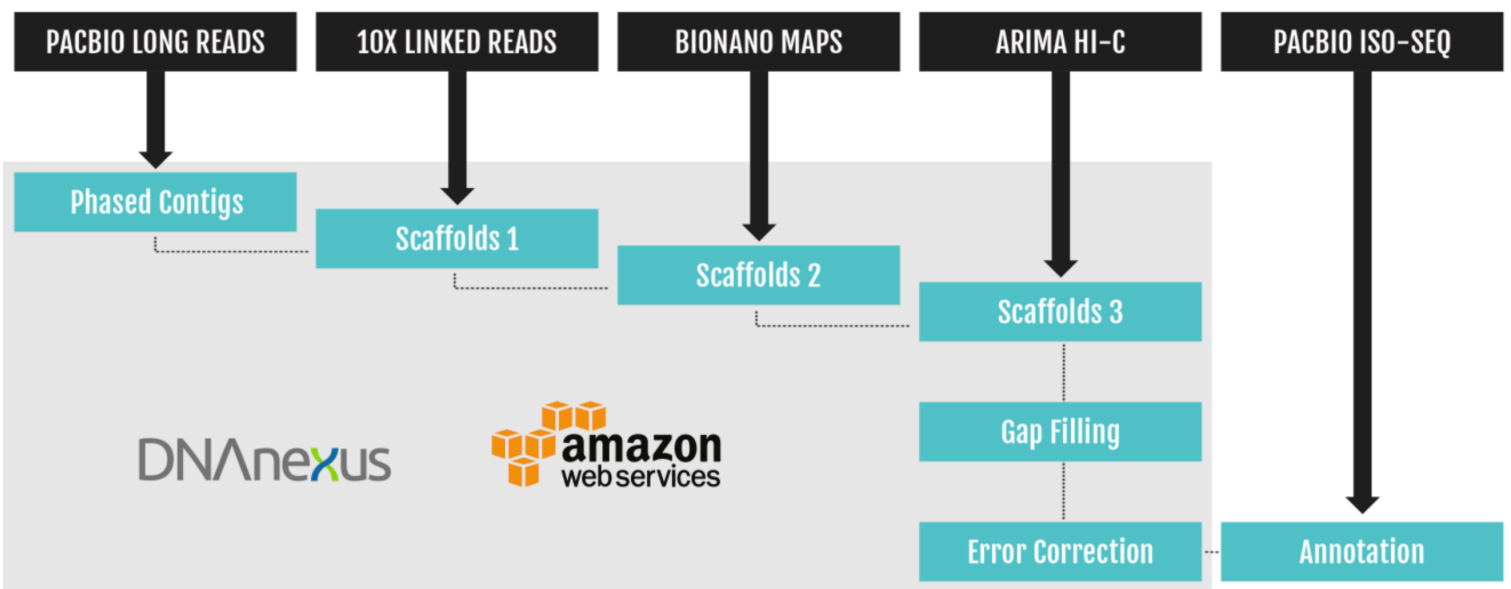
FISH



4

INVERTEBRATES

OVER 260 SPECIES



vertebratengenomesproject.org

Vertebrate genomes project

Long reads
Building blocks



Linked reads
Scaffold



Optical maps
Scaffold



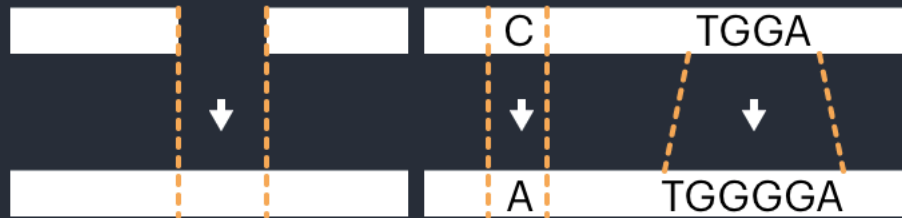
Hi-C
Scaffold



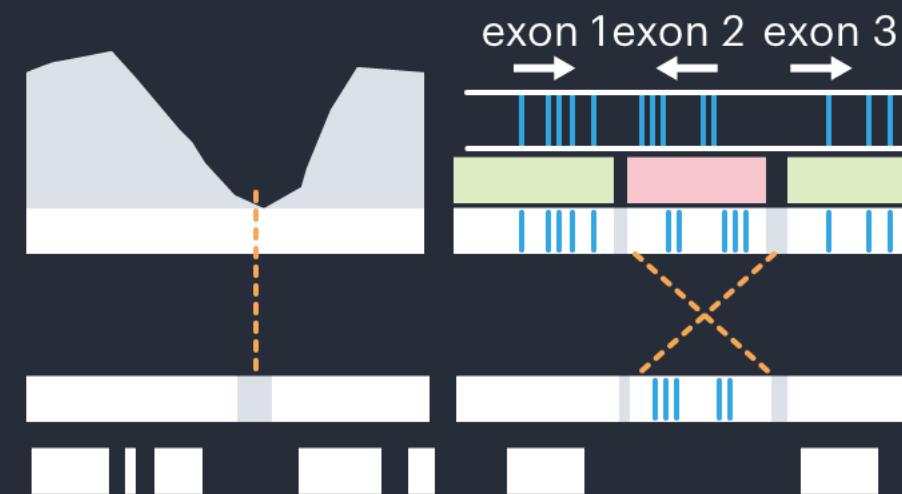
Gap-filling & polishing



Draft assembly



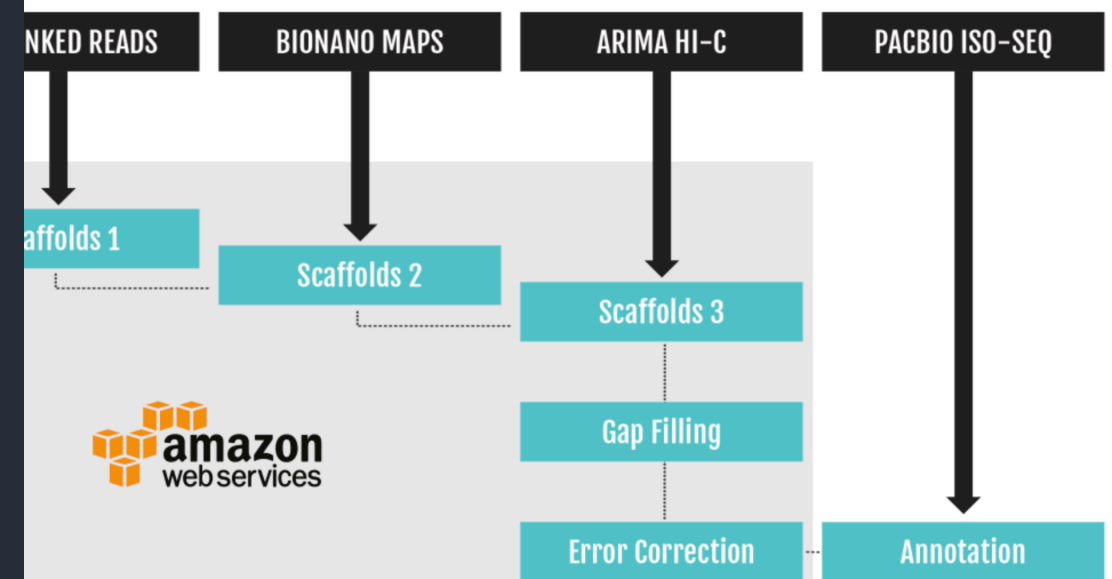
Curated assembly



ERROR-FREE REFERENCE GENOMES OF 260 SPECIES REPRESENTING ALL VERTEBRATE ORDERS WITH A
GO (MYA) OR GREATER FROM THEIR MOST RECENT COMMON ORDINAL ANCESTOR, INCLUDING HUMAN AND
N. WE WILL SEQUENCE THE HETEROGAMETIC SEX (WHEN IT EXISTS) SO THAT BOTH SEX CHROMOSOMES



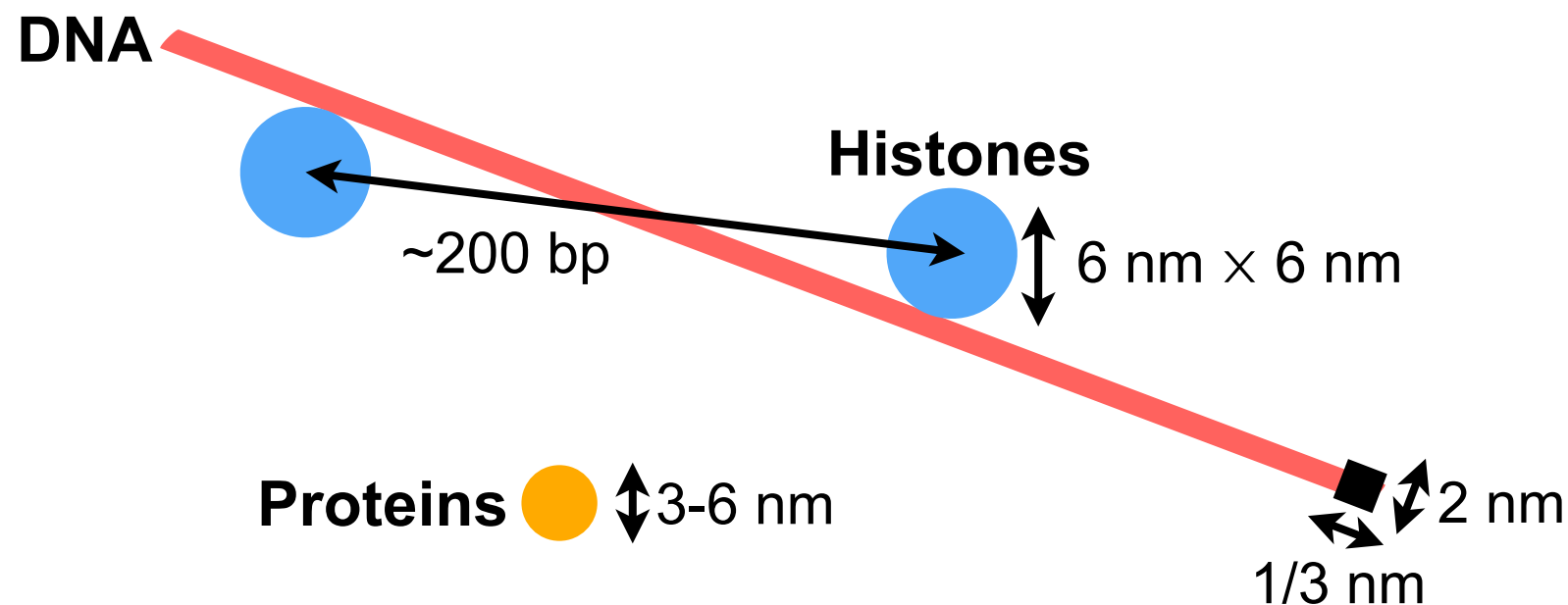
OVER 260 SPECIES



vertebrategenomesproject.org

Using numbers

- **Human body:** 10^{14} cells
- **Size of a mammalian cell:** 100-10'000 μm^3 (yeast: 20-100 μm^3)
- **Size of a nucleus:** 300 μm^3 (10% of cell)



Using numbers

- **Cell cycle time:** ~1 day (yeast: 200 min.)
- **Transcription:** 50-70 nt/sec., average 3-4 kb/min
- **Translation:** 10-20 aa/sec
- **Number of mRNA per cell:** 300'000 (human), 15'000 (yeast), 1380 (e. coli)
- **Mutation rate:** 10^{-8} per basepair, per duplication
- **Misincorporation rate:**
 - $10^{-4} - 10^{-5}$ per nucleotide (transcription)
 - $10^{-3} - 10^{-4}$ per amino acid (translation)

Exercises

AAC 231

