

INTRODUCTION to COMPUTATIONAL GENOMICS



Josep F. Abril, PhD

`jabril@ub.edu`

Computational Genomics Lab
<https://compgen.bio.ub.edu/>

Summary

- Historical Context.
- Growth of Sequence Data.
- Sequencing in the Post-Genome Era.
- Sequence Annotation.
 - Basic Concepts.
 - Data Formats in CG.
 - Genomic Sequences Stats.
 - Sequences as Strings.

Index Medicus [1879]
National Library of Medicine (NLM)

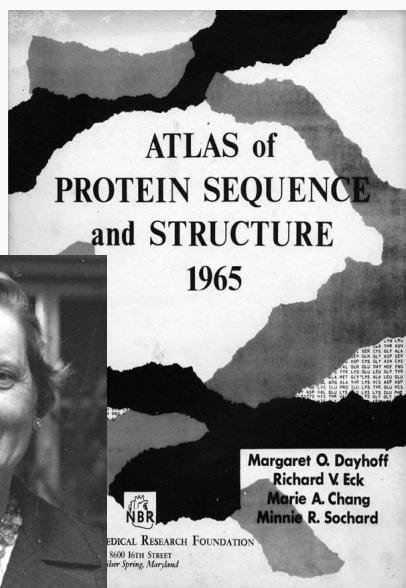
Raymond Gosling & Rosalind Franklin
DNA X-ray
diffraction pattern [1952]



James Watson &
Francis Crick **DNA**
structure model [1953]

First protein sequence
bovine **insulin** [1953]

X-ray cristalography 3D structure
myoglobin protein [1959]



Pioneering “Bioinformatics”
Margaret Dayhoff compiled
The Atlas of Protein Sequence and Structure
(~300 proteins) [1965], from those sequences
built first substitution matrices
(Point Accepted Mutations: PAM)



Ada Lovelace first “program” [1843]
for Charles Babbage Analytical Engine



Alan Turing's Machines [1936]

ENIAC [1943 – 1955[†]]

~18000 vacuum tubes
Electronic Numerical Integrator And Computer

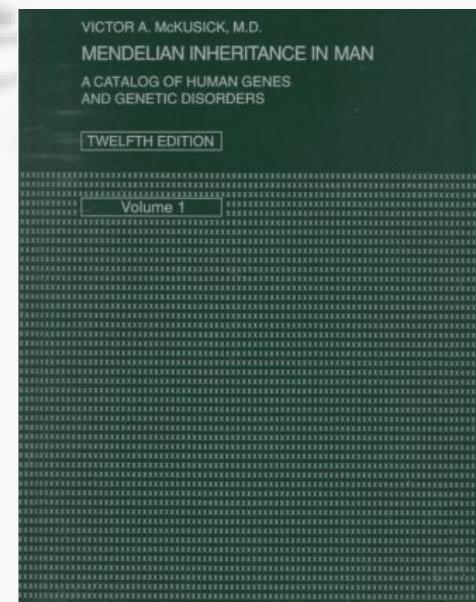
High-level Programming Languages
FORTRAN [1957]

Structured Query Languages
SQL [1969]

ARPAnet [1969]

UNIX [1970]

Dr. Victor A. McKusick
Mendelian Inheritance
in Man (MIM) [1966]
A catalog of mendelian
traits and disorders.



Protein Data Bank PDB [1971]

First database of protein structures

MEDLINE [1971]

Medical Literature Analysis and Retrieval System Online

Maxam & Gilbert vs Sanger DNA sequencing methods [1975]

First complete genome sequenced
Virus phi-X174 (~5400bp) [1977]

GenBank and EMBL [1982]

sequence databases
as of EMBL, june 1982,
contained **582 sequences**,
summing up to **600000bp**

First Human Gene Sequenced *Huntington's disease* [1983]

Kary B Mullis → PCR technique [1985]

Swiss-Prot [Amos Bairoch, 1986] Protein Sequence Database

National Center for Biotechnology Information (NCBI) [1988]

Human Genome Project Launched [1990]

Entrez [1991]

(NCBI's text search tool)

GenBank project

transitioned to NCBI [1992]

First bacterial genome sequenced

Haemophilus influenzae [1995]

Saccharomyces cerevisiae [1996]

PubMed (Public Medline) Launched [1997]

Caenorhabditis elegans genome [1998]

First human chromosome, chr22 [1999]

GEO and dbSNP [1999]

Drosophila melanogaster shotgun genome [2000]

Homo sapiens genome draft [2001]

Open Access Journals [PLoS, 2003]

Project ENCODE launched [2003]

DNA Origami [2006]

Illumina sequencing technology [2006]

Diploid genome of human single individuals [2007]

Synthetic bacterial genomes [2008]

Nanopore sequencing technology [1999/2014]

Synthetic yeast full genome [2017]

Earth BioGenome project [2018]

T2T: first complete, gapless sequence of a human genome [2022/2023]



Larry Page & Sergey Brin [1996]

genscan gene-finding program [1997]

EMBOSS [1998]

European Molecular Biology
Open Software Suite

Genome Browsers

ENSEMBL & UCSC-GB
[2001/2002]

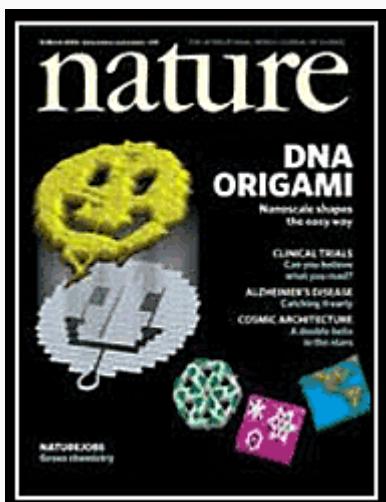
BioPython [2000]

BioConductor [2001]

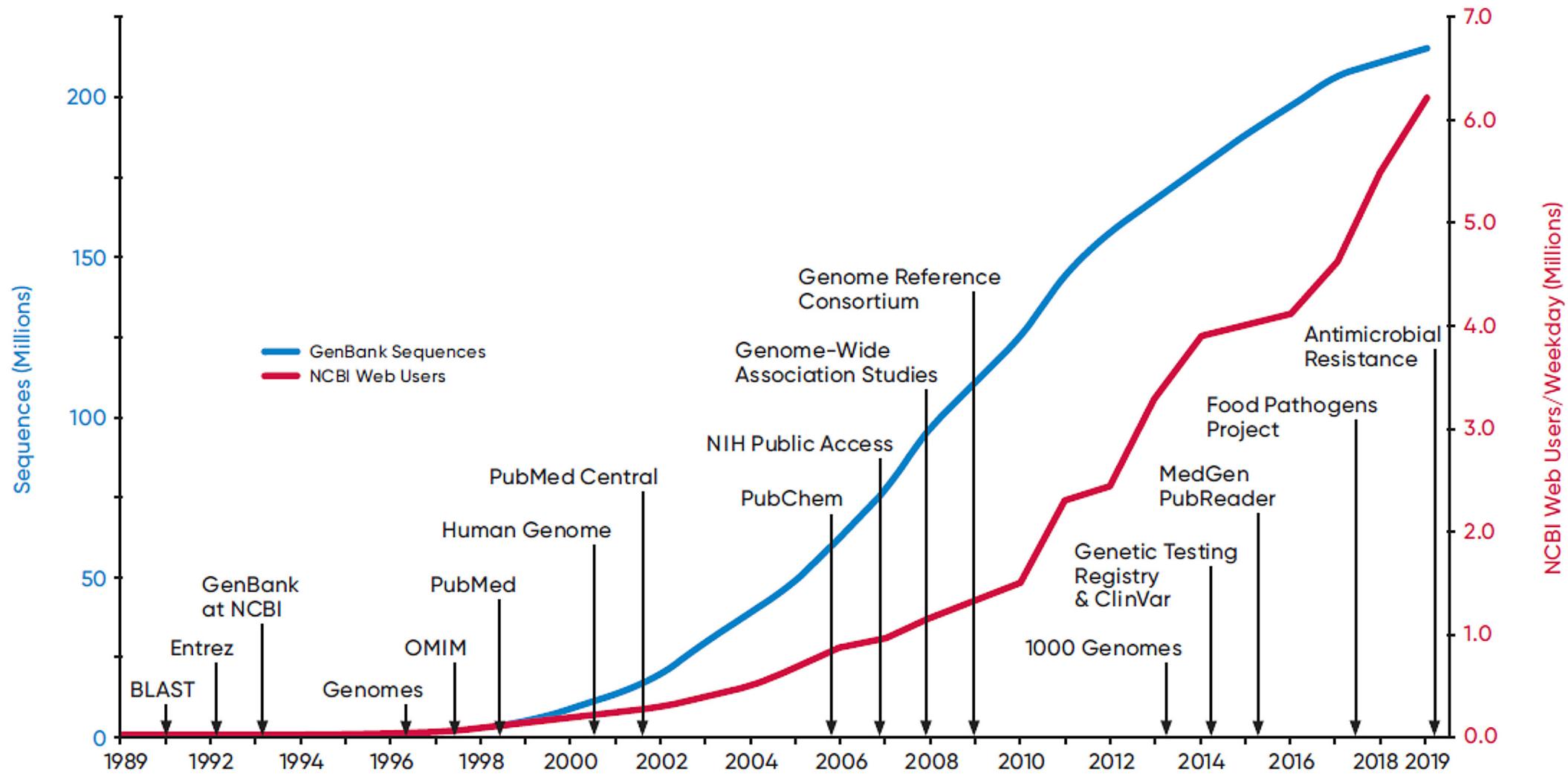
BioPERL [2002]

General AI

Deep Learning
& Big Data
[2012]

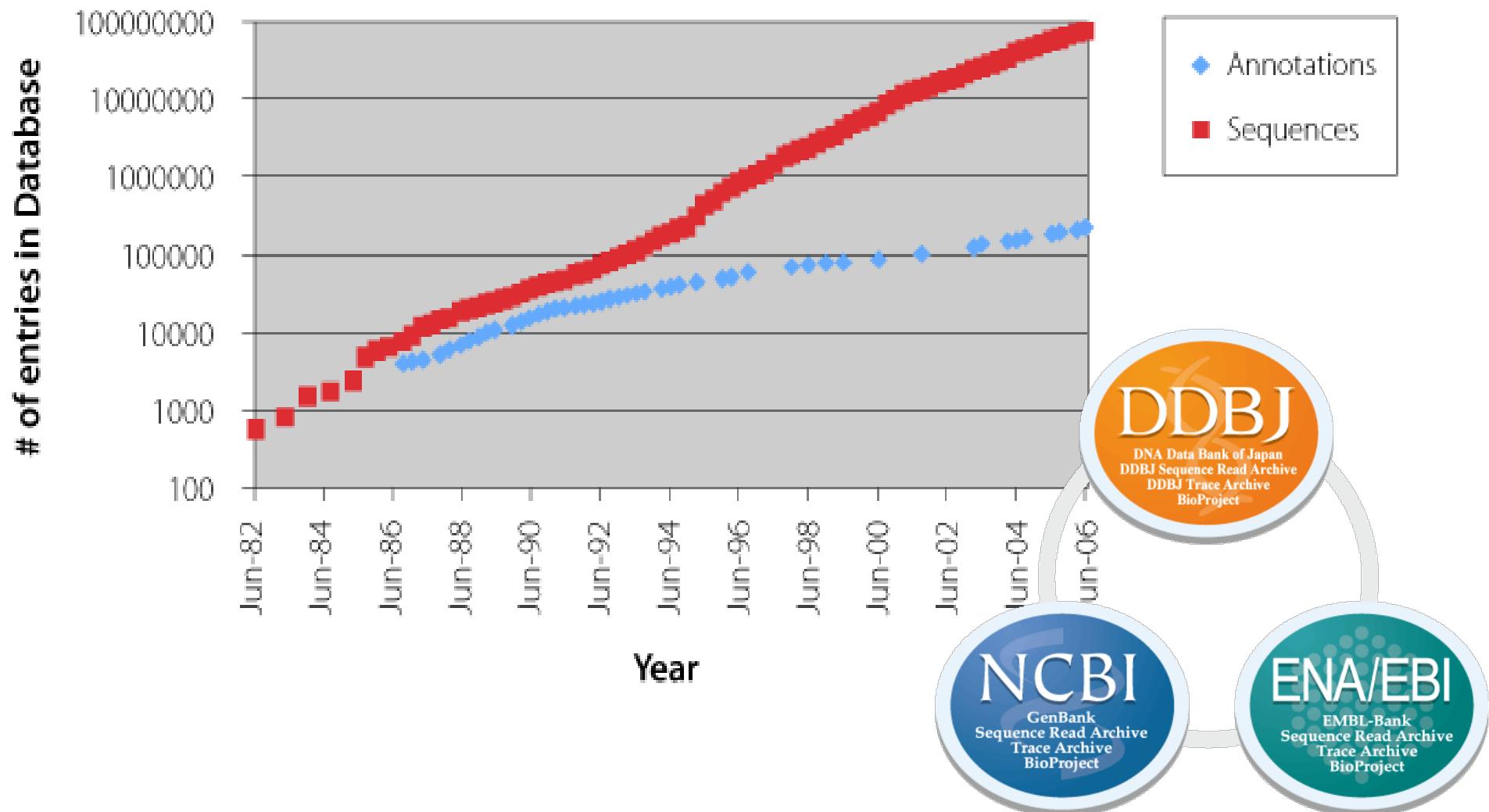


Growth of Sequence Databases



Growth of Sequence Databases Revisited

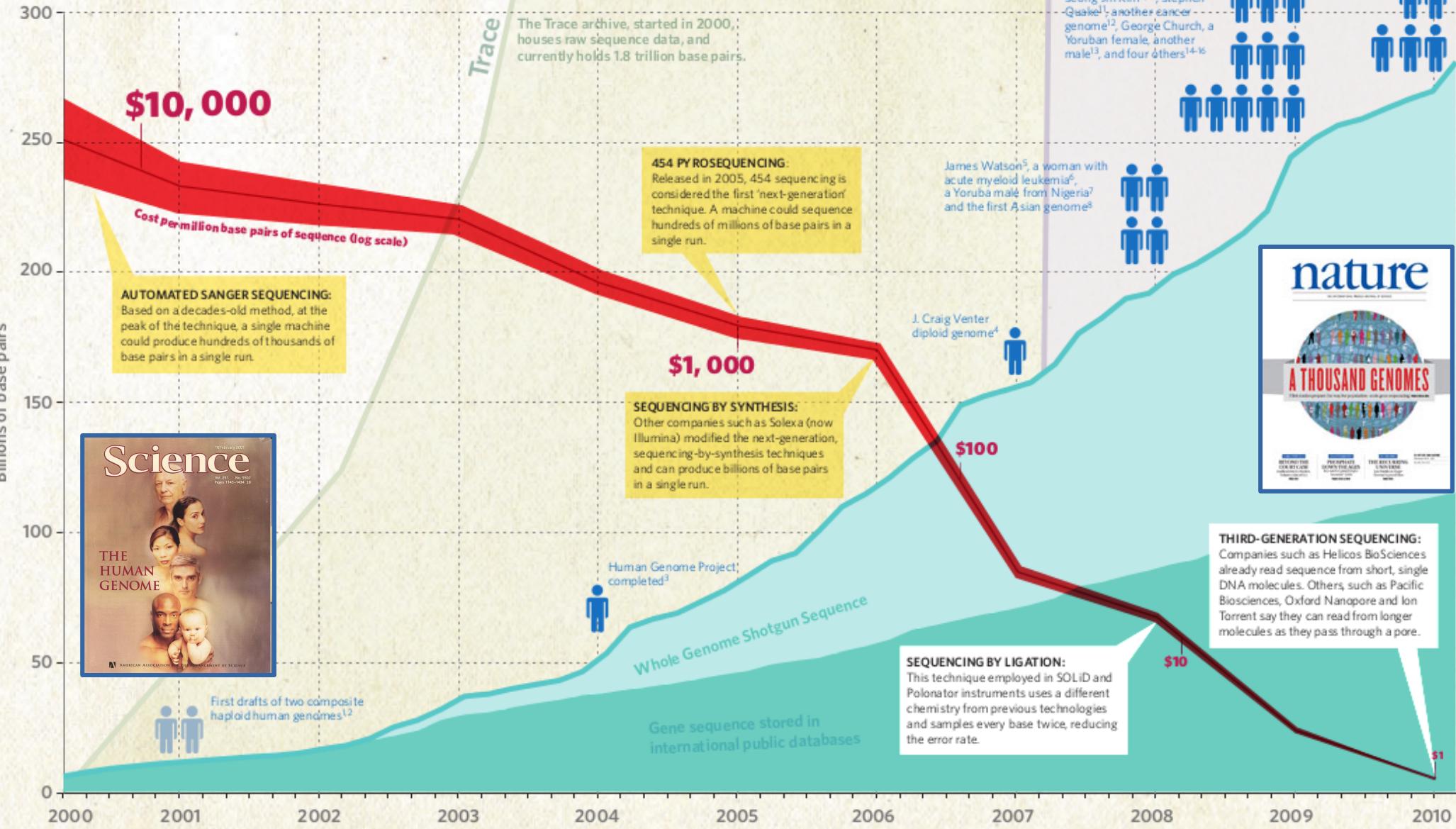
GenBank [nucleotide]



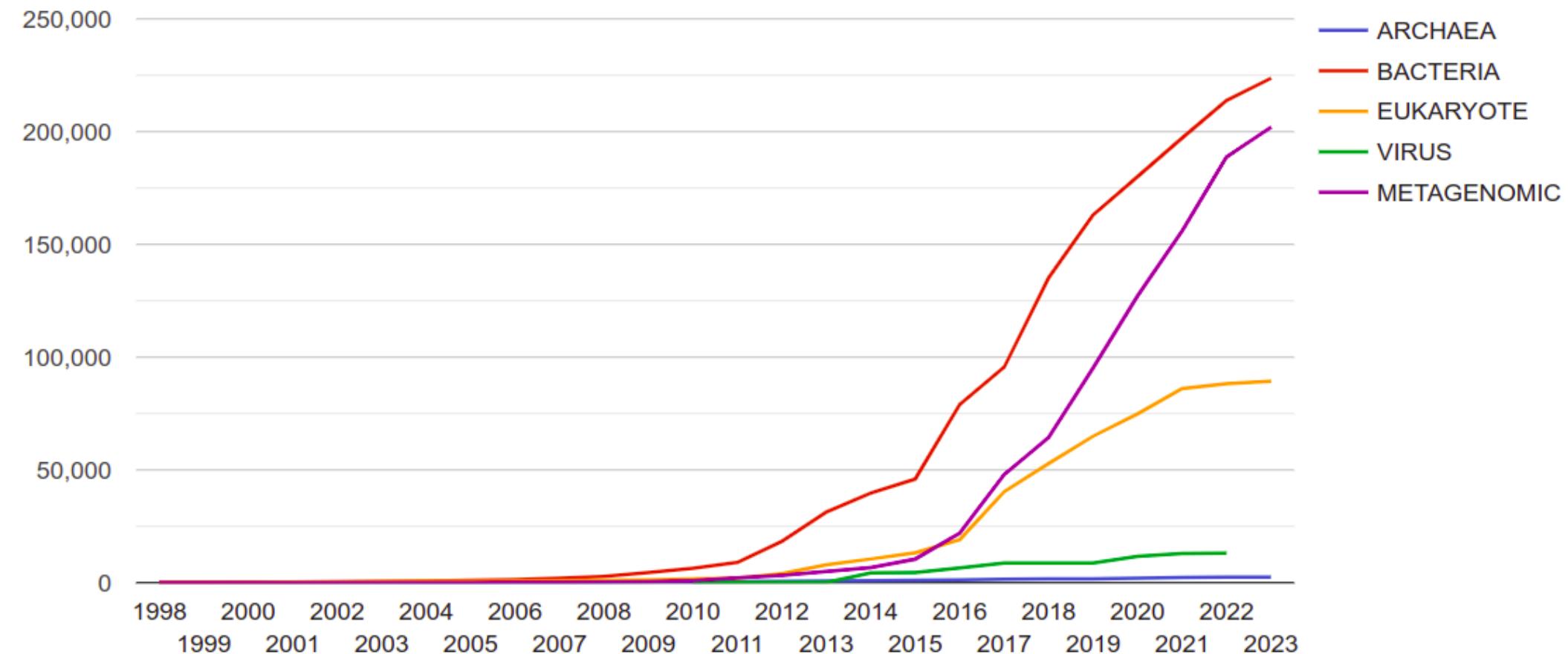
NGS Data Explosion

The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

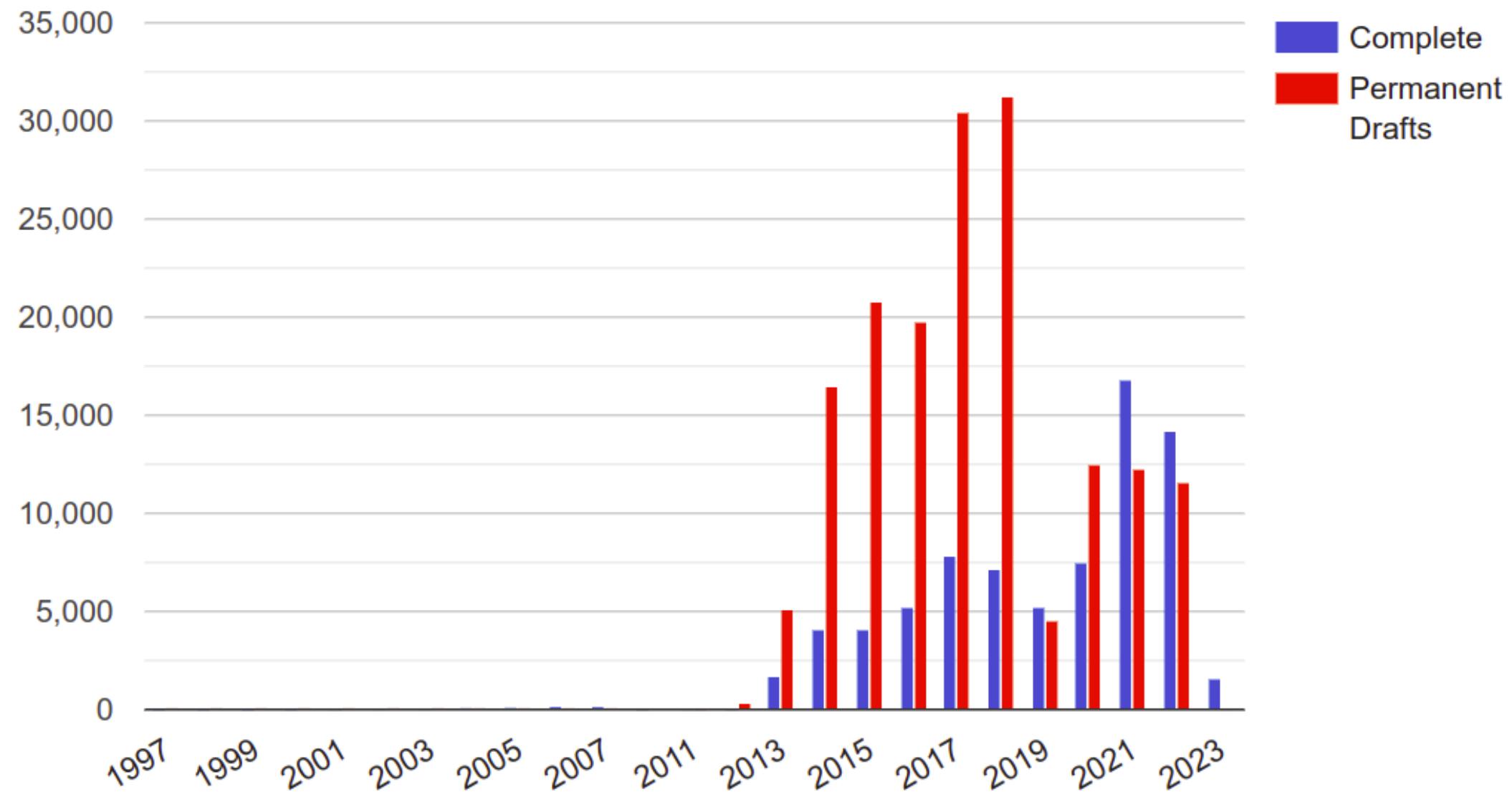
Nature, 464:670-671, 2010.



Sequenced Genomes



Genome Completeness



<https://gold.jgi.doe.gov/statistics>

How much complete a genome is?



PubMed
Central

PubMed Central
biomedical and life sciences journal

finishing human genome

Search for journal titles

Browse PMC journals:

A-B C-H

Receive notice of new journal issues from PMC: join the PMC News
PMC News RSS feed

All the articles in PMC are freely available online (on a rolling basis). Some journals go back many years. Find out what that means.

PMC's utilities include an Online Submission System for the full-text of some article types, PMC search tools for PMC searches and linking to your site, and more ...

Looking for a modern journal? Visit NLM's Journal Publishing

It's about preservation and accessibility. See our complete run of back issues of PMC.

NCBI Resources How To

Science

HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

f t in g m

The complete sequence of a human genome

SERGEY NURK [ID](#), SERGEY KOREN [ID](#), ARANG RHIE [ID](#), MIKKO RAUTIAINEN [ID](#), ANDREY V. BZIKADZE [ID](#), ALLA MIKHEENKO, MITCHELL R. VOLLMER [ID](#), NICOLAS ALTEMOSE [ID](#), LEV URALSKY [ID](#), ARIEL GERSHMAN [ID](#), SERGEY AGANEZOV [ID](#), SAVANNAH J. HOYT [ID](#), MARK DIEKHANS [ID](#), GLENNIS A. LOGSDON [ID](#), MICHAEL ALONGE [ID](#), STYLIANOS E. ANTONARAKIS [ID](#), MATTHEW BORCHERS [ID](#), GERARD G. BOUFFARD [ID](#), SHELISE Y. BROOKS, GINA V. CALDAS, NAE-CHYUN CHEN [ID](#), HAOYU CHENG [ID](#), CHEN-SHAN CHIN [ID](#), WILLIAM CHOW [ID](#), LEONARDO G. DE LIMA [ID](#), PHILIP C. DISHUCK [ID](#), RICHARD DURBIN [ID](#), TATIANA DVORKINA, IAN T. FIDDES [ID](#), GIULIO FORMENTI [ID](#), ROBERT S. FULTON, ARKARACHAI FUNGTAMMASAN [ID](#), ERIK GARRISON [ID](#), PATRICK G. S. GRADY [ID](#), TINA A. GRAVES-LINDSAY [ID](#), IRA M. HALL [ID](#), NANCY F. HANSEN [ID](#), GABRIELLE A. HARTLEY, MARINA HAUKNES [ID](#), KERSTIN HOWE [ID](#), MICHAEL W. HUNKAPILLER, CHIRAG JAIN, MITEN JAIN [ID](#), ERICH D. JARVIS [ID](#), PETER KERPEDJIEV, MELANIE KIRSCH [ID](#), MIKHAIL KOLMOGOROV [ID](#), JONAS KORLACH [ID](#), MILINN KREMITZKI [ID](#), HENG LI [ID](#), VALERIE V. MADURO [ID](#), TOBIAS MARSCHALL [ID](#), ANN M. MCCARTNEY, JENNIFER McDANIEL [ID](#), DANNY E. MILLER [ID](#), JAMES C. MULLIKIN [ID](#), EUGENE W. MYERS [ID](#), NATHAN D. OLSON [ID](#), BENEDICT PATEN [ID](#), PAUL PELUSO, PAVEL A. PEVZNER [ID](#), DAVID PORUBSKY [ID](#), TAMARA POTAPOVA [ID](#), EVGENY I. ROGAEV, JEFFREY A. ROSENFELD [ID](#), STEVEN L. SALZBERG [ID](#), VALERIE A. SCHNEIDER, FRITZ J. SEDLAZECK [ID](#), KISHWAR SHAFIN [ID](#), COLIN J. SHEW, ALAINA SHUMATE [ID](#), YING SIMS, ARIAN F. A. SMIT [ID](#), DANIELA C. SOTO [ID](#), IVAN SOVIĆ [ID](#), JESSICA M. STORER [ID](#), AARON STREETS [ID](#), BETH A. SULLIVAN [ID](#), FRANÇOISE THIBAUD-NISSEN [ID](#), JAMES TORRANCE [ID](#), JUSTIN WAGNER, BRIAN P. WALENZ [ID](#), AARON WENGER [ID](#), JONATHAN M. D. WOOD [ID](#), CHUNLIN XIAO [ID](#), STEPHANIE M. YAN [ID](#), ALICE C. YOUNG [ID](#), SAMANTHA ZARATE [ID](#), Urvashi SURTI, RAJIV C. MCCOY [ID](#), MEGAN Y. DENNIS [ID](#), IVAN A. ALEXANDROV [ID](#), JENNIFER L. GERTON [ID](#), RACHEL J. O'NEILL [ID](#), WINSTON TIMP [ID](#), JUSTIN M. ZOOK [ID](#), MICHAEL C. SCHATZ [ID](#), EVAN E. EICHLER [ID](#), KAREN H. MIGA [ID](#), AND ADAM M. PHILLIPPI [ID](#)

[fewer](#) [Authors Info & Affiliations](#)

SCIENCE • 31 Mar 2022 • Vol 376, Issue 6588 • pp. 44-53 • DOI: 10.1126/science.abj6987

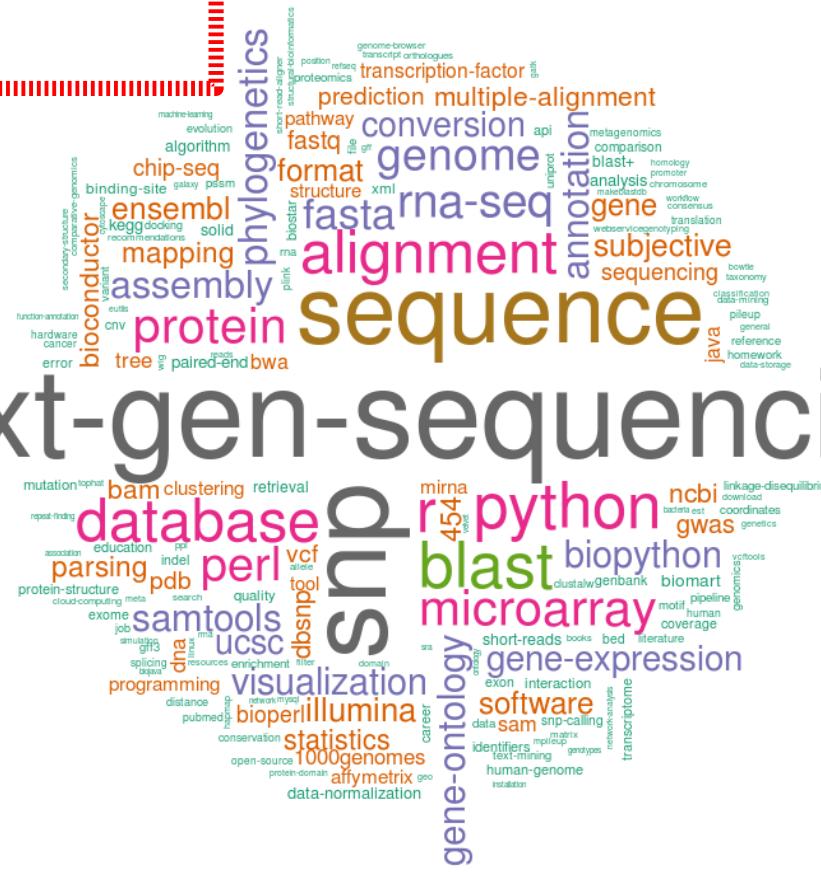
Nature 585, 79–84 (2020) | Cite this article

Privacy Policy | Disclaimer | Accessibility | Help | NCBI Home

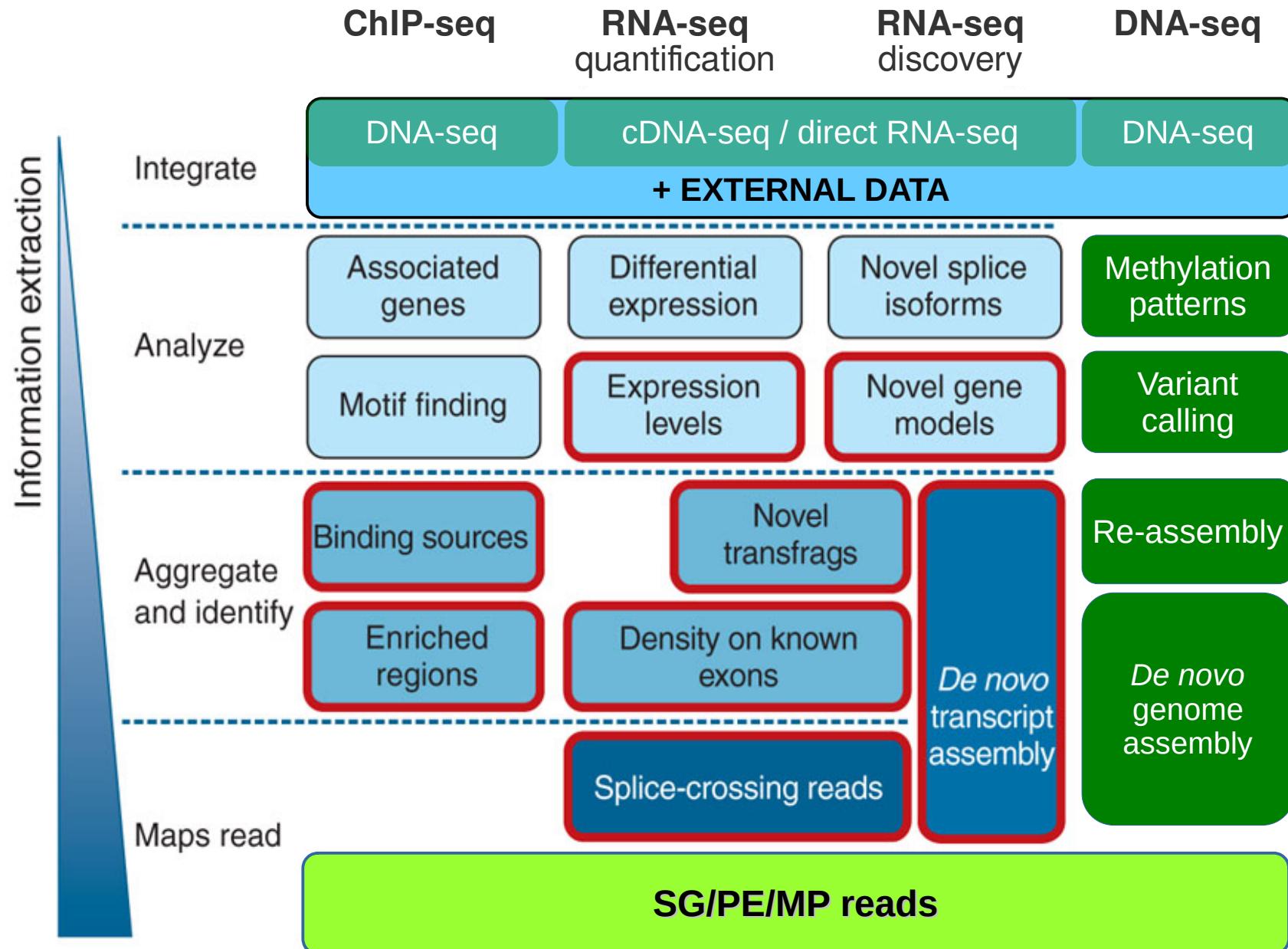
<http://www.ncbi.nlm.nih.gov/pmc/>

SEQUENCING OVERVIEW

next-gen-sequencing



Seqs, Seqs & more Seqs

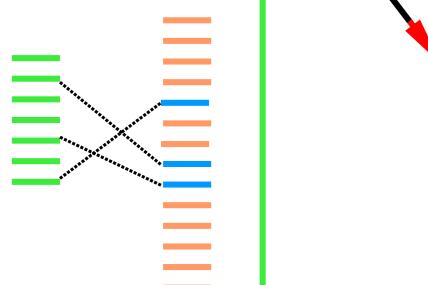


Assembly vs Homology

SUBJECTs



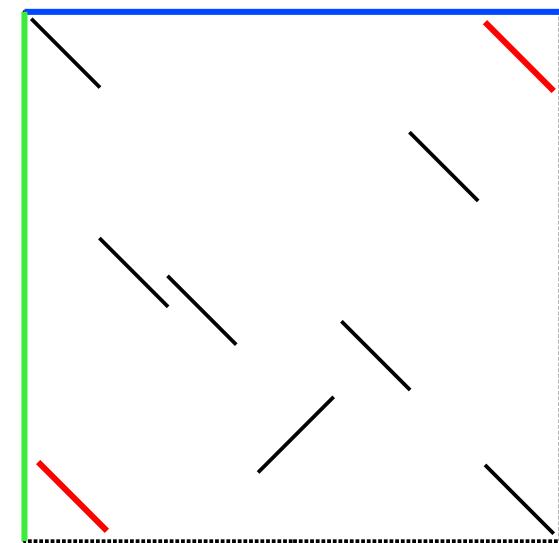
QUERY —



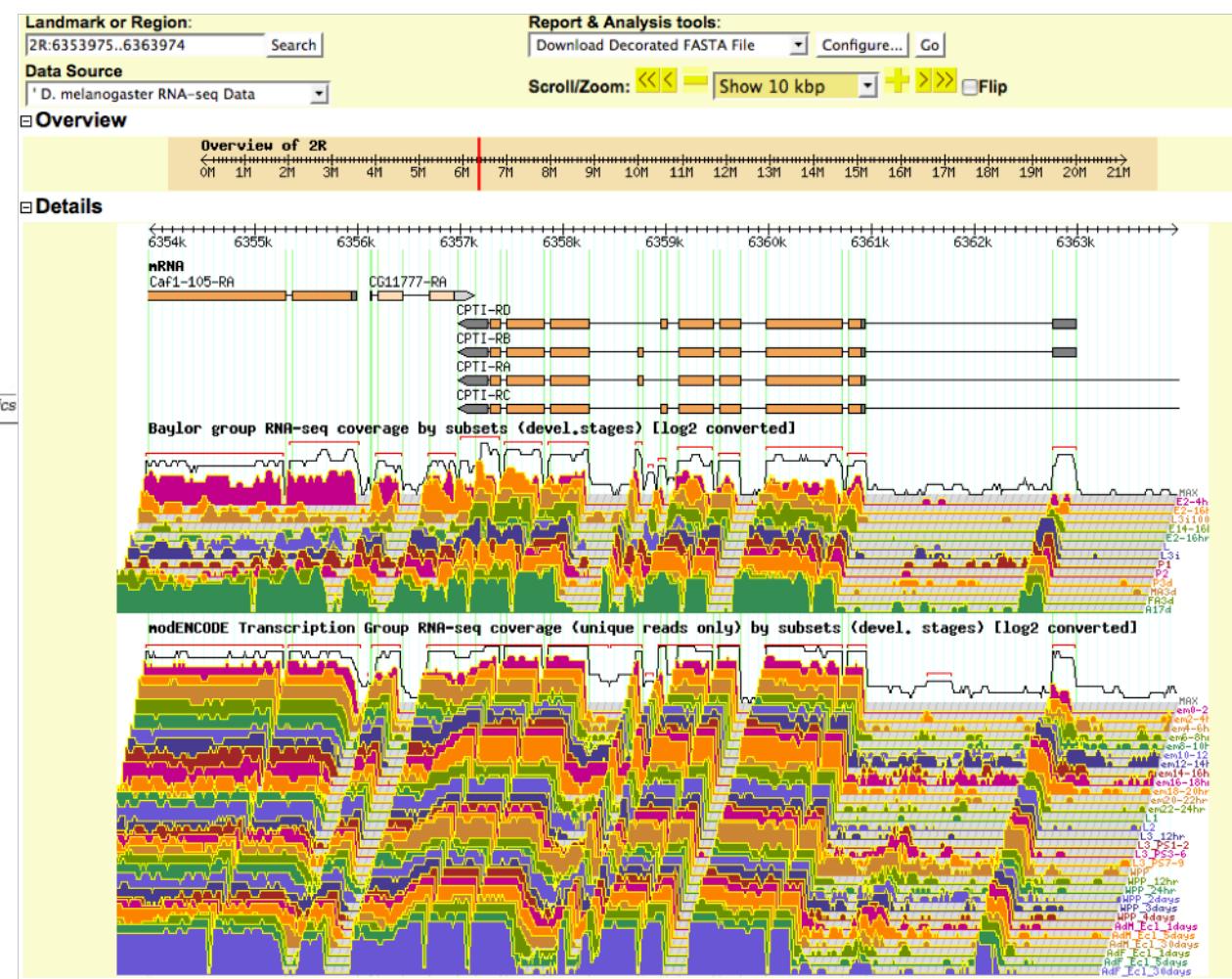
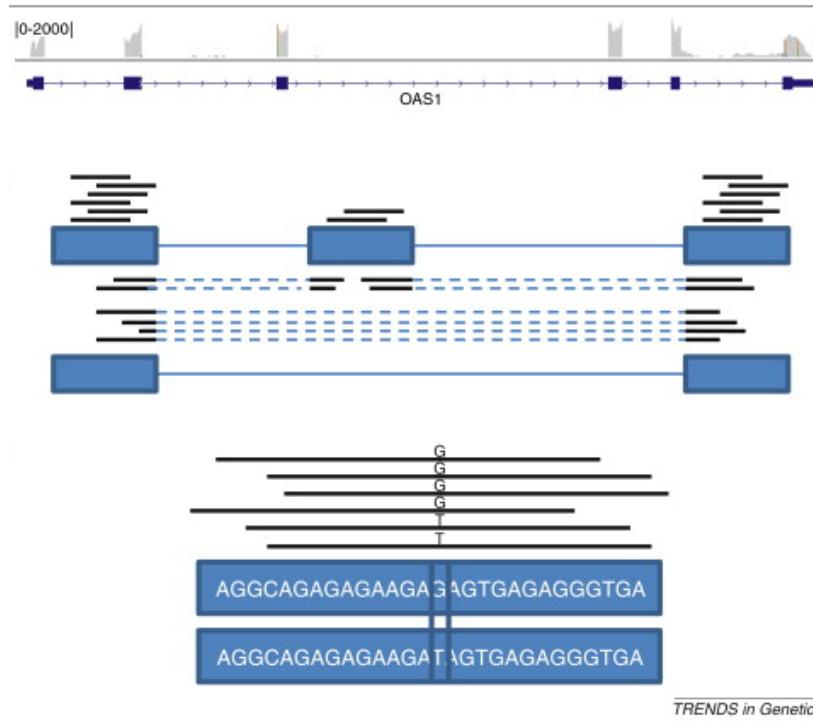
TRACES



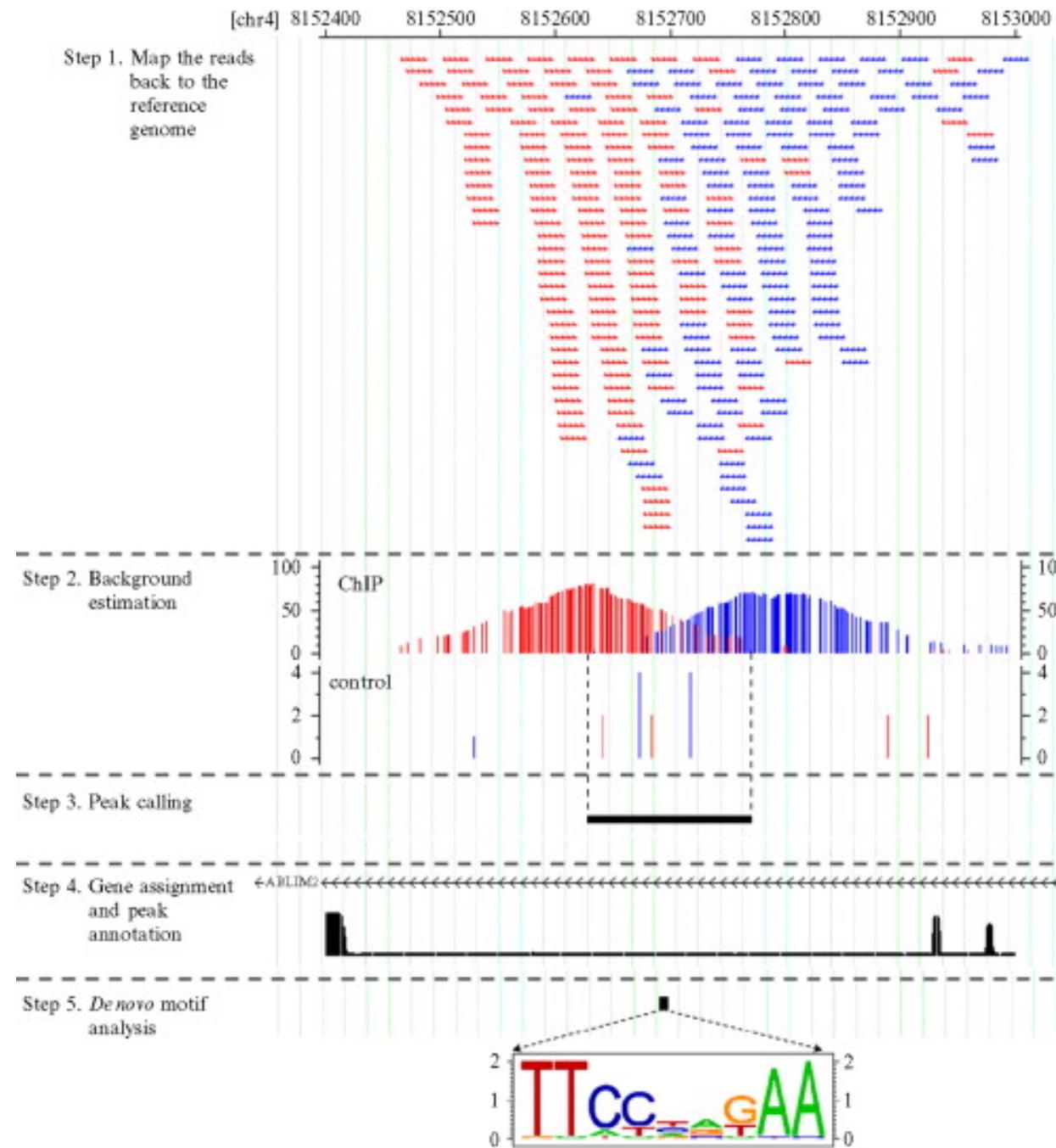
ALL against ALL



RNA-Seq Overview

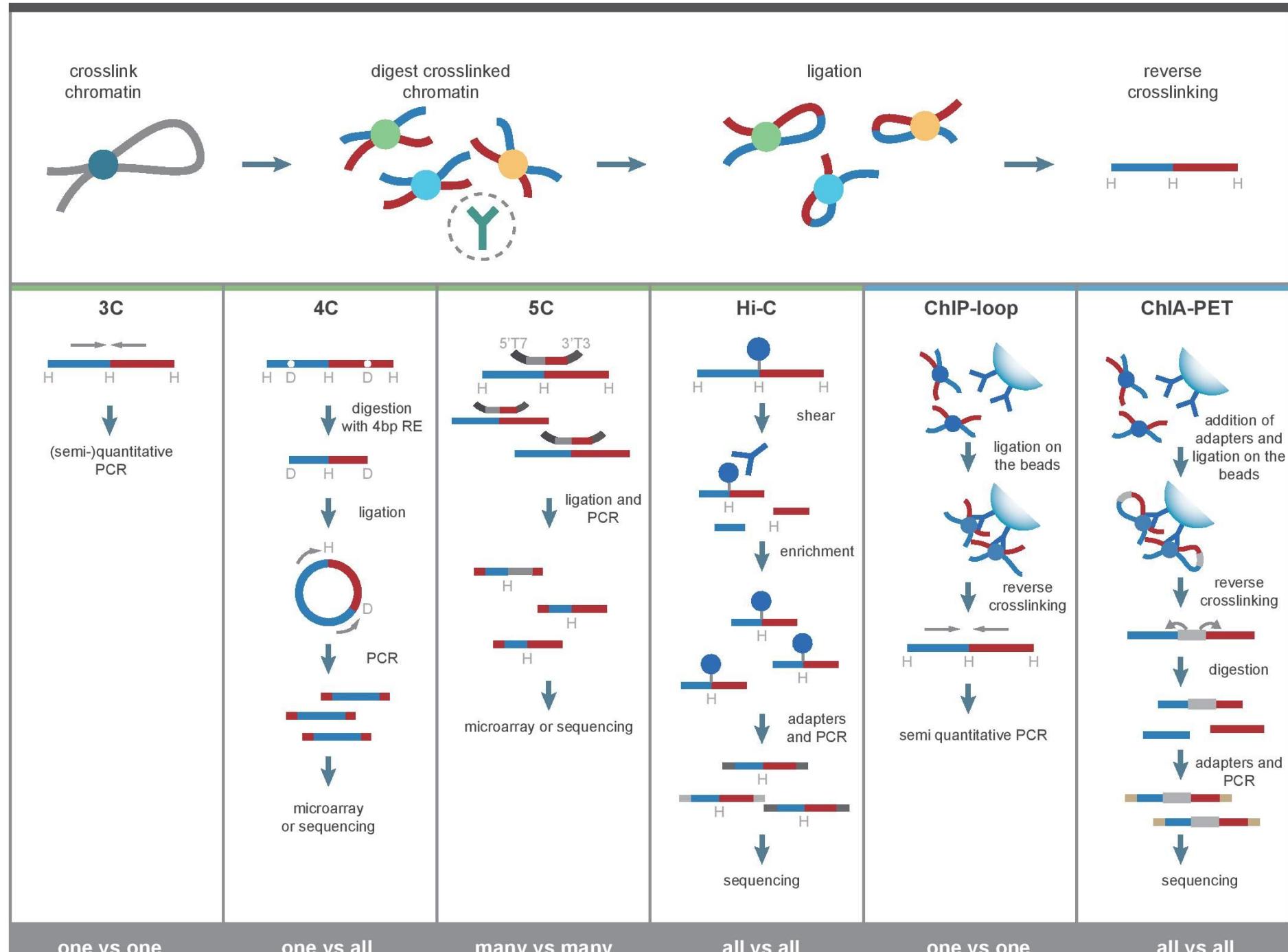


From Reads to Peaks to Signals

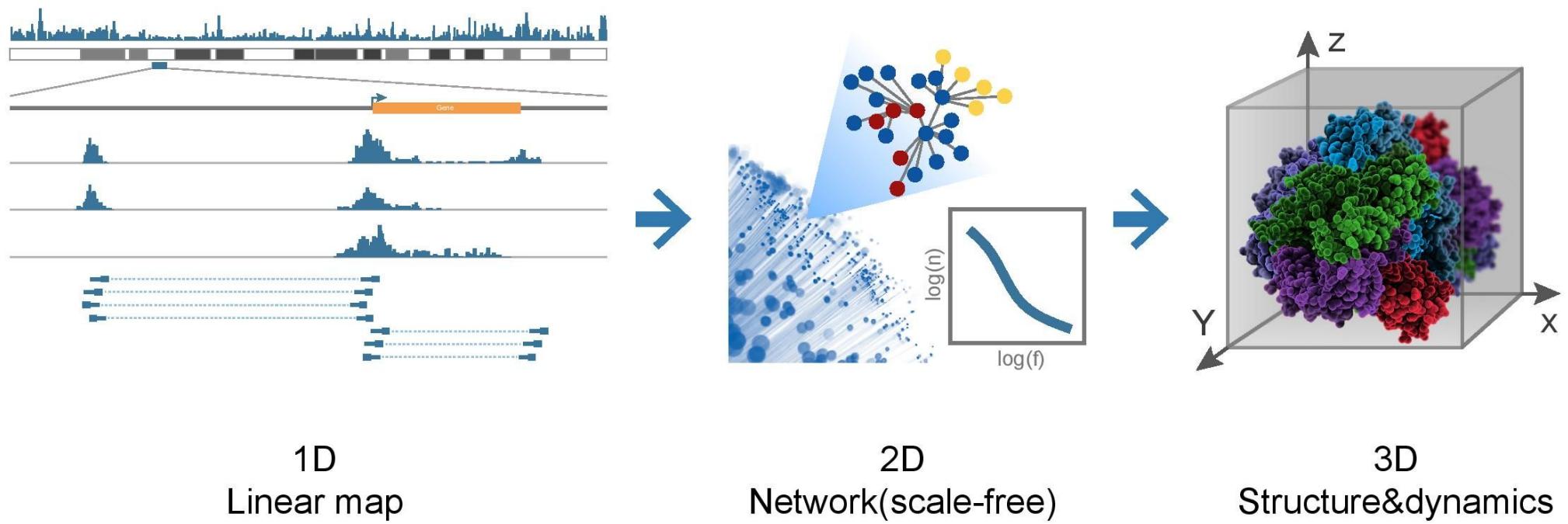


Chromatin Conformation Overview

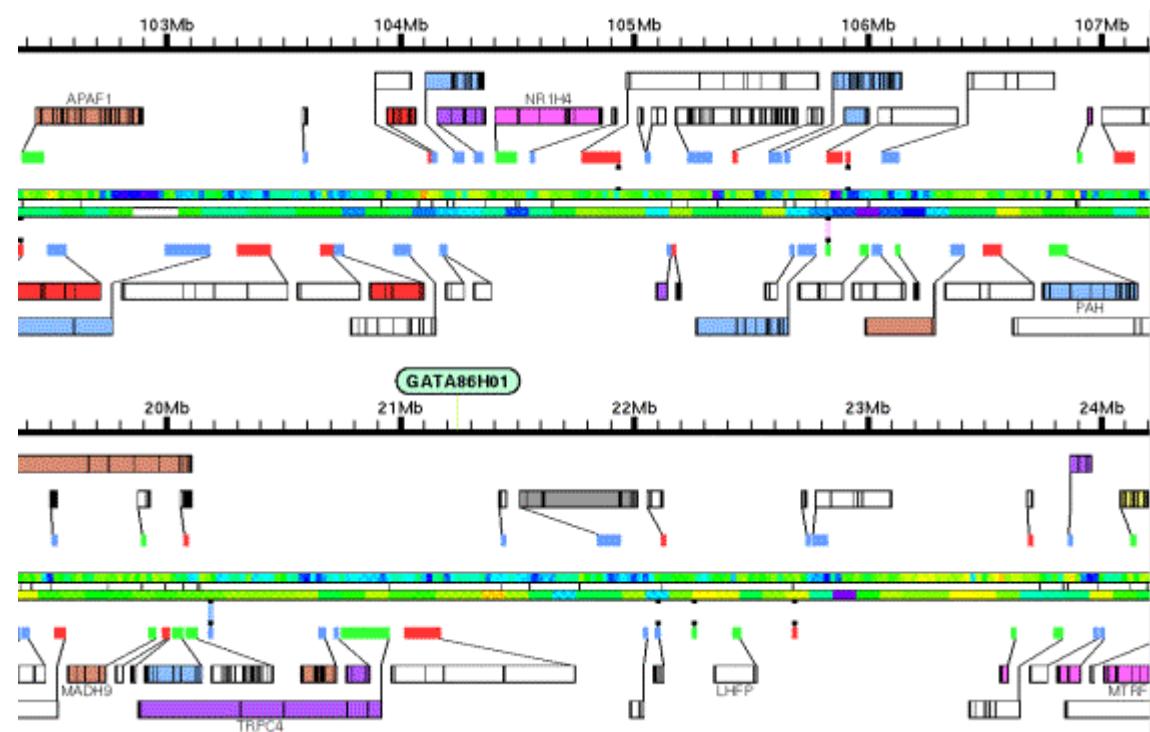
Li et al *BMC Genomics* 2014



Chromatin Conformation Over Genome



SEQUENCE ANNOTATION



Annotating Sequences: Does it matters ?



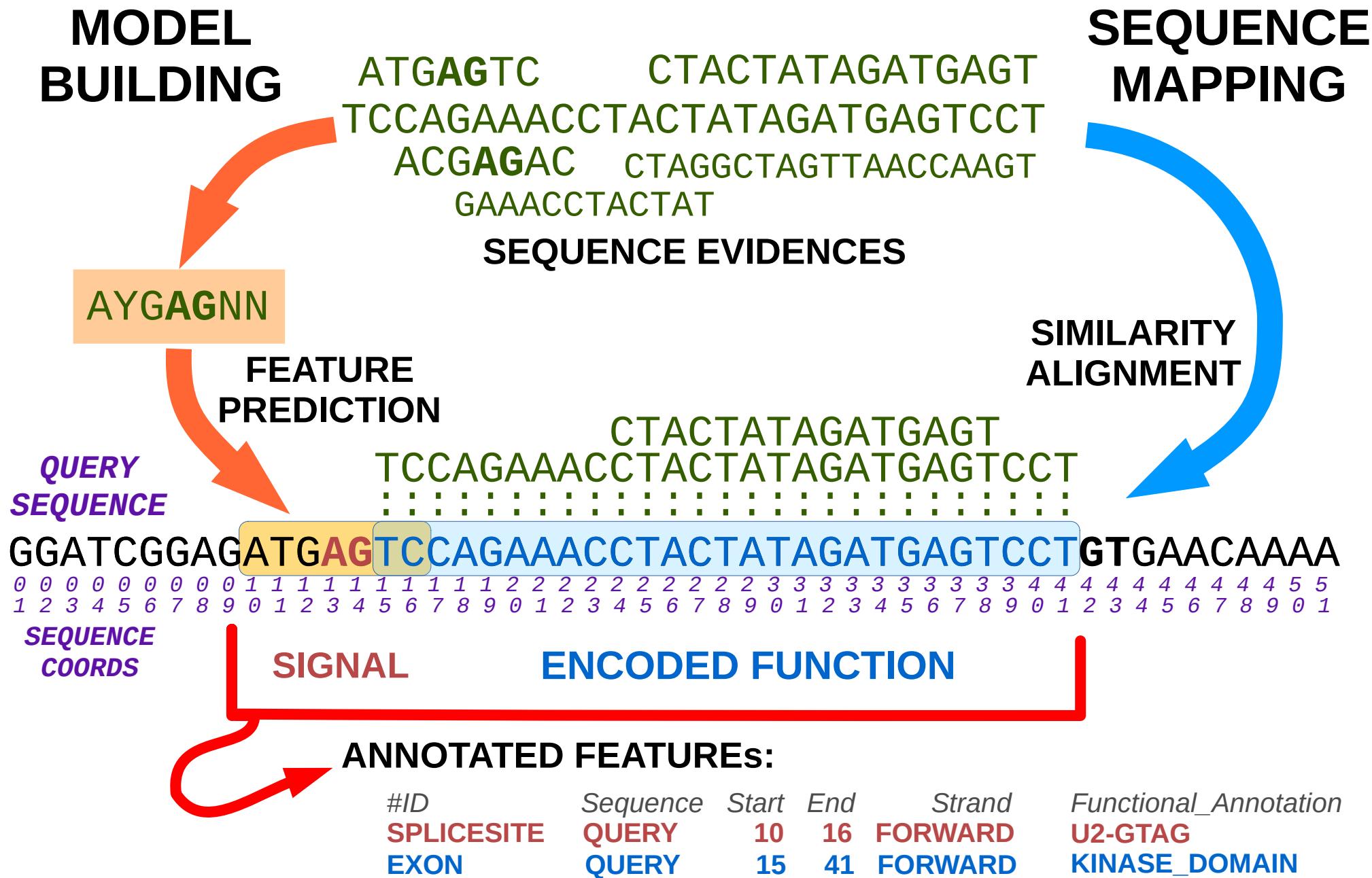
"The genome sequence of an organism is an information resource unlike any that biologists have previously had access to.

But the **value** of the **genome** is only as good as its **annotation**.

It is the annotation that bridges the gap from the sequence to the biology of the organism."

Stein L. *Nat Rev Genet*, 2(7):493-503, 2001

Sequence Annotation: What Does It Mean?



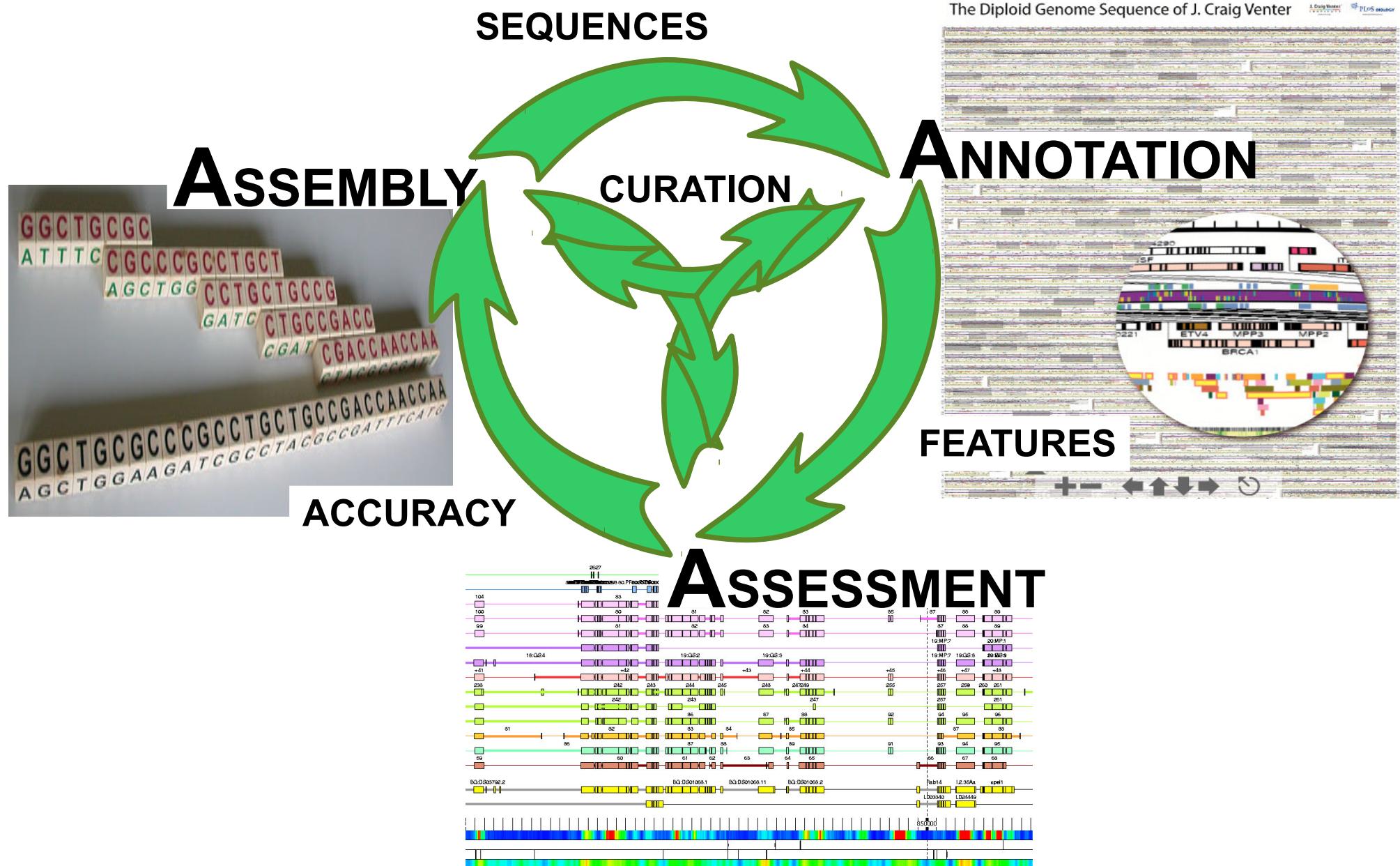
ANNOTATING SEQUENCES



chr2	refseq	mRNA	1	13027	.	+	.	transcript_id	"UGT1A1"
chr2	refseq	exon	1	879	.	+	.	transcript_id	"UGT1A1"
chr2	refseq	start_codon	16	18	.	+	.	transcript_id	"UGT1A1"
chr2	refseq	CDS	16	879	.	+	0	transcript_id	"UGT1A1"
chr2	refseq	exon	6762	6893	.	+	.	transcript_id	"UGT1A1"
chr2	refseq	CDS	6762	6893	.	+	1	transcript_id	"UGT1A1"
chr2	refseq	exon	7577	7664	.	+	.	transcript_id	"UGT1A1"
chr2	refseq	CDS	7577	7664	.	+	1	transcript_id	"UGT1A1"
chr2	refseq	exon	7948	8167	.	+	.	transcript_id	"UGT1A1"
chr2	refseq	CDS	7948	8167	.	+	2	transcript_id	"UGT1A1"
chr2	refseq	exon	11990	13027	.	+	.	transcript_id	"UGT1A1"
chr2	refseq	CDS	11990	12287	.	+	0	transcript_id	"UGT1A1"
chr2	refseq	stop_codon	12285	12287	.	+	.	transcript_id	"UGT1A1"



The Annotation Problem



Assembly vs Annotation VERSIONS

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

Login · Register

Human (GRCh37) ▾



Human assembly and gene annotation

Assembly

This site provides a data set based on the February 2009 *Homo sapiens* high coverage assembly GRCh37 from the [Genome Reference Consortium](#). This assembly is used by UCSC to create their hg19 database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

It also includes nine [haplotypic regions](#), mainly in the MHC region of chromosome 6.

[Watch a video on YouTube](#) about patches and haplotypes in the Human genome.

Gene annotation

The Ensembl human gene annotations have been updated using Ensembl's pipeline. The updated annotation incorporates new protein and cDNA sequences publicly available since the last GRCh37 genebuild (March 2009).

In release 70 (January 2013), we continue to display a joint gene set based on automatic annotation from Ensembl and the manually curated annotation from Gencode. The Consensus Coding Sequence set corresponds to [GENCODE](#) release 15. The Consensus Coding Sequence has been mapped to the annotations. More information about the [CCDS project](#).

Updated manual annotation from Havana is merged into the Ensembl annotation. Transcripts from the two annotation sources are merged if they share the same boundaries (i.e. have identical splicing pattern) with slight differences in the transcript coordinates. Importantly, all Havana transcripts are included in the final Ensembl/Havana gene set. In this release, 23532 Ensembl gene models and 49095 Havana genes were merged to create the final set of 60620 genes.

- [Detailed information on genebuild \(PDF\)](#)

Statistics

Summary

Gene counts (Primary assembly)

Coding genes:	20,848
Non coding genes:	22,486
Pseudogenes:	13,430
Gene transcripts:	195,409

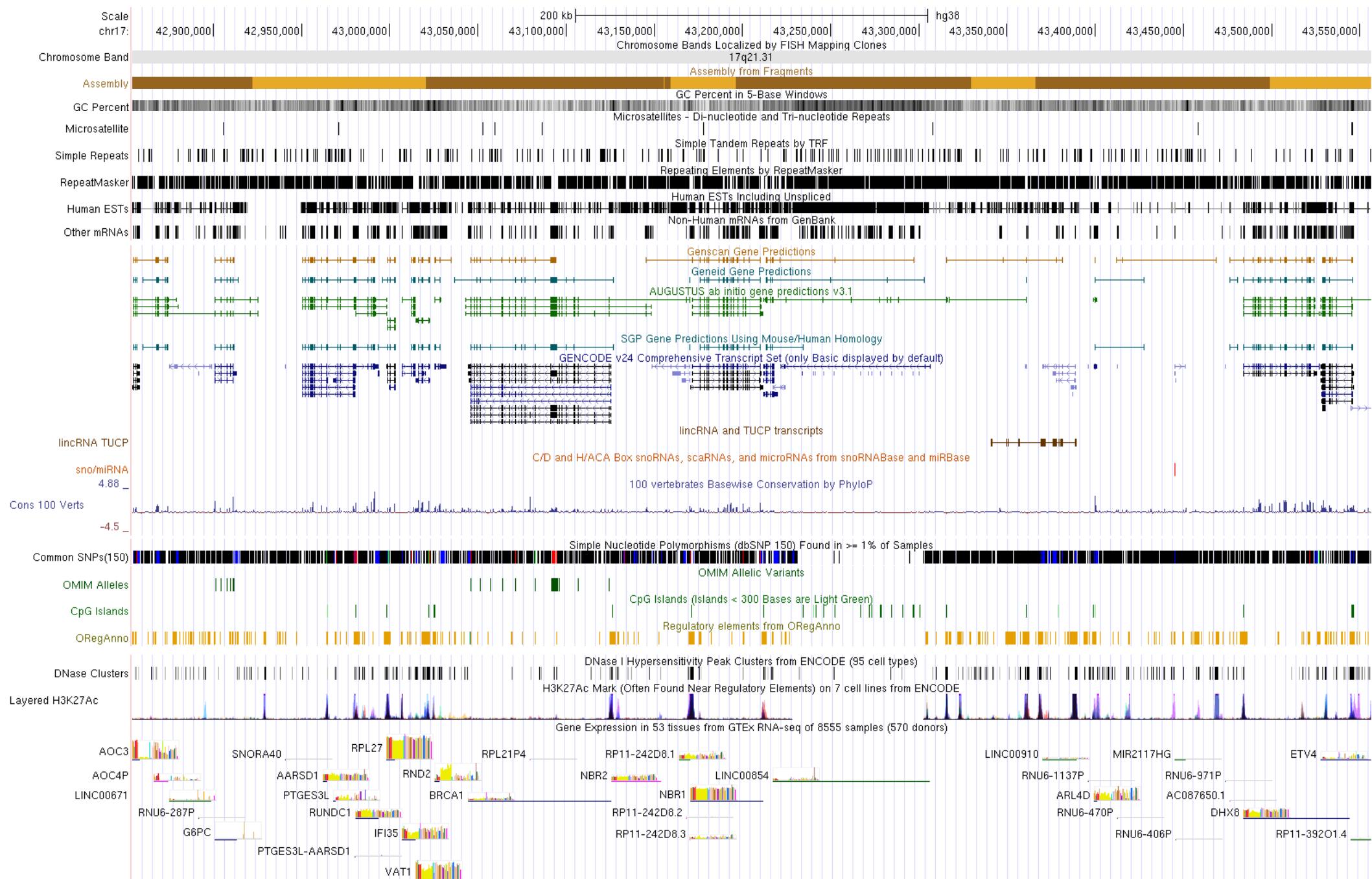
Gene counts (Alternate sequences)

Coding genes:	2,442
Non coding genes:	1,325
Pseudogenes:	1,538
Gene transcripts:	17,863

Other

Genscan gene predictions:	48,186
Short Variants (SNPs, Indels, somatic mutations):	54,474,297
Structural variants:	10,152,873

Annotation Layers



Gene-Finding vs Gene Annotation

Gene prediction
(SNAP)



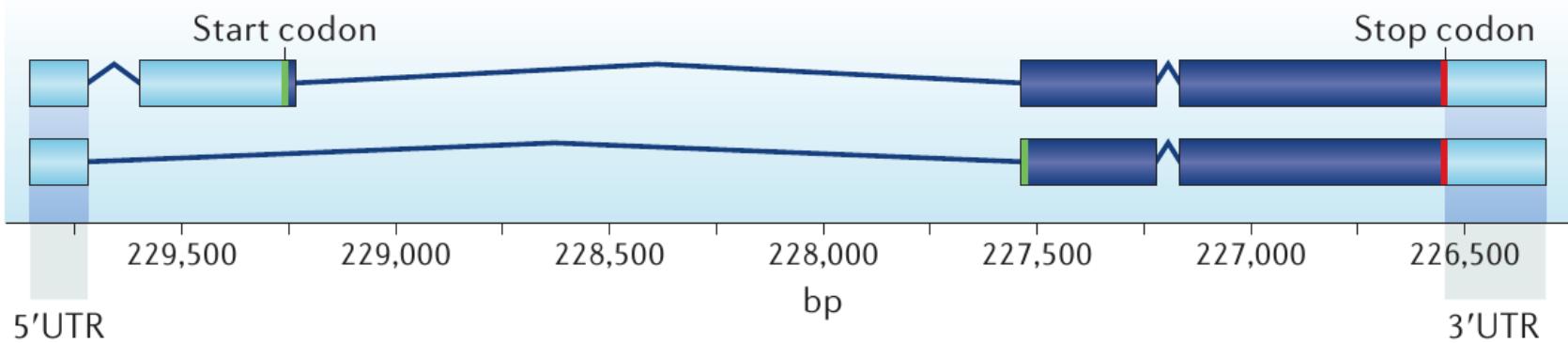
mRNA or EST evidence
(Exonerate)



Protein evidence
(BLASTX)



Gene annotation resulting
from synthesizing all
available evidence
(two alternative splice forms)



Yandell & Ence, *Nat Rev Genet*, 13:329, 2012.

Surfing a Web of Annotations

The image shows a desktop environment with four Mozilla Firefox browser windows open side-by-side:

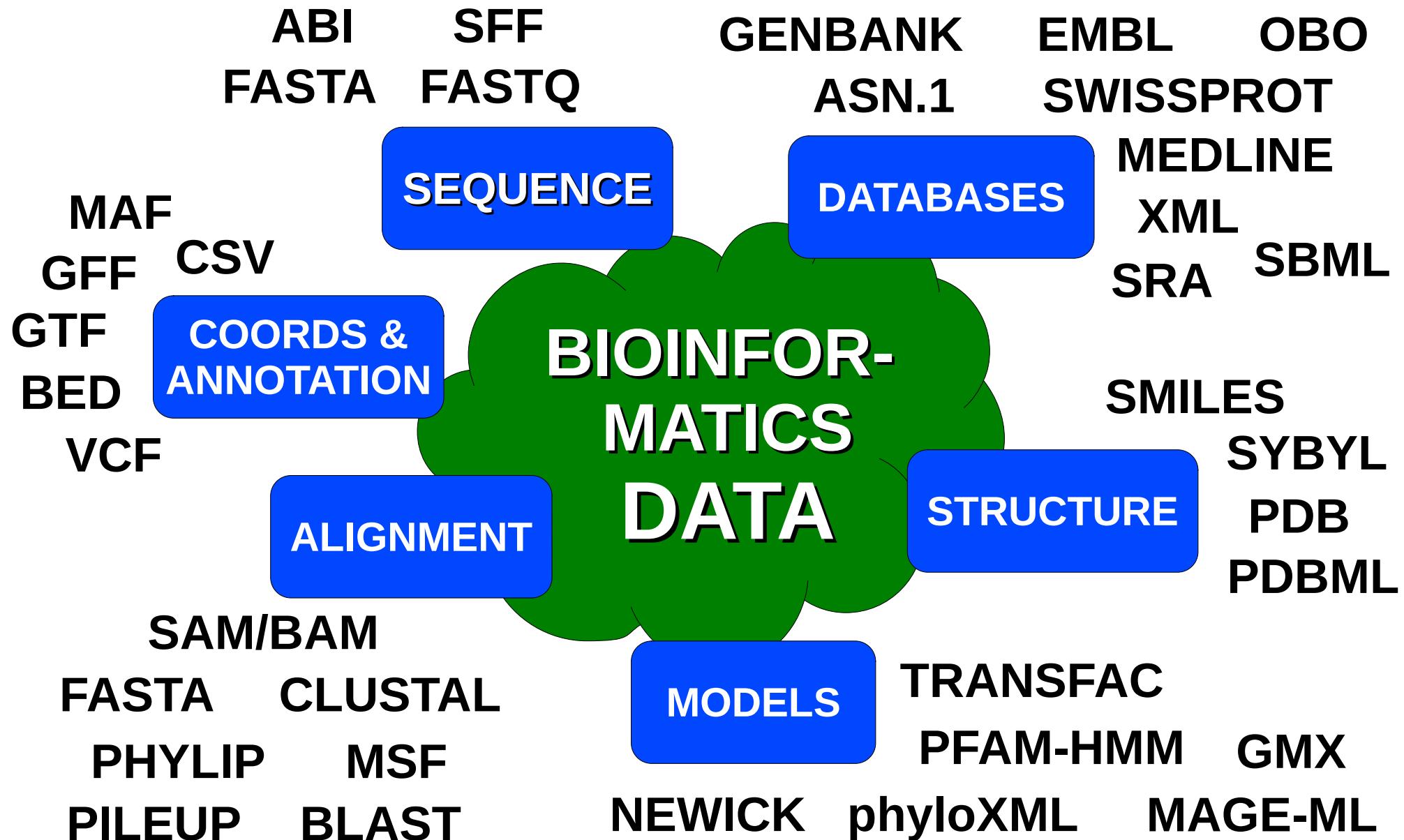
- Entrez cross-database search - Mozilla Firefox**: Shows the Entrez search interface.
- BioMart (MartView) - Mozilla Firefox**: Shows the BioMart search interface.
- Ensembl Genome Browser - Mozilla Firefox**: Shows the Ensembl genome browser interface.
- Pseudogene Analysis of Prokaryotic and Eukaryotic Organisms - Mozilla Firefox**: Shows the Pseudogene.org genome analysis interface.
- GENE PREDICTIONS on GENOMES - Mozilla Firefox**: Shows the IMIM Gene Predictions page.

The Pseudogene.org window is the most prominent, displaying its main content area. The IMIM Gene Predictions window is also clearly visible, showing its header and some descriptive text about gene predictions.

DATA FORMATS

TTTATACTGATAAATTTAACTGCTATAGCTTGGAGTTAGAGATAAGTACAAAAAT
AAAATCCATGACAAATGAAAACATCAAACATTCCAAAATTGAAAAGTCTAAAATAACCAA
ATTATGTCTAACGCCAAAAAGGAAATGATATCTTTCAAGAACATATAAGCGGC
TGAAGAACATTTGAAAGCAAATTAGCTAAAATTGCTAAAACAAGAGGGAAAAAAATGA
ATTAATGAGAGCATTGTTGTCAATGAAATTGTGGAAGTTAAAATGACAAGAGTTGA
AAAATAGAAAATGGACTAATGAAGCGACTGGTATTGAAGCTGACGCTTGAATTACAGT

A Plethora of Bioinformatics Formats



TABULAR vs MULTILINE SEQUENCE RECORDS



Sequence Ontology (SO)

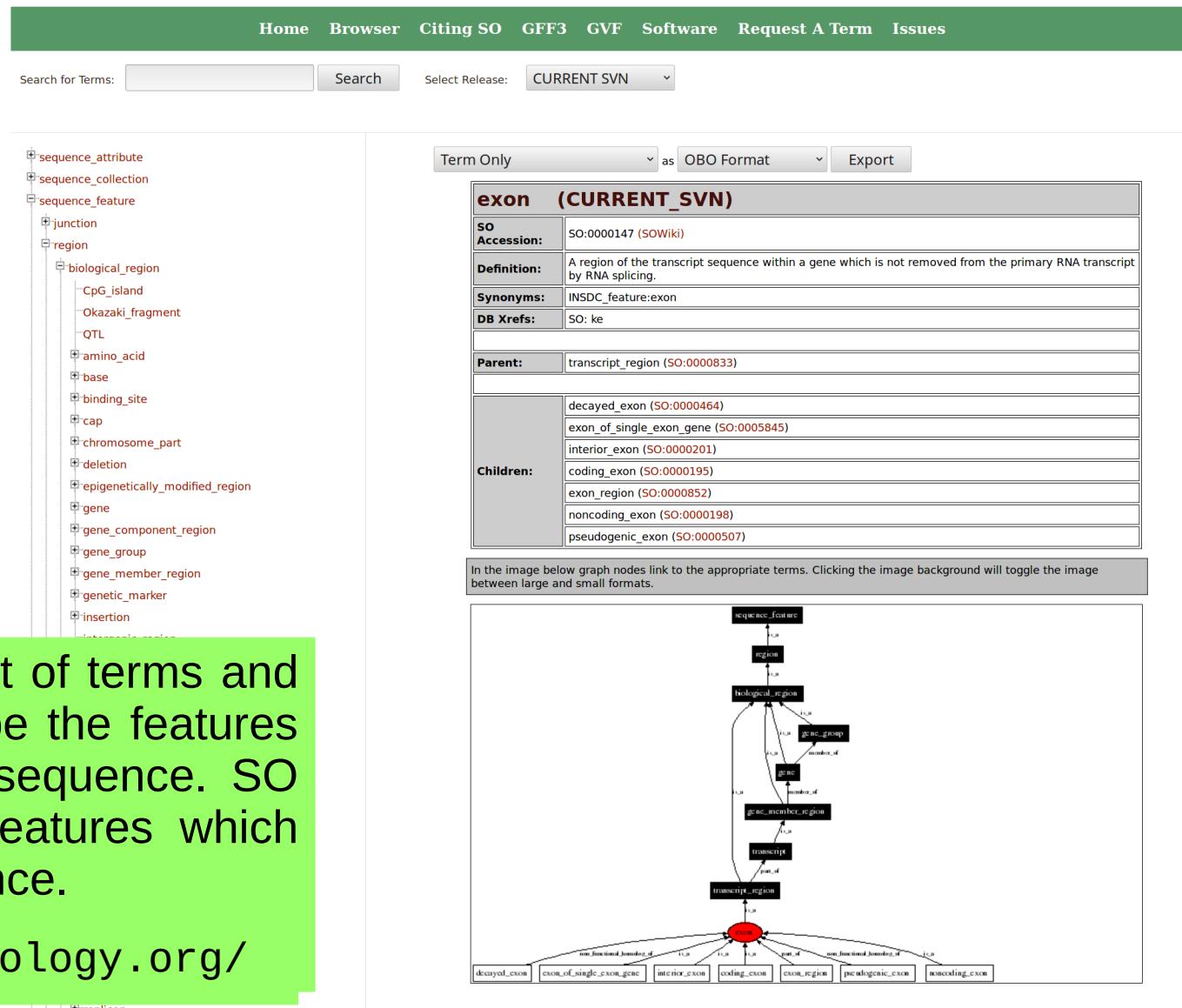
DEFINE and DESCRIBE DIFERENCES AMONG ANNOTATION FEATURES

GENOMIC
cDNA
EST
GENE
primary TRANSCRIPT
mRNA
EXON
UTR
CDS
ORF

1

Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence.

<http://www.sequenceontology.org/>



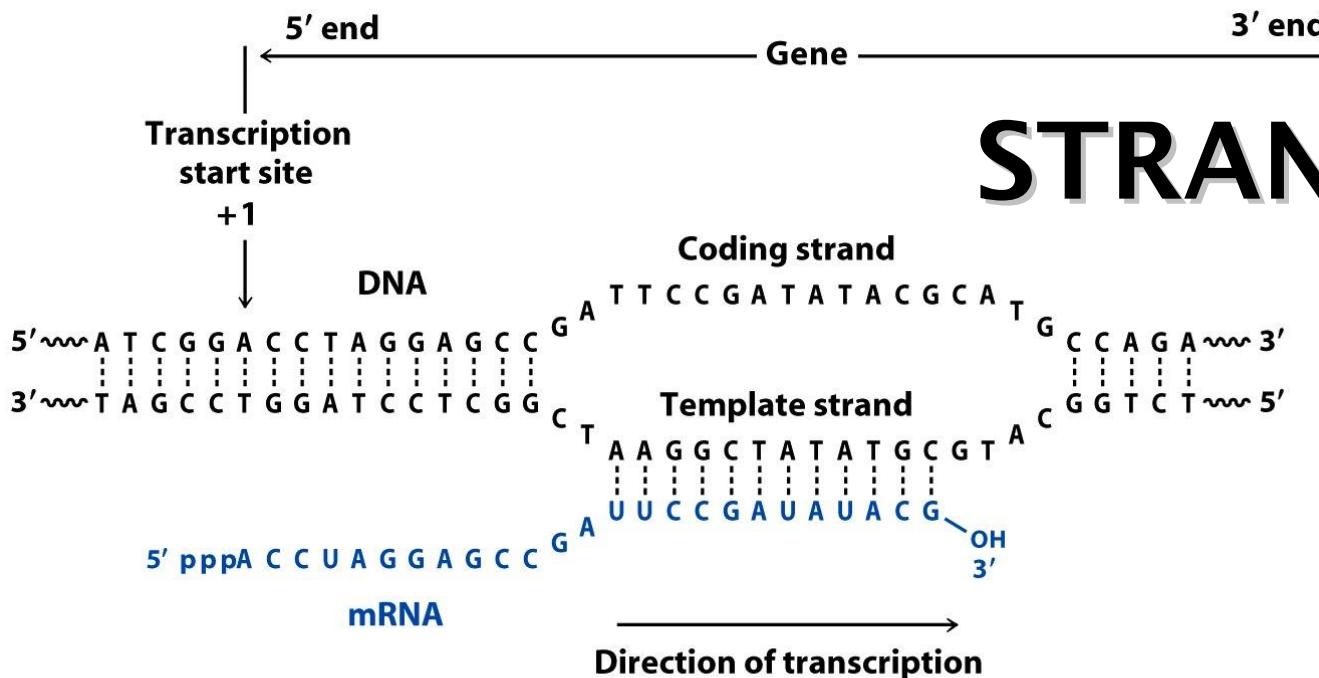


Figure 21-5 Principles of Biochemistry, 4/e
© 2006 Pearson Prentice Hall, Inc.

SEQUENCE

Forward [5' → 3']

TGTGAATACAAGGCAGAACCTAGACCATGTTAAAGAATCAAAAACAGGTT

FORWARD
[+]

Complement

BASE PAIR COMPLEMENTARITY

ACACTTATGTTCCGTCTGGATCTGGTACAATTCTTAGTTTGTC

Reverse

Reverse- Complement

[Forward 3' → 5'] AACCTGTTTGATTCTTAACATGGTCTAGGTTCTGCCTGTATT

ANNOTATIONS

FORWARD
[+]

ANNOTATION FORMATS

RAW COORDS

```
>NM_005101 chr1  
88946 89980 mRNA  
88946 89020 utr  
89021 89023 cds  
89431 89925 cds  
89926 89980 utr  
  
<NM_021170 chr1  
74412 75537 mRNA  
74412 74505 utr  
74506 74879 cds  
74973 75060 cds  
75313 75420 cds  
75421 75537 utr
```

GENERAL FEATURE FORMAT (GFF)

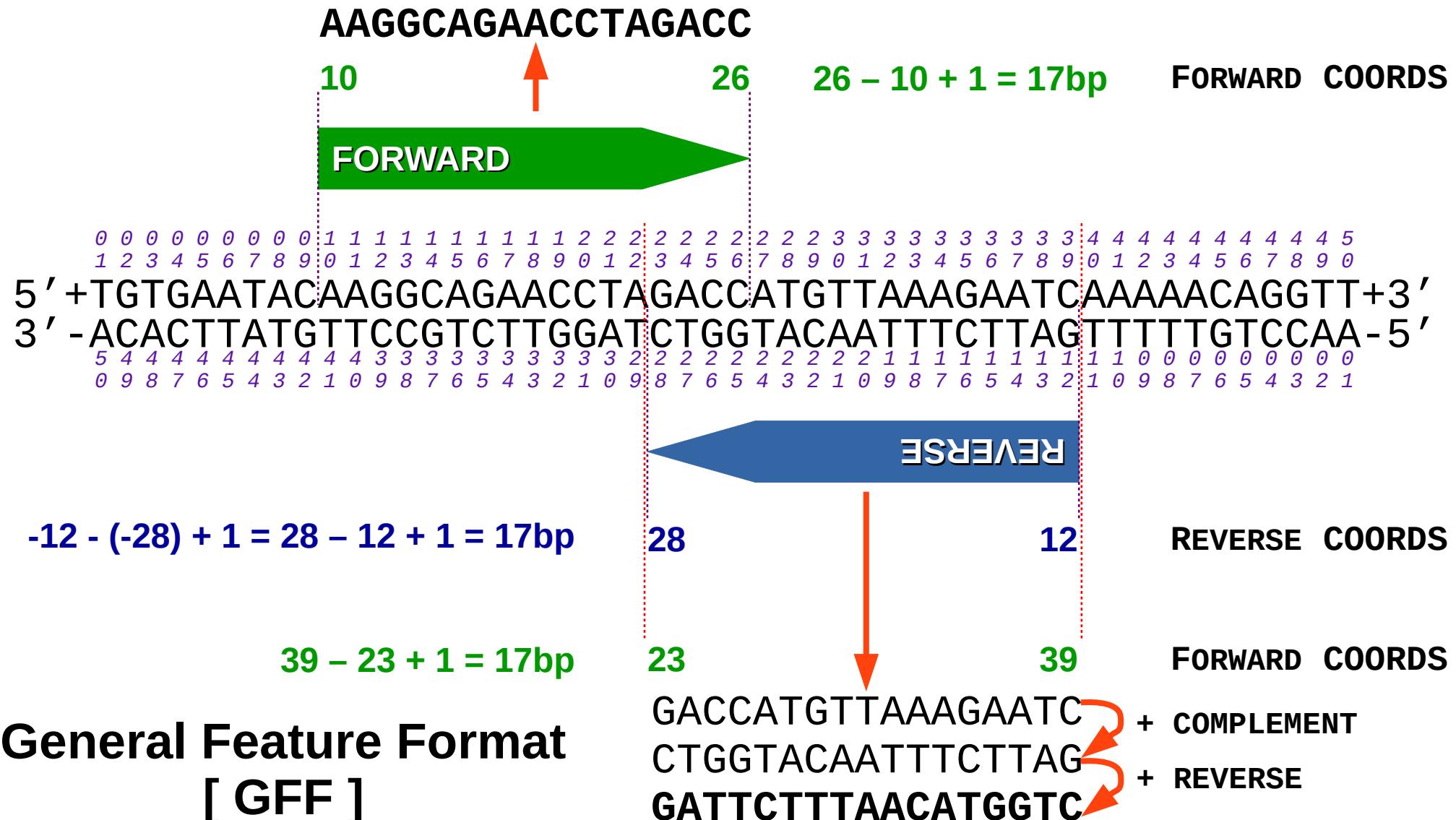
```
# Seq: chr1 Con: NM_021170  
chr1 refseq mRNA 74412 75537 . - . NM021170  
chr1 refseq utr 74412 74505 . - . NM021170  
chr1 refseq CDS 74506 74879 . - 2 NM021170  
chr1 refseq CDS 74973 75060 . - 0 NM021170  
chr1 refseq CDS 75139 75234 . - 0 NM021170  
chr1 refseq CDS 75313 75420 . - 0 NM021170  
chr1 refseq utr 75421 75537 . - . NM021170
```

UCSC BED

```
# GENE SEQUENCE STRAND mRNA_START mRNA_STOP CDS_START CDS_STOP EXONS EXONS_START EXONS_END  
  
NM_024796 chr1 - 801451 802749 801942 802434 1 801451, 802749,  
NM_021170 chr1 - 74411 75537 74505 75420 4 \ 74411, 74972, 75138, 75312, 74879, 75060, 75234, 75537,  
...
```

Strands & Annotation Coords

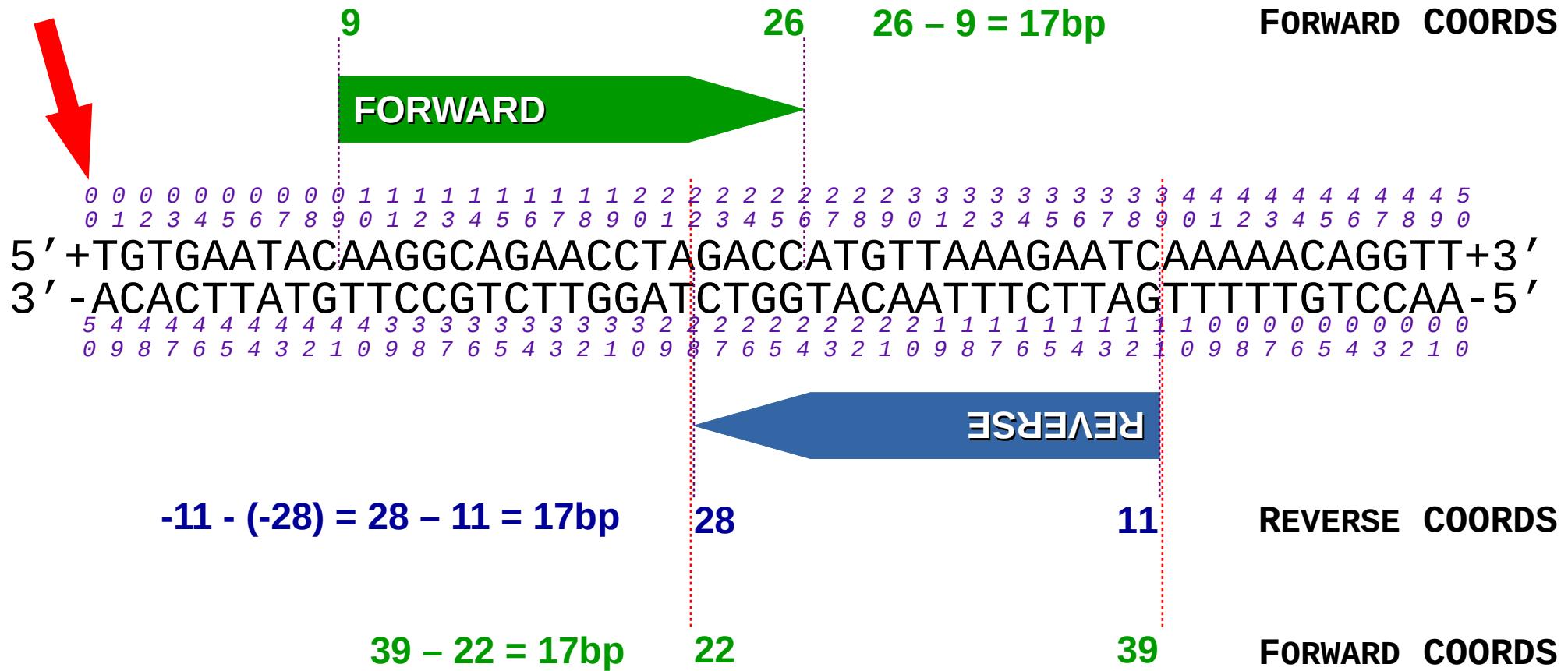
FULLY-CLOSED INTERVALS / 1-BASED



General Feature Format [GFF]

Strands & Annotation Coords

HALF-OPEN INTERVALS / 0-BASED



UCSC BED Format

More examples at <http://genome.ucsc.edu/blog/the-ucsc-genome-browser-coordinate-counting-systems/>

COMPLEX RECORDS: GENBANK

```
LOCUS      NM_021170 893 bp mRNA linear PRI 24-SEP-2005
DEFINITION Homo sapiens hairy and enhancer
               of split 4 (Drosophila) (HES4), mRNA.
ACCESSION  NM_021170
VERSION    NM_021170.2 GI:20127596
.
.
.
SOURCE     Homo sapiens (human)
FEATURES   Location/Qualifiers
source    1..893
          /organism="Homo sapiens"
          /mol_type="mRNA"
          /db_xref="taxon:9606"
          /chromosome="1"
          /map="1p36.33"
gene      1..893
          /gene="HES4"
          /db_xref="GeneID:57801"
.
.
.
CDS        118..783
          /gene="HES4"
          /note="bHLH factor Hes4"
          /codon_start=1
          /product="hairy and enhancer of split 4"
          /protein_id="NP_066993.1"
          /db_xref="GI:10863967"
          .
          /translation="MAADTPGKP SASP MAGAP ASAS RTPDKP
                        LLPGLTRALPAAPRAGP QGP GGP WRPWLR"
.
.
.
ORIGIN
  1 gggaaagaat gcggagccgg gttcacacac cccgcggcgg cgaggccta aataggaaaa
  61 cggcctgagg cgcgcgcccc cctggagccg ggatccgccc taggggctcg gatcgccg
.
.
.
  901 ccgttctagg gccgtggcct ttgccgagac tgttagcagag aaaacgtatt tattattcca
  961 ga
//
```

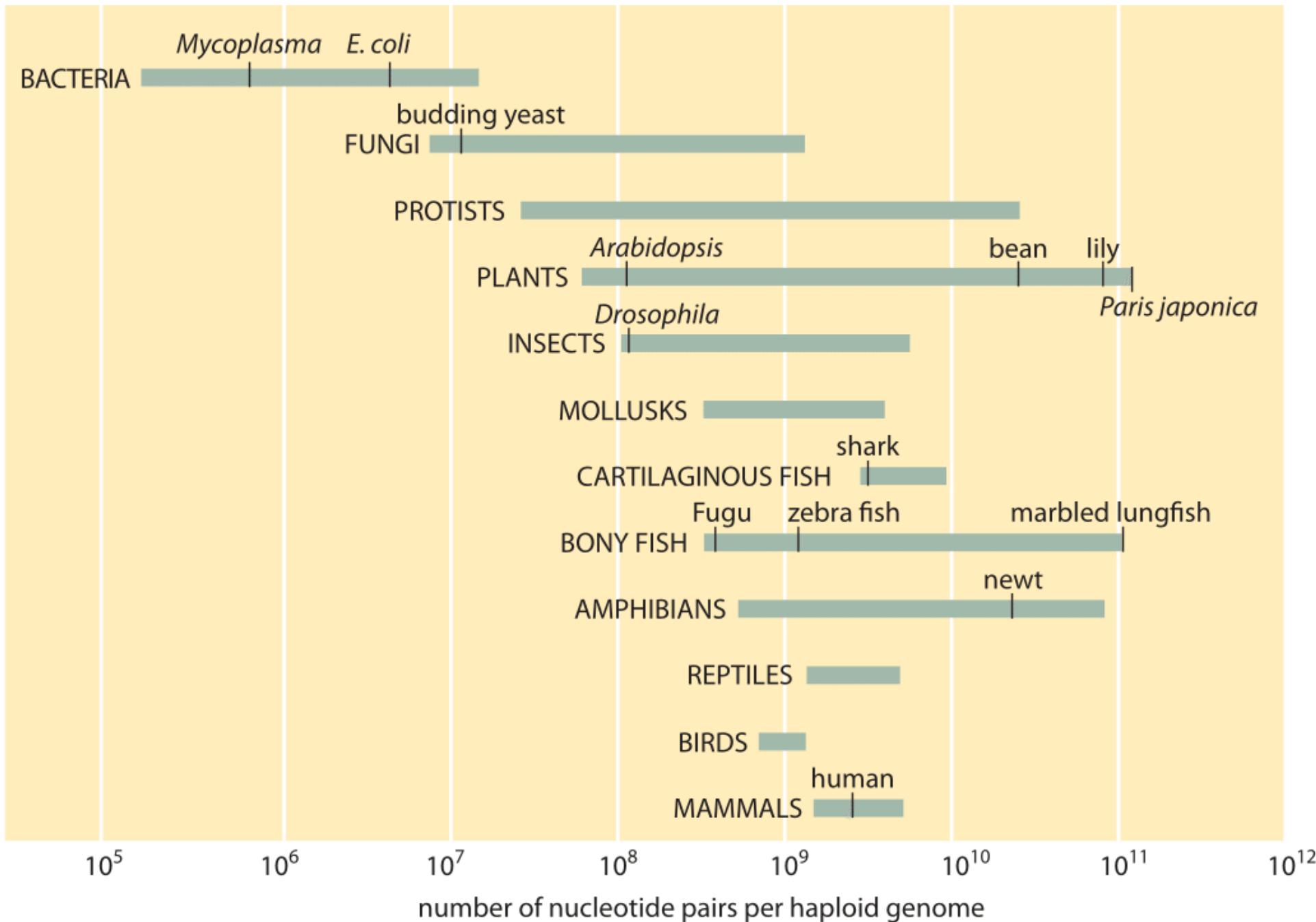
SEQUENCE STATISTICS



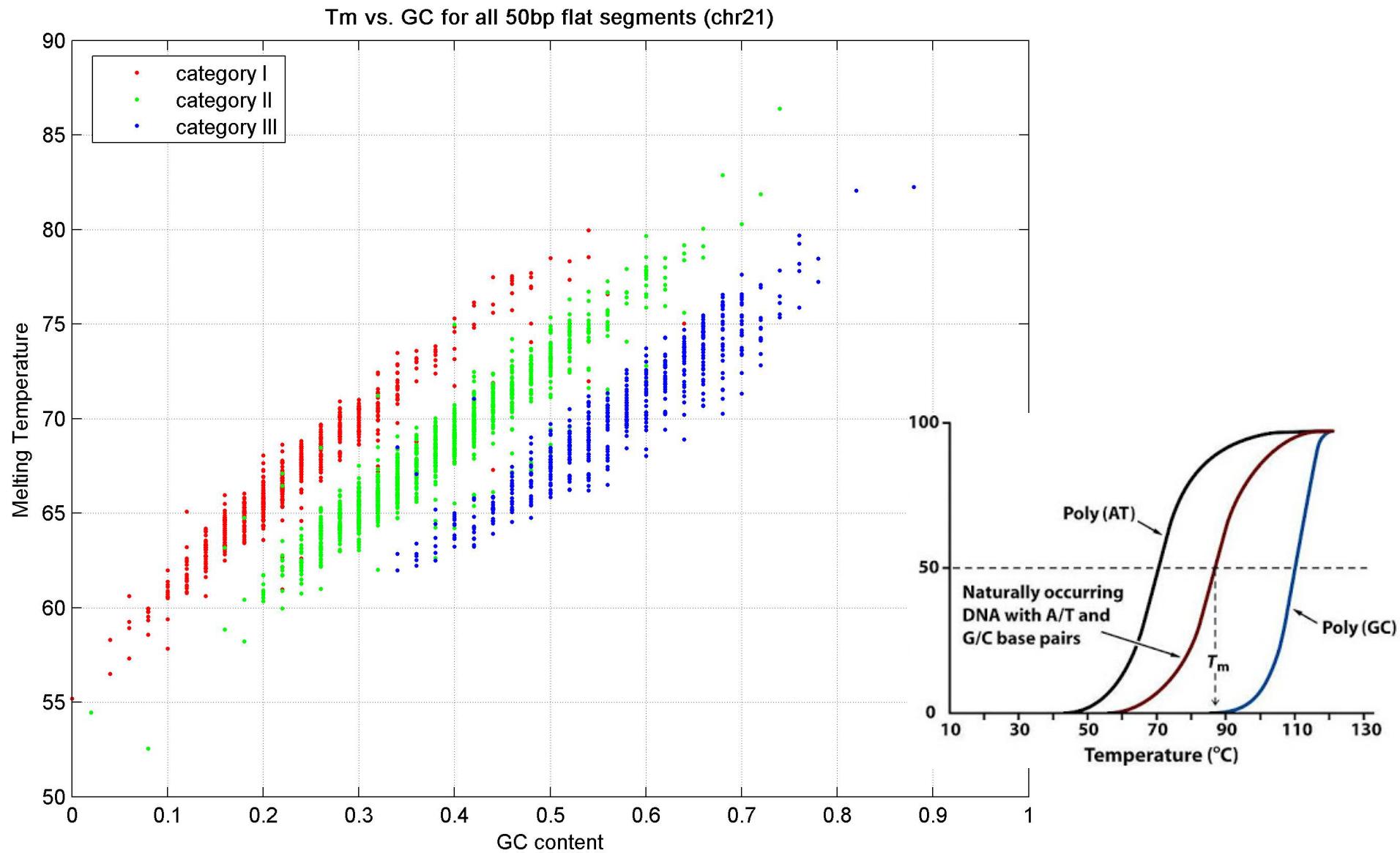
Basic Genome Stats

SPECIES	LENGTH (bp)	#GENES	GC Content
HIV Virus (ssRNA)	9,780	9	42.5
Lambda bacteriophage (dsDNA)	49,500	73	49.9
<i>Mycoplasma genitalium</i>	580,000	476	31.7
<i>Haemophilus influenzae</i>	1,800,000	2,098	37.9
<i>Escherichia coli</i>	4,609,000	5,449	50.6
<i>Saccharomyces cerevisiae</i>	14,625,900	6,692	40.7
<i>Plasmodium falciparum</i>	23,326,900	5,362	19.3
<i>Caenorhabditis elegans</i>	98,302,800	20,222	35.4
<i>Arabidopsis thaliana</i>	118,891,000	27,655	38.9
<i>Drosophila melanogaster</i>	143,726,000	13,931	42.1
<i>Takifugu rubripes</i>	391,485,000	18,523	45.8
<i>Homo sapiens</i>	3,257,320,000	20,376	41.5
<i>Mus musculus</i>	3,251,250,000	22,628	41.8
<i>Triticum aestivum</i>	15,344,700,000	110,790	46.1

C-VALUE Paradox

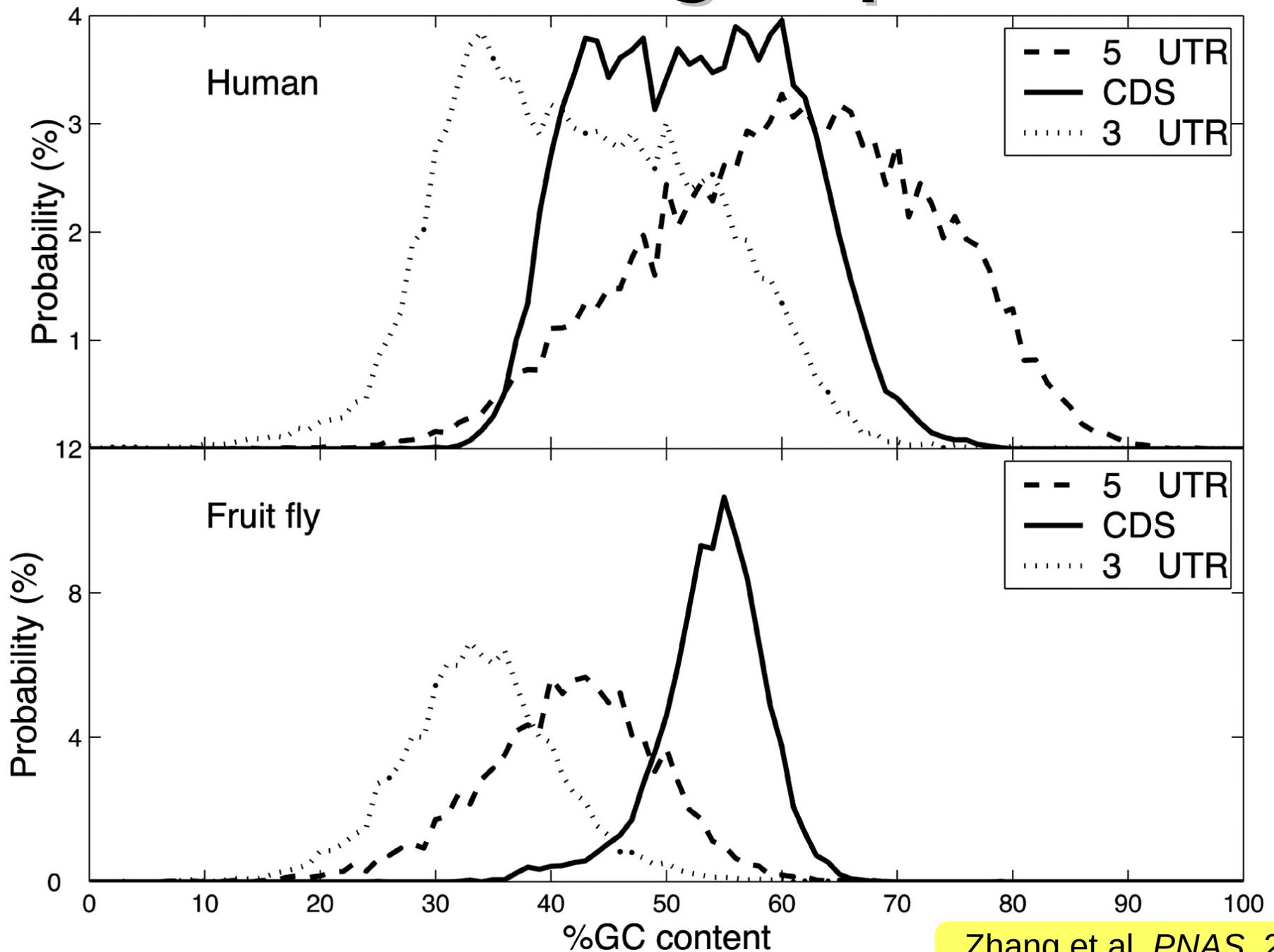


GC Content Affects Tm



Liu et al. PloS Comp Biol 2007

%GC on Coding Sequences



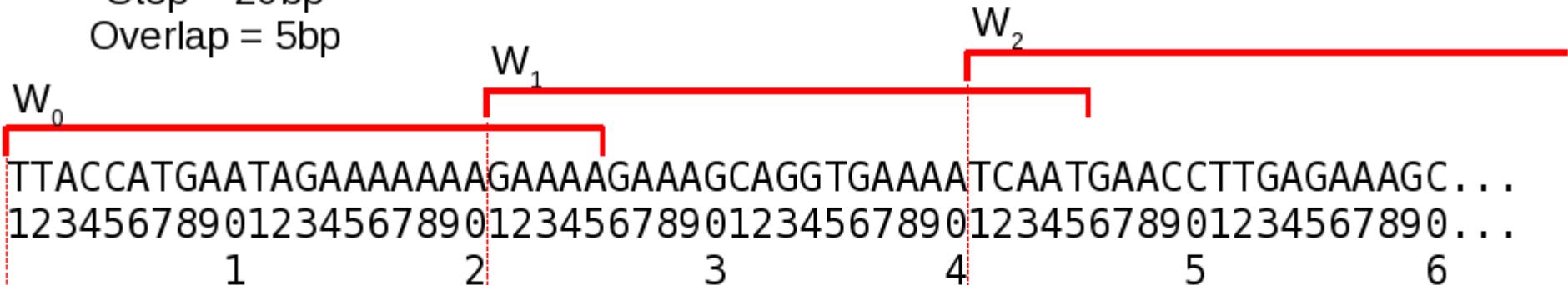
RUNNING WINDOWS

Window

Length = 25bp

Step = 20bp

Overlap = 5bp

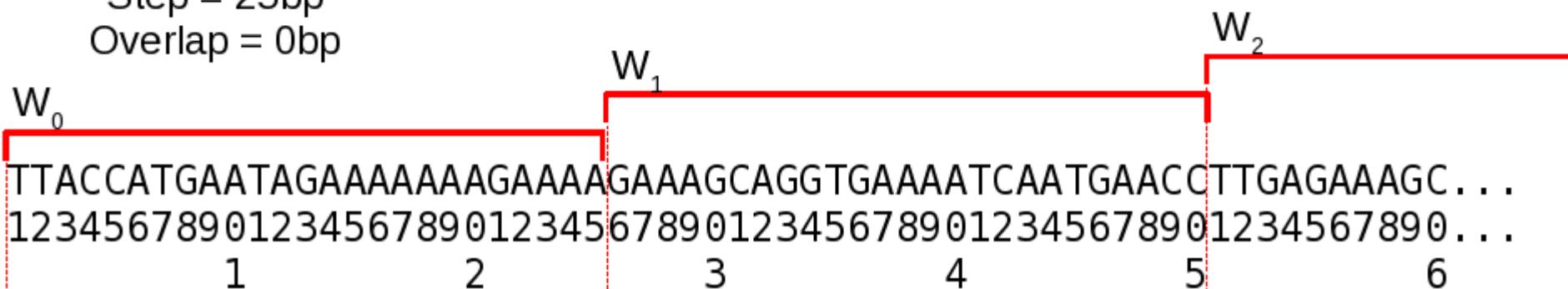


Window

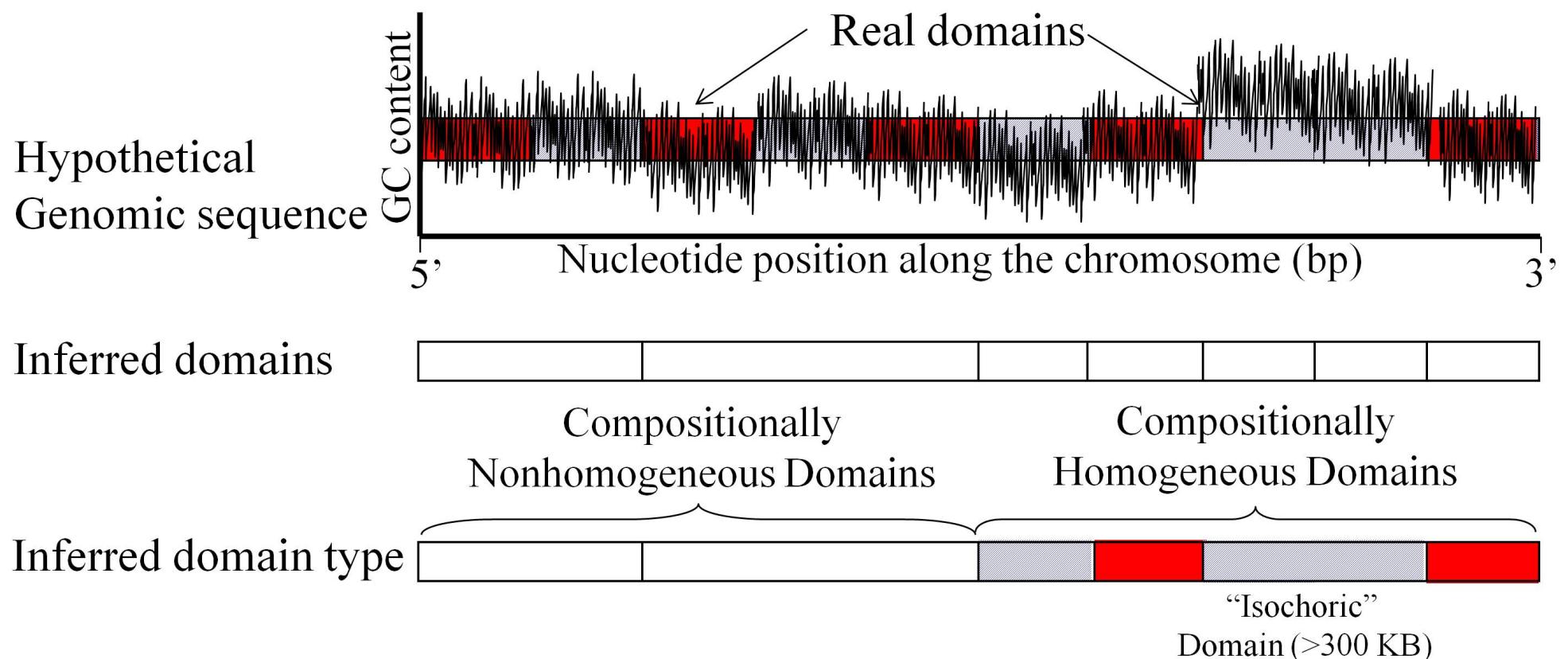
Length = 25bp

Step = 25bp

Overlap = 0bp

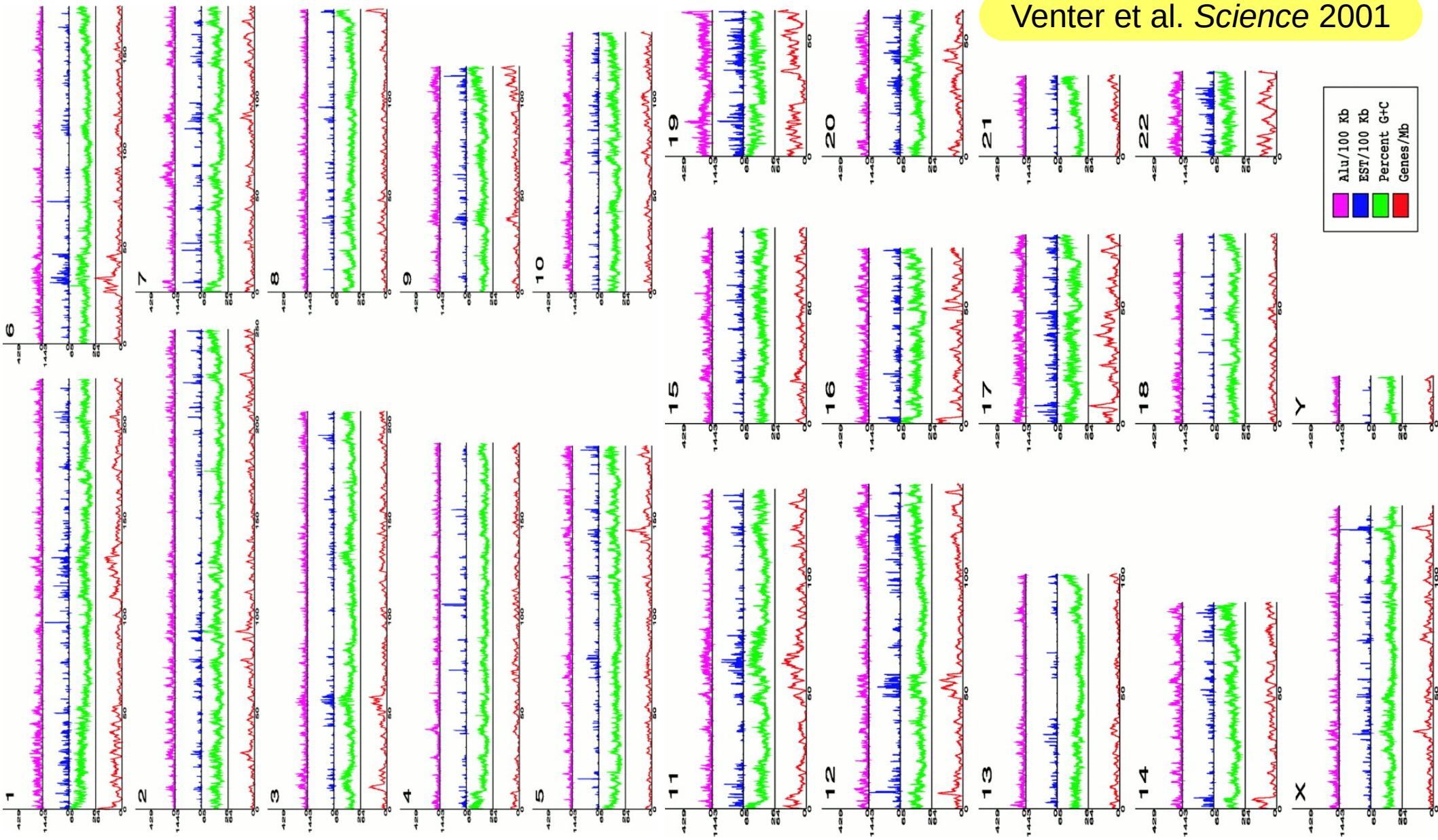


Genomic GC Content Variation

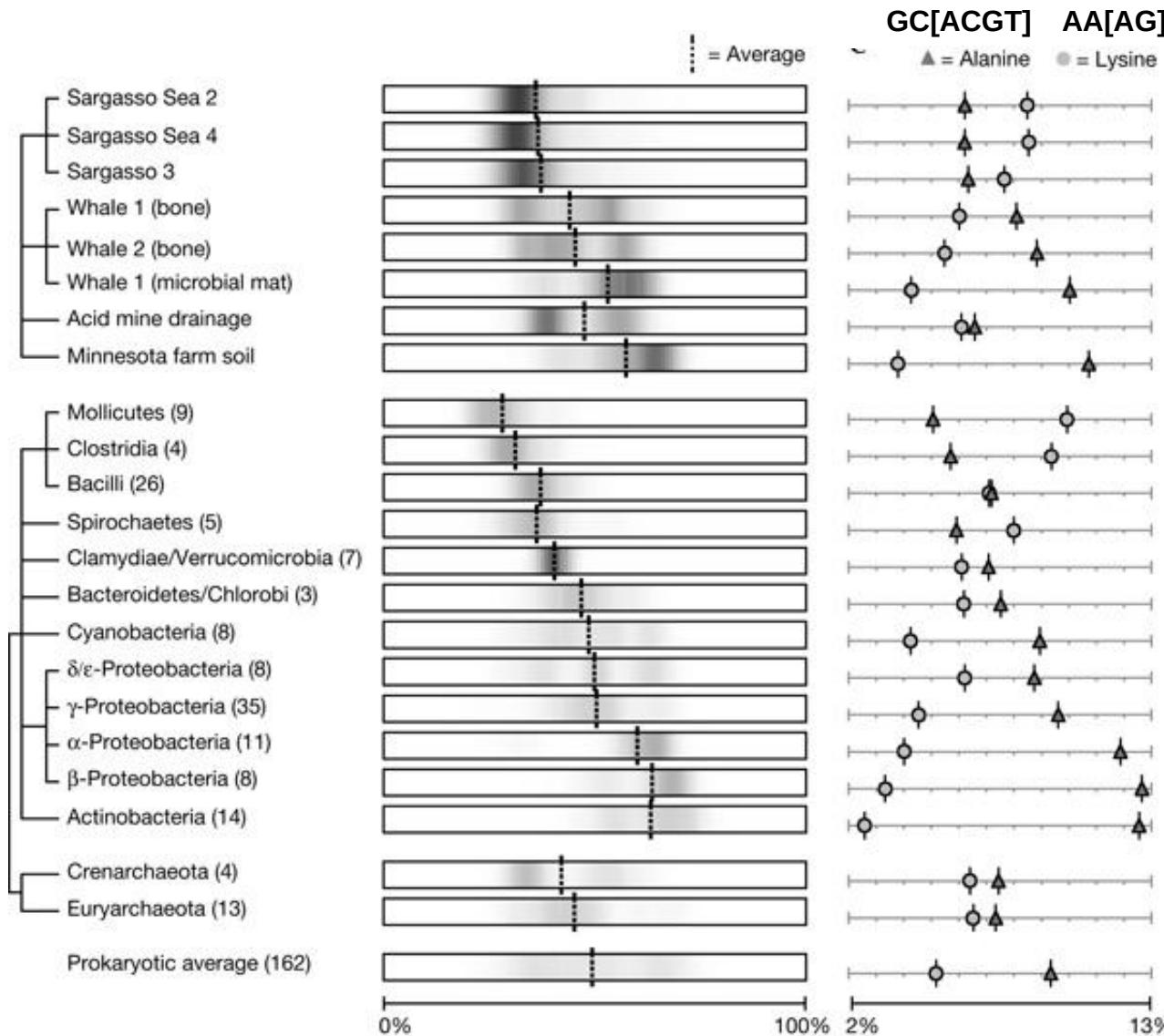


Genome-wide %GC Variation

Venter et al. *Science* 2001



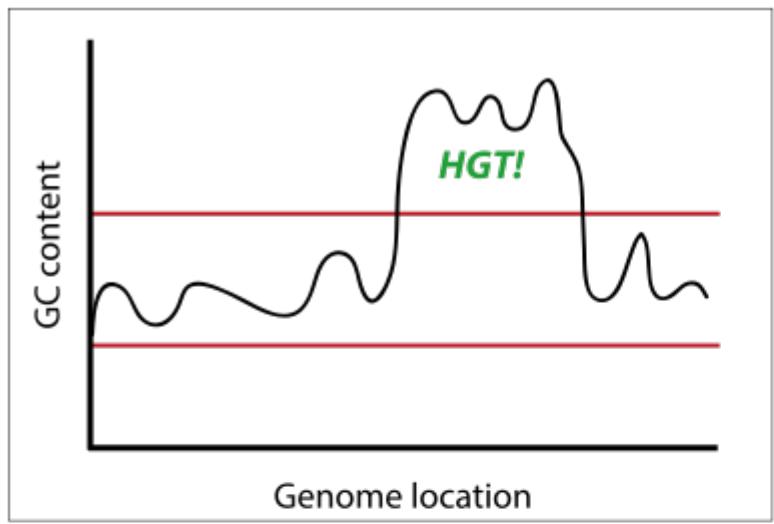
%GC and AA Composition



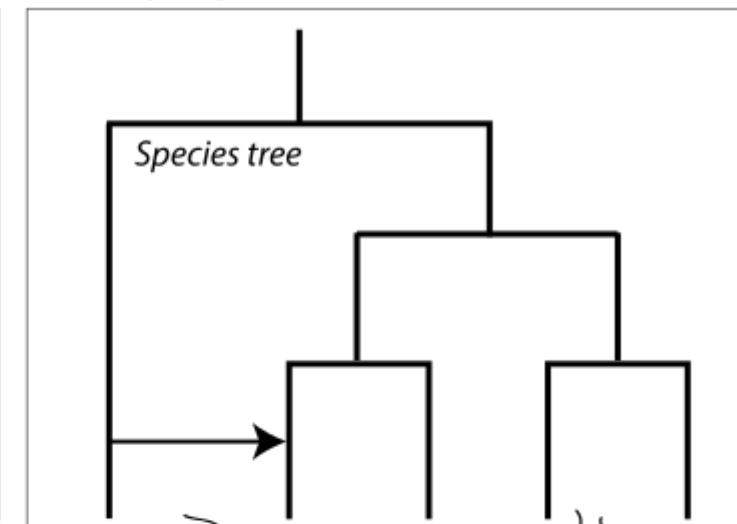
Foestner et al. EMBO Rep. 2005

%GC to Detect HGT Events

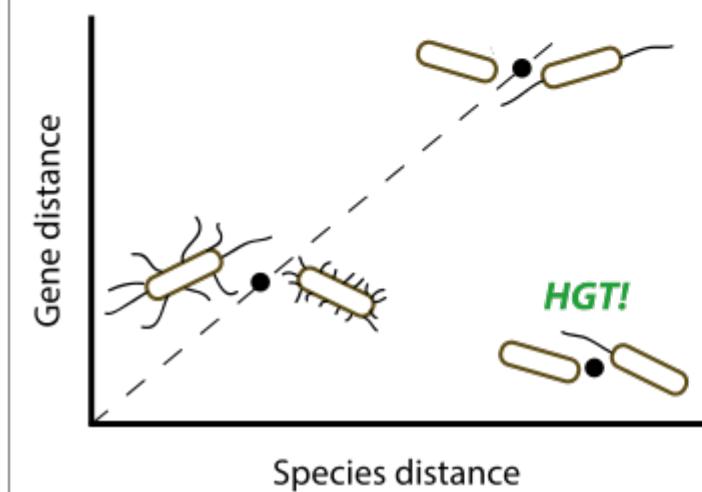
1. Parametric methods



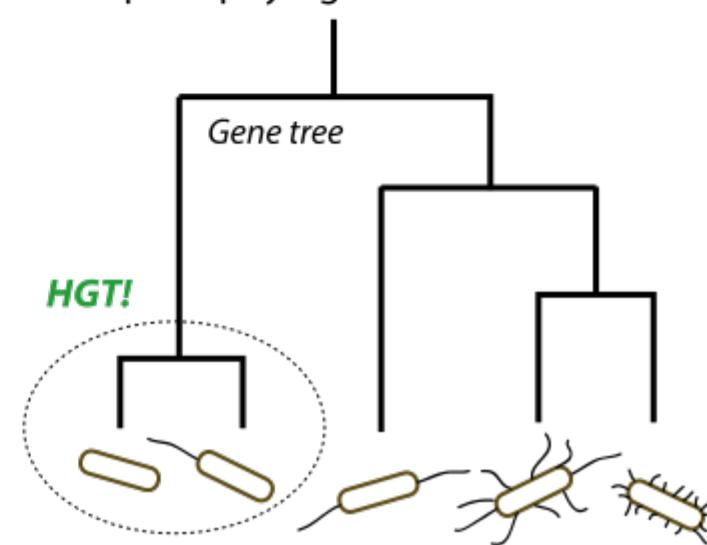
2. Phylogenetic methods



2a. Implicit phylogenetic methods



2b. Explicit phylogenetic methods



Ravenhall et al., PLoS Comput Bio, 2015

Strings of Symbols

We define an **alphabet** Σ , as a finite set of individual symbols:

$$\begin{aligned}\sum_{DNA\text{. IUPAC}} &= \{ a, c, g, t, u, r, y, m, k, w, s, b, d, h, v, n \} \\ \sum_{AminoAcids} &= \{ a, r, n, d, c, q, e, g, h, i, l, k, m, f, p, s, t, w, y, v, b, z, x \}\end{aligned}$$

A **string** S is an ordered list of contiguous symbols taken from an alphabet; where $|S|$ denotes the string **length**, the number of characters in string S .

$$S = (\alpha_1, \dots, \alpha_{|S|}); \quad \forall i \in [1, |S|]; \quad \alpha_i \in \Sigma$$

A **substring** $S[i..j]$ is a contiguous set of symbols from S , starting at position i and ending at position j of S . In particular:

$S[1..i]$, where $i \leq |S|$, is the **prefix** of string S ; and

$S[i..|S|]$, where $1 \leq i$, is the **suffix** of string S

Sequences as Strings

Given an **alphabet** for nucleotide sequences:

$$\Sigma_{DNA} = \{ a, c, g, t \}$$

A nucleotide **sequence** can be then modeled as a string of symbols taken from the DNA alphabet with a given length n .

$$seq_{DNA} = (\alpha_1, \dots, \alpha_n); \quad \forall i \in [1, n]; \quad \alpha_i \in \Sigma_{DNA}$$

A **k -mer** can be understood as a word or a substring within such sequence string, having a fixed length of k symbols (say here nucleotides).

$$kmer_{DNA} = (\beta_1, \dots, \beta_k);$$

$$\forall j \in [1, k]; \quad k \leq n; \quad \beta_j \in \Sigma_{DNA}; \quad kmer_{DNA} \subset seq_{DNA}$$