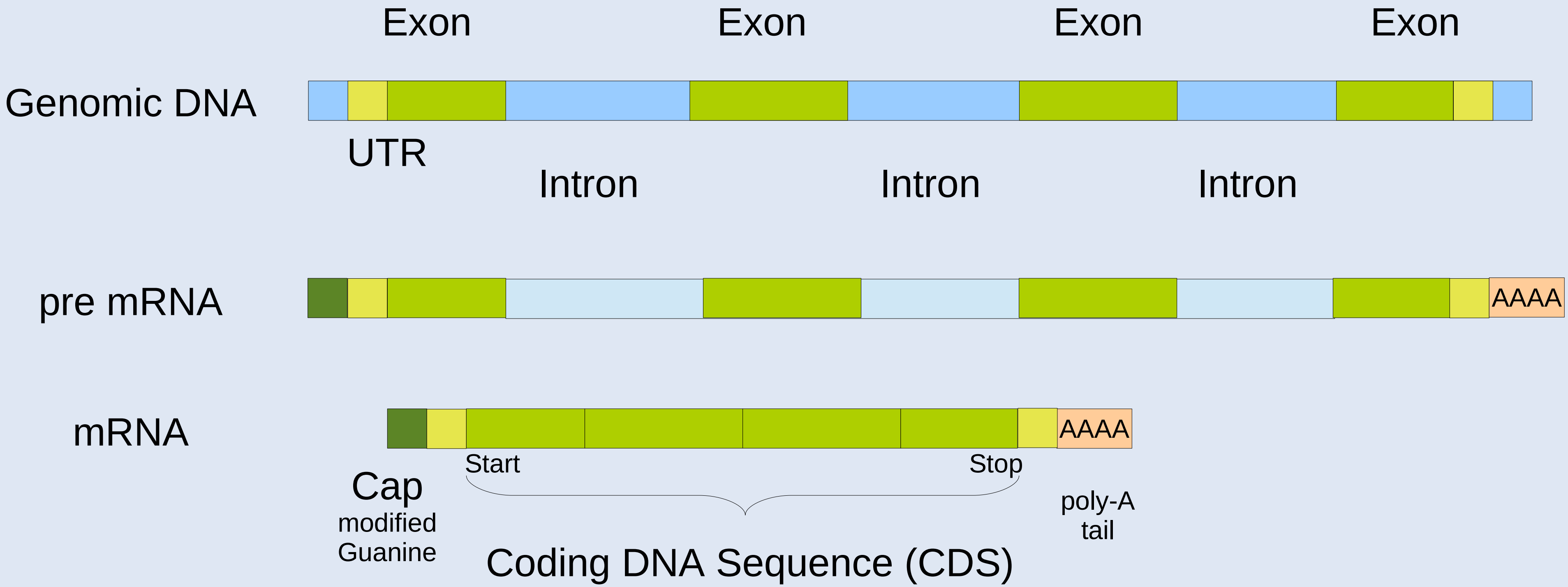
Topic 3 - Sequence Databases

Coding Sequence, cDNA, etc

cDNA is complementary DNA

This is DNA obtained by reverse-transcribing mRNA (messenger RNA)



Protein Primary Sequence

The ”primary sequence” of a protein is the linear sequence of amino acids in that protein

Proteins can be phosphorylated or glycosylated but this does not change its primary sequence.

DNA can also be modified: methylation, acetylation, phosphrylation, etc. and such changes are important in epigenetics.

DNA sequence stays the same.

Sequence Databases

The first nucleotide database was the European Molecular Biology Laboratory's EMBL database in Heidelberg, Germany; now called ENA.

GenBank followed, initially in Los Alamos, NM.

The third major nucleotide database was the DNA Data Bank of Japan

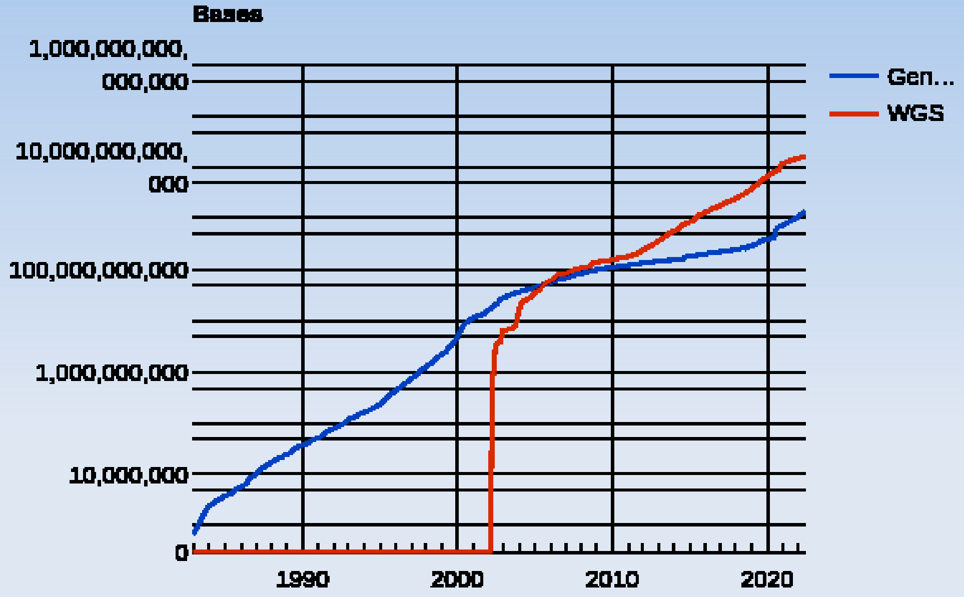
Databases

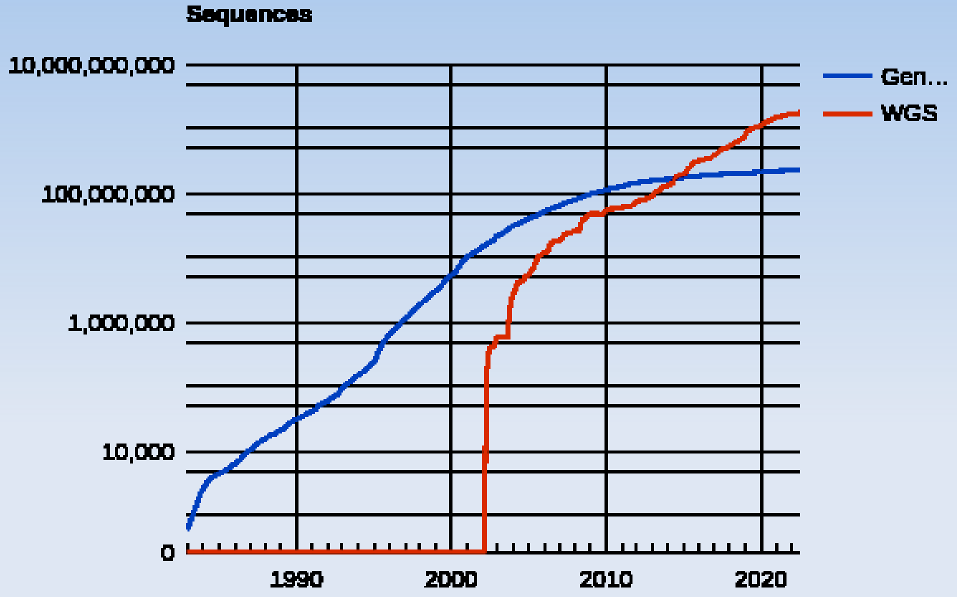
A database is a way of storing information with rules for searching and retrieving stored data.

The first automated database was the US 1890 Census data – precedes computers!

Database ~ tables of data.

Sequence Database





https://www.ncbi.nlm.nih.gov/genbank/statistics/

Growth
40 years of GenBank seqs,
20 years of Whole Genome Shotgun seqs

Databases and Applications

Soon after the web started up, biological databases also started setting up web interfaces.

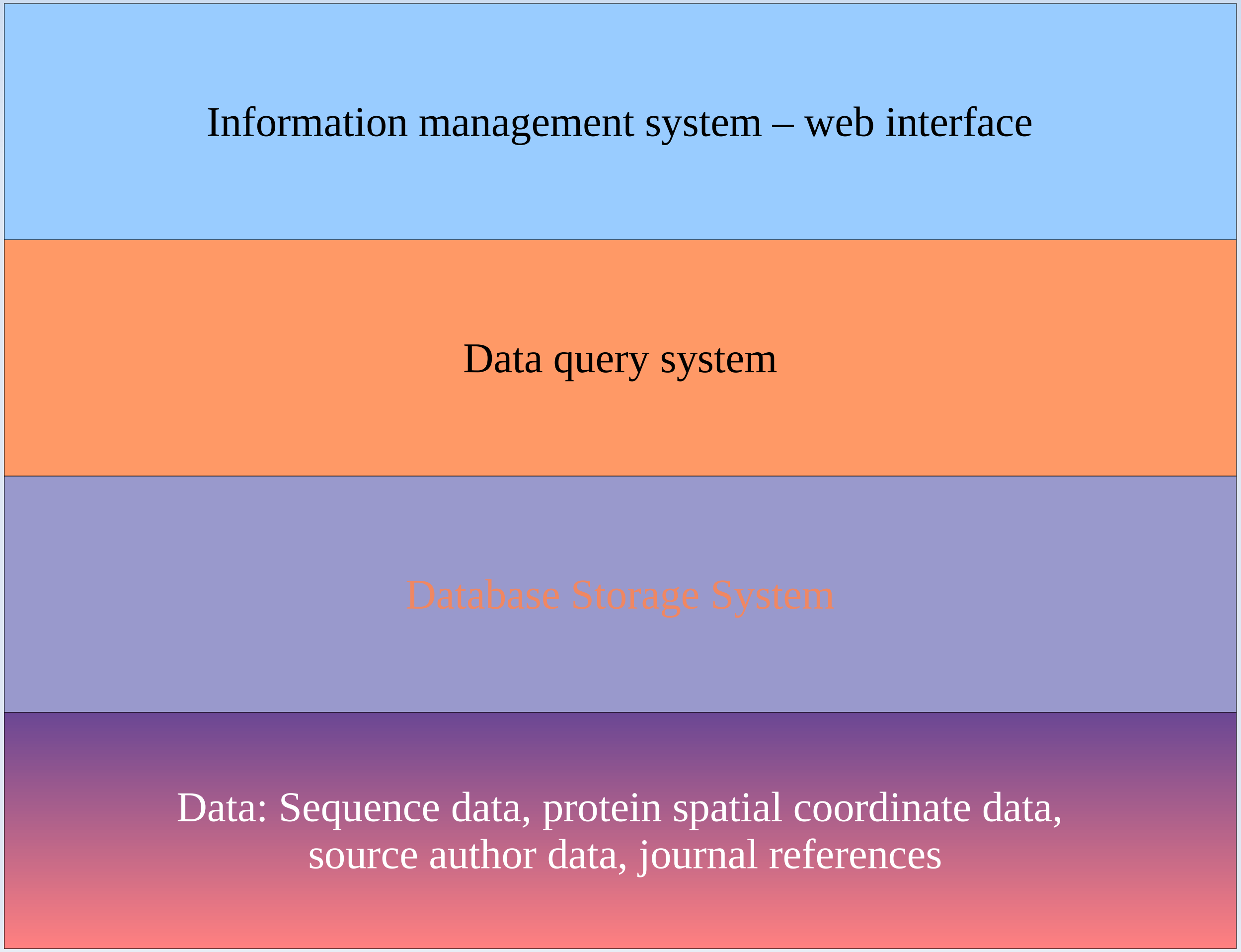
Initially these web sites were limited to searching and displaying the results of searches

Currently, most ”databases” have a data part and a number of applications for analying the data.

These applications can have complex algorithms.

Sequence Databases

There are many layers of software needed for databases of this scale.



Information management system – web interface

Data query system

Database Storage System

Data: Sequence data, protein spatial coordinate data, source author data, journal references

Other important databases

The RCSB Protein Data Bank started out as a place where 3D structures could be shared

Protein Information Resource

ExPASy – Expert Protein Analysis System

REBASE – Proteins that cut DNA

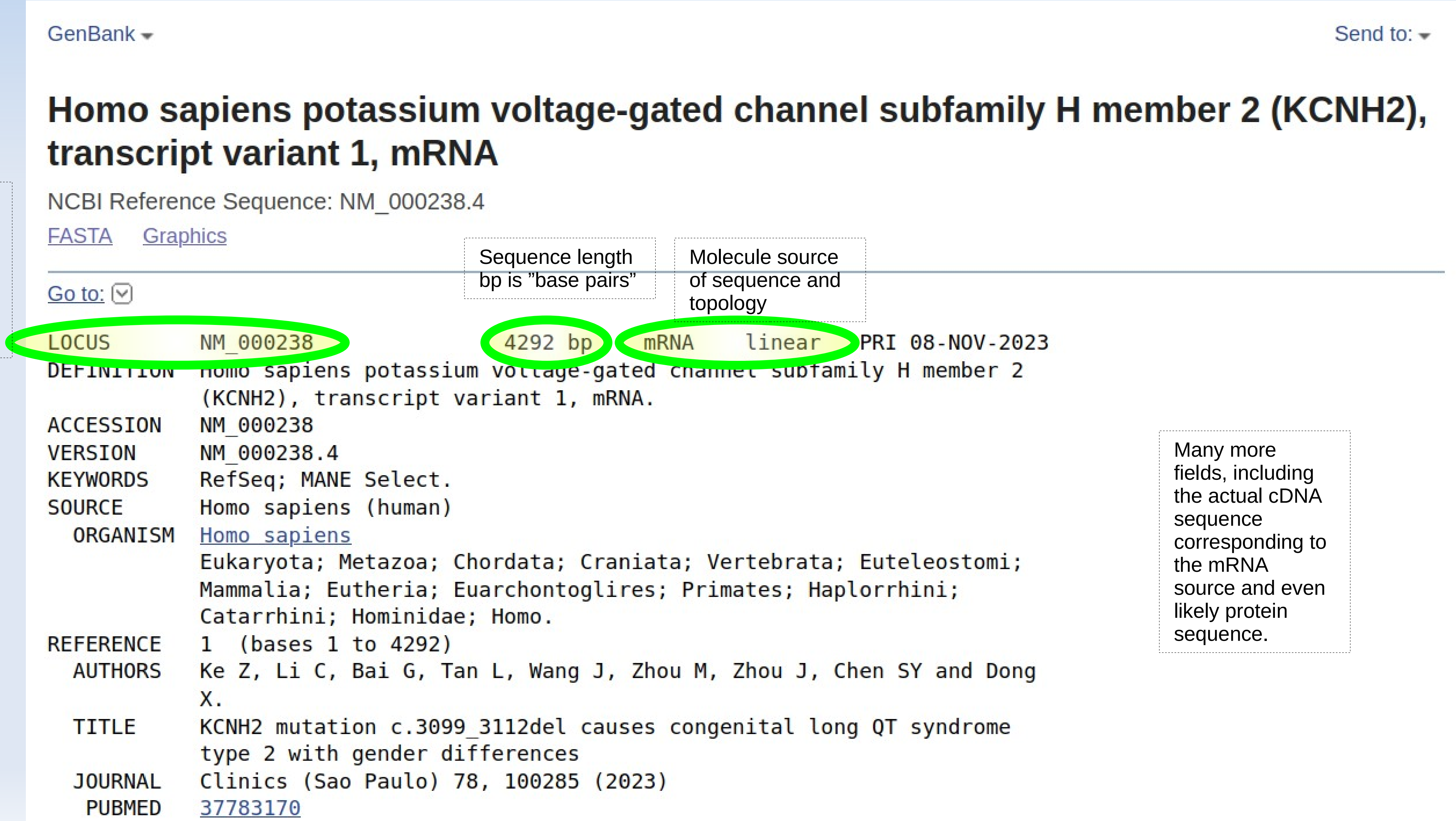
Searching for genes

Let's see what we can find out about a particular gene:

1) Go to the main page for NCBI: http://www.ncbi.nlm.nih.gov/
2) Choose Nucleotide for database and enter the word KCNH2, click Search
3) Click ”RefSeq transcripts” and find the search result with the accession ID NM_000238.4
4) Clicking that link pulls up a lot of information about this gene, including protein primary sequence and cDNA sequence.

Example Nucleotide Sequence

The NCBI Reference Sequence NM_000238.4:



GenBank ▾ Send to: ▾

Homo sapiens potassium voltage-gated channel subfamily H member 2 (KCNH2), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000238.4

FASTA Graphics

Go to: ▾

```
LOCUS       NM_000238               4292 bp    mRNA    linear   PRI 08-NOV-2023
DEFINITION  Homo sapiens potassium voltage-gated channel subfamily H member 2
            (KCNH2), transcript variant 1, mRNA.
ACCESSION   NM_000238
VERSION     NM_000238.4
KEYWORDS    RefSeq; MANE Select.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 4292)
  AUTHORS   Ke Z, Li C, Bai G, Tan L, Wang J, Zhou M, Zhou J, Chen SY and Dong
            X.
  TITLE     KCNH2 mutation c.3099_3112del causes congenital long QT syndrome
            type 2 with gender differences
  JOURNAL   Clinics (Sao Paulo) 78, 100285 (2023)
   PUBMED   37783170
```

The ”locus” used to be descriptive but is the same as the accession number now.

Sequence length bp is ”base pairs”

Molecule source of sequence and topology

Many more fields, including the actual cDNA sequence corresponding to the mRNA source and even likely protein sequence.

GenBank databases

GenBank is actually a collection of different databases (anyone can submit a sequence):

- Nucleotides
  - Reference Gene sequences
  - Chromosomes
  - Assembled genomes
  - Predicted (algorithmically from genome annotation) mRNA
  - ”Non-coding” RNA
- Protein
  - Predicted Protein

RefSeq database in GenBank

The goal of RefSeq is to provide a single reference sequence for each type molecule (DNA, mRNA, protein) for a single gene:

- Should include include biological attributes of gene
- Should reflect the current knowledge of sequence data and biology
- Include a wide range of species with focus on human and other mammals
- It is supposed to be non-redundant – so that we see only one search result for each gene.
- Constantly curated – can change at any time!

GenBank Accession numbers

Each sequence has an ID called ”Accession number”

Curated, experimentally verified:

NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
NG_123456	Reference Genomic Sequence
NT_123456	Genomic Sequence

XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA

CODIS

You may have seen ”investigators” on CSI, NCIS, and Law & Order match DNA from a suspect with data in the CODIS database.

CODIS is the ”Combined DNA Index System” - a DNA database set up by the FBI and used to identify people based on DNA profile.

2015: 12 million offender, 2 million arrestee profiles used in 300,000 crime investigations.

2021: 20 million DNA profiles

If people's genes are so similar, how can we identify a person using DNA?

Short Tandem Repeats - STR

An STR is a sequence of genomic DNA with a highly repetitive DNA.

1-6 bp repeating 10-50 times

– e.g. ATATATAT or GAAGAAGAA

STRs are often in non-coding regions → most STRs are biologically silent

High mutation rate → happens early in embryo development → each person is likely to have a ”unique” set of STRs.

Microsatellites

Microsatellites are STRs.

In humans, the 5 most abundant microsatellites are:

A, AC, AAAN, AAN, and AG (N is C, G, or T)

Microsatellites are found throughout the human genome

There are variations in the number of tandem repeats (VNTR).

Microsatellites

Microsatellites are used as markers to construct genetic maps

And for forensic medicine because they are polymorphic – they have different number of repeats that can be used to identify individuals and close relatives.

Microsatellites

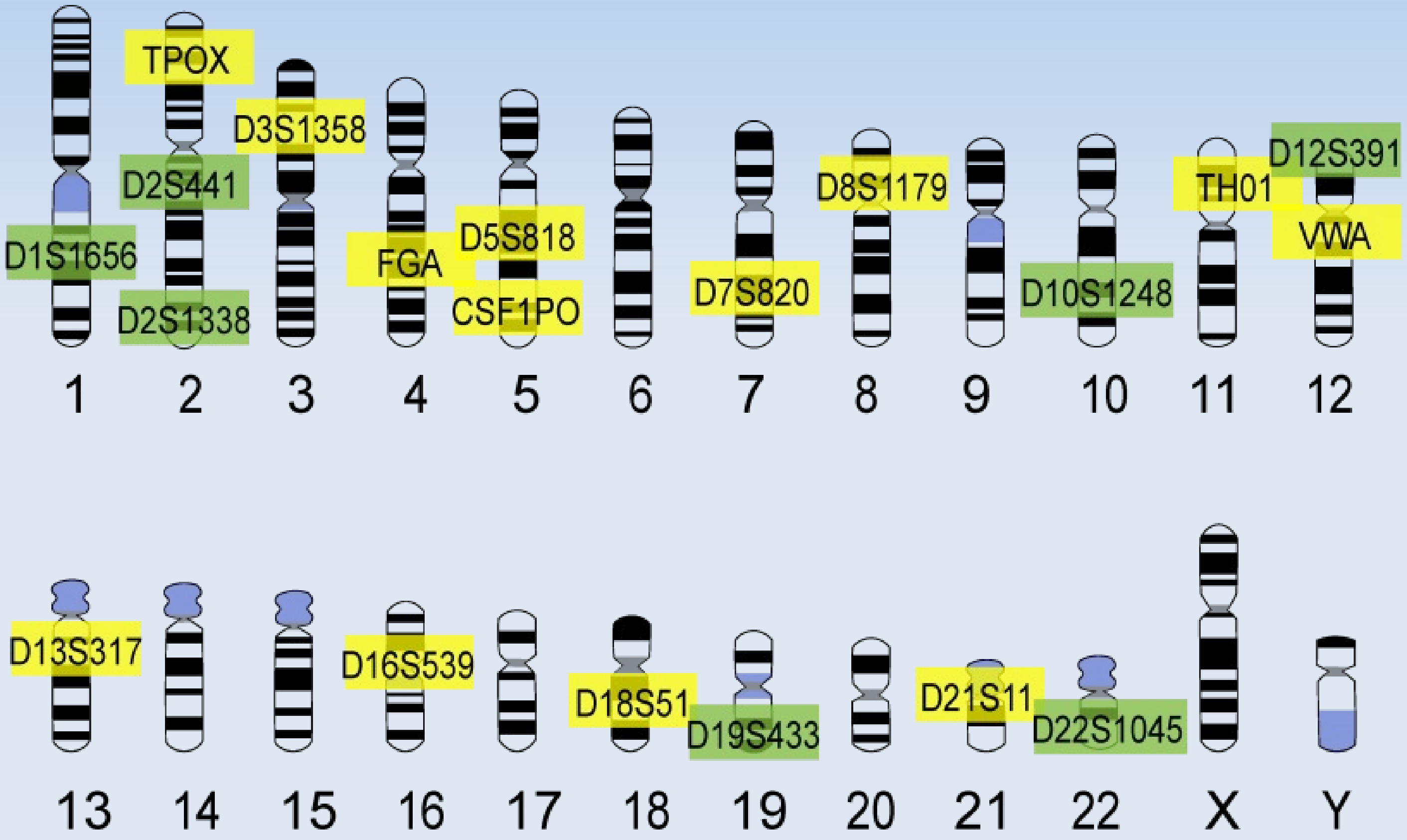
Some triplet repeats are associated with diseases:

Spinal and Bulbar Muscular Atrophy (SBMA) is due to CAG repeats in an androgen receptor gene

Fragile X Syndrome is due to a CGG repeat in FMR1 gene in the X chromosome

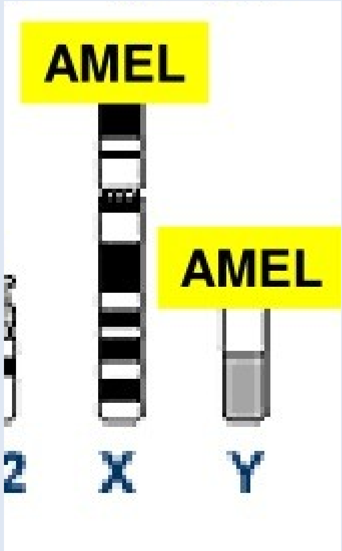
CODIS uses 20 STR loci + AMEL

STR loci used in U.S. Combined DNA Index System (CODIS)



🟨 13 original CODIS Core STR Loci

🟩 Additional 7 CODIS Core STR Loci added Jan 2017



AMEL – Amelogenin to determine sex (gender)