Database	Data Type	Curated	Organism Scope	URL
NCBI GenBank	DNA / RNA	Yes	Global	Link
EMBL-EBI ENA	DNA / RNA	Yes	Global	Link
DDBJ	DNA / RNA	Yes	Global	Link
RefSeq (NCBI)	DNA / RNA / Protein	Yes	Global	Link
RNAcentral	Non-coding RNA	Partially	Global	Link
UniProt	Protein	Yes (Swiss-Prot)	Global	Link
PDB	Protein / DNA / RNA (3D)	Yes	Global	Link
Pfam	Protein Domains	Yes	Global	Link
InterPro	Protein Signatures	Yes	Global	Link
Human Protein Atlas	Protein Expression	Yes	Human	Link
STRING	Protein Interactions	Yes	Global	Link
Ensembl	Genomes / Genes / Proteins	Yes	Global	Link
UCSC Genome Browser	Genomes / Genes	Yes	Human, Model Organisms	Link

Here's a list of commonly used model organisms in biology, their reference genomes, associated databases, and some key features relevant for biological research:

🧬 1. *Escherichia coli* (E. coli)

- **Strain: K-12 MG1655**
- **Genome Size: ~4.6 Mb**
- **Database: NCBI Genome, EcoCyc**
- **Key Features:**
  - **Prokaryotic model**
  - **Widely used in molecular biology and genetics**
  - **Fast growth and easy manipulation**

🥯 2. *Saccharomyces cerevisiae* (Baker's yeast)

- **Genome Size: ~12 Mb**
- **Database: Saccharomyces Genome Database (SGD)**
- **Key Features:**
  - **Eukaryotic unicellular model**
  - **Used in cell cycle, genetics, and aging research**
  - **First eukaryotic genome to be fully sequenced**

🐛 3. *Caenorhabditis elegans*

- **Genome Size: ~100 Mb**
- **Database: WormBase**

- **Key Features:**
  - **Transparent nematode**
  - **Model for development, neurobiology, and apoptosis**
  - **First multicellular organism with a sequenced genome**

🐞 4. *Drosophila melanogaster* (Fruit fly)

- **Genome Size: ~180 Mb**
- **Database: FlyBase**
- **Key Features:**
  - **Short life cycle**
  - **Extensive genetic tools and mutant libraries**
  - **Key in developmental biology and genetics**

🐟 5. *Danio rerio* (Zebrafish)

- **Genome Size: ~1.5 Gb**
- **Database: ZFIN, Ensembl**
- **Key Features:**
  - **Transparent embryos**
  - **Model for vertebrate development and disease**
  - **Easy for genetic manipulation**

🐭 6. *Mus musculus* (House mouse)

- **Genome Size: ~2.7 Gb**
- **Database: Mouse Genome Informatics (MGI), Ensembl**
- **Key Features:**
    - **Mammalian model**
    - **Shares ~99% genes with humans**
    - **Extensive use in immunology, neurobiology, cancer research**

🐵 7. *Homo sapiens* (Human)

- **Genome Size: ~3.2 Gb**
- **Database: NCBI Genome, Ensembl, UCSC Genome Browser**
- **Key Features:**
    - **Medical relevance**
    - **Extensive functional annotation**
    - **Genome-wide association studies (GWAS), personalized medicine**

🌿 8. *Arabidopsis thaliana*

- **Genome Size: ~135 Mb**
- **Database: TAIR (The Arabidopsis Information Resource)**
- **Key Features:**
    - **Model plant**
    - **Short lifecycle, small genome**
    - **Extensively used in plant genetics, development, and stress physiology**

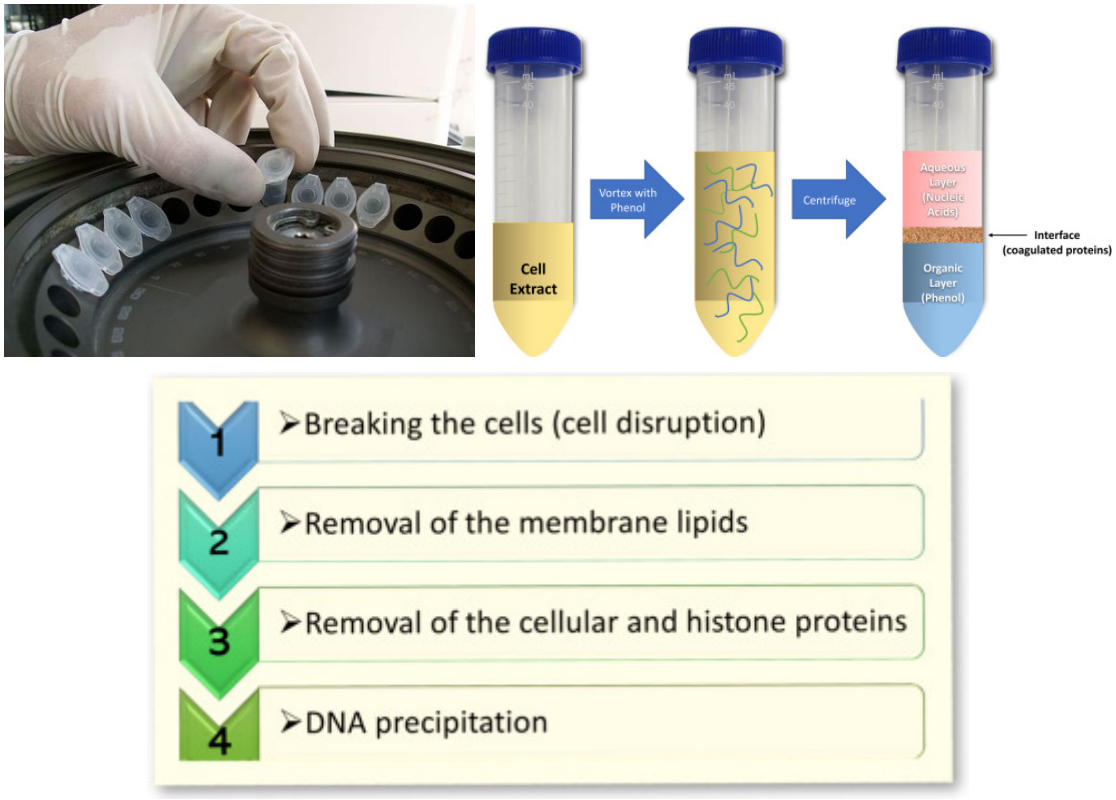
🦠 9. *Mycobacterium tuberculosis*

- **Strain: H37Rv**
- **Genome Size: ~4.4 Mb**
- **Database: TubercuList, NCBI Genome**
- **Key Features:**
    - **Pathogen model**
    - **Studied for tuberculosis pathogenesis and drug resistance**

🦠 10. *Plasmodium falciparum*

- **Genome Size: ~23 Mb**
- **Database: PlasmoDB**
- **Key Features:**
    - **Malaria-causing parasite**
    - **Complex life cycle**
    - **Target for drug and vaccine development**

PCR: Polymerase Chain Reaction/ Primer BLAST/ FASTA format



The polymerase chain reaction (PCR) was invented in the 1980s by Kary Mullis and his colleagues at Cetus Corporation. PCR is a laboratory technique that amplifies specific DNA sequences.

PCR (Polymerase Chain Reaction) mimics the natural process of DNA replication in a test tube, allowing for the rapid amplification of specific DNA sequences.
https://www.youtube.com/watch?v=TNKWgcFPHqw

Polymerase Chain Reaction (PCR):

PCR is a widely used molecular biology technique that amplifies specific DNA fragments, even from very small or degraded samples. It's simple, cost-effective, and used in applications like research, diagnostics, and forensic science.

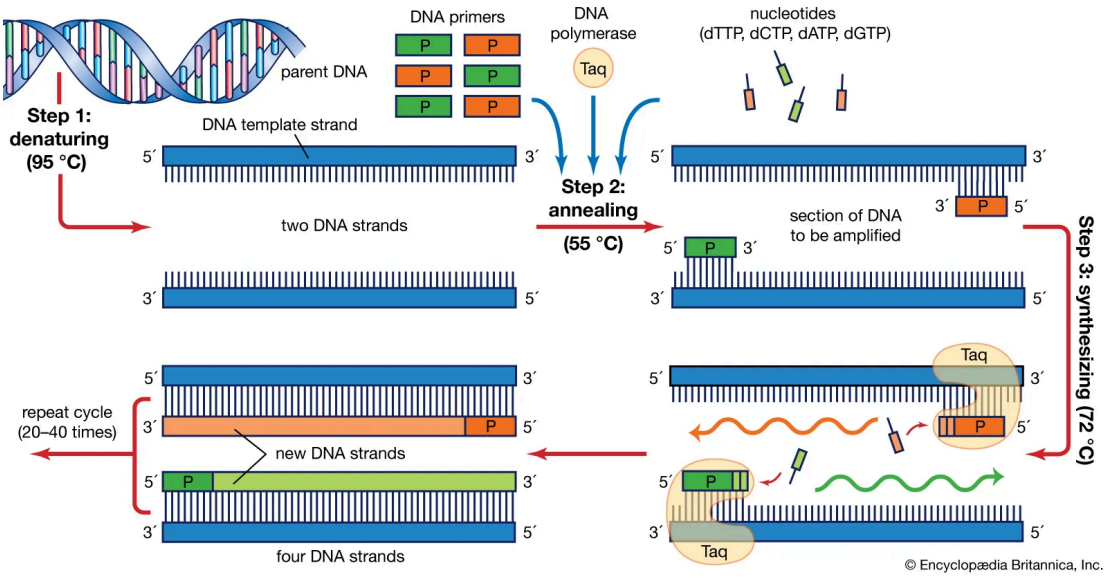
PCR Process:

1. **Initial Denaturation:** DNA is heated to 95°C to separate double-stranded DNA into single strands.
2. **Cycling:**
    - **Denaturation:** DNA is heated to 95°C to melt into single-stranded DNA.
    - **Annealing:** The temperature is lowered to allow primers to bind to the DNA at 45–60°C.
    - **Extension:** The temperature is raised to 72°C for DNA polymerase to extend the primers and replicate the DNA.

3. **Repetition:** These steps are repeated for 30–40 cycles to exponentially amplify the target DNA.

https://www.youtube.com/watch?v=2KoLnIwoZKU

PCR's simplicity and ability to amplify DNA from minimal and degraded samples make it invaluable in a range of applications.



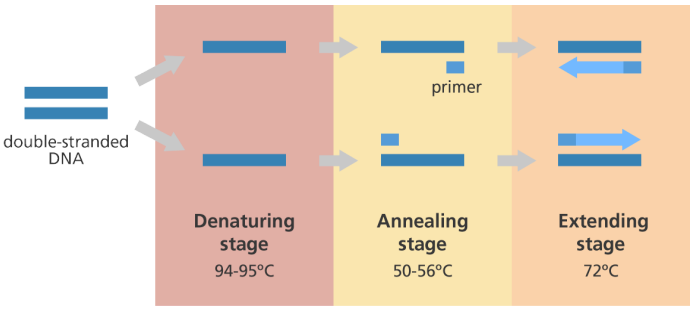


Figure 7. The thermal cycler is used to carry out PCR reaction. Image by Walter Suza.

Taq Polymerase Overview:

- **Discovery:** Taq polymerase was isolated from *Thermus aquaticus* in Yellowstone in 1976 by Chien et al.
- **PCR Development:** In 1985, PCR using the Klenow fragment of E. coli DNA polymerase was introduced. Taq polymerase was later used in PCR (1988) due to its heat stability, making PCR practical.
- **Revolutionizing PCR:** Taq polymerase allows PCR to work without needing to add fresh enzymes after each denaturation step. Perkin-Elmer Cetus provided PCR kits and thermal cyclers.

- **Recombinant Taq Polymerase:** The Taq polymerase gene was cloned and expressed in E. coli, leading to the production of AmpliTaq DNA polymerase for commercial use.
- **Overproduction:** Codon optimization of the Taq gene improved its production by over 10-fold, leading to higher commercial availability.

PCR amplifies a specific region:

The PCR reaction uses primers to target and replicate only a specific section of the DNA template, resulting in a much shorter DNA fragment than the original DNA.

Gel electrophoresis separates DNA fragments by size:

When running DNA on a gel, smaller fragments travel further through the gel matrix than larger fragments, allowing for visualization of different DNA sizes based on their migration distance.

Before PCR:

A DNA sample before PCR would show a **smear on the gel because it contains a mixture** of very long DNA fragments from the genome.

After PCR:

After PCR, the gel will show a **distinct band corresponding to the amplified fragment size,** as the PCR reaction has produced a large quantity of identical, short DNA fragments.

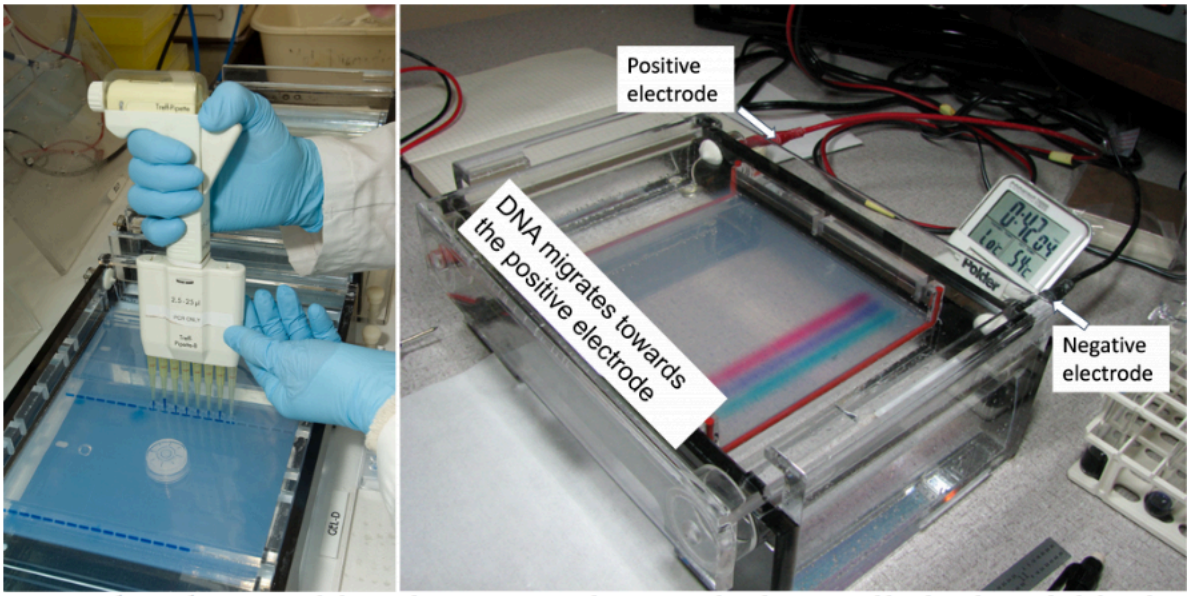


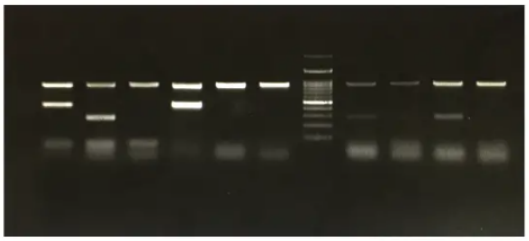
Figure 8. Electrophoresis: A gel electrophoresis set-up with agarose gel with DNA and loading dye on the left and the power supply on the right. Image Source: Michael, CC BY 2.0, via Wikimedia Commons and U. S. Department of Agriculture, CC BY 2.0, via Wikimedia Commons.

Types of PCR:

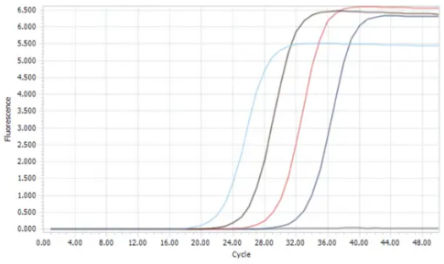
- **Conventional PCR**: Basic amplification.
- **Multiplex PCR**: Amplifies multiple targets at once.
- **Nested PCR**: Increases specificity using two sets of primers.
- **Reverse Transcriptase PCR (RT-PCR)**: Converts mRNA to cDNA for amplification.
- **Quantitative PCR**: Quantifies DNA or RNA.
- **Hot-start PCR**: Reduces nonspecific amplification by heating components before adding polymerase.
- **Touchdown PCR**: Gradually decreases annealing temperature for better primer binding.
- **Assembly PCR**: Synthesizes long DNA segments by joining oligonucleotides.
- **Methylation-specific PCR**: Detects DNA methylation patterns.
- **LAMP Assay**: Amplifies DNA using a loop-mediated technique.
- **Real-time PCR**: Quantifies PCR product as it amplifies.

Similarities between PCR and qPCR:

Both methods are used to amplify or synthesize the DNA. Both the techniques are based on the temperature-based amplification. The machine used for both techniques is known as a thermocycler. Both RNA and DNA can be amplified in PCR as well as qPCR however, to amplify the RNA a different type of DNA polymerase is used.



Results of PCR



Results of qPCR

PCR vs. qPCR: Key Differences and Applications

1. **Primers and Probes**:
    - **PCR**: Uses **simple, non-labeled primers**.
    - **qPCR**: Utilizes **labeled probes** with both a **quencher dye** and a **reporter dye** for fluorescence detection.
2. **Amplification Process**:
    - **PCR**: Involves three main steps—**denaturation**, **annealing**, and **extension**. The results are typically analyzed after the reaction is completed, often using **gel electrophoresis** to visualize the DNA fragments.
    - **qPCR**: Also includes the basic three steps, but the key difference is **quantification**. It measures **fluorescence** during the **exponential phase** of amplification, offering real-time data about the amount of DNA.

3. **Results Detection**:
    - **PCR**: Results are visible after the reaction, typically as distinct **DNA bands** or **amplicon bands** on an **agarose gel electrophoresis**.
    - **qPCR**: Results are recorded as **fluorescence emissions**, which are converted into **peaks** on a graph. These peaks indicate positive amplification and can be monitored during the reaction.
4. **Resolution**:
    - **PCR**: A **low-resolution technique**, producing discrete bands after the reaction.
    - **qPCR**: A **high-resolution technique** that provides more detailed, quantitative data during the amplification process.
5. **Time and Expertise**:
    - **PCR**: A more **time-consuming** process, taking **3 to 4 hours** for preparation and execution. Analyzing results requires **high expertise**, especially in interpreting gel electrophoresis data.
    - **qPCR**: **Faster**, typically completed in **1 to 1.5 hours**, and requires less technical expertise to interpret results, as the machine provides the data in real-time.
6. **Type of Data**:
    - **PCR**: **Qualitative**—detects the presence or absence of a specific sequence or mutation.
    - **qPCR**: **Quantitative**—measures the **amount** of DNA or RNA during the amplification process.
7. **Applications**:
    - **PCR**: Primarily used for **mutation detection**, **genetic analysis**, and **amplifying DNA templates** for further analysis (e.g., sequencing).
    - **qPCR**: Used for **quantifying DNA/RNA**, **gene expression studies**, **microbial identification**, **quantification of ancient DNA**, and **detecting pathogens** in various fields such as clinical diagnostics, food safety, and forensic analysis.
8. **Technical Principle**:
    - **PCR**: A **basic amplification technique** suitable for detecting mutations and confirming the presence of specific alleles.
    - **qPCR**: A more **advanced method** that quantifies the DNA or RNA based on the **hydrolysis** of the **probe** during amplification. This provides both qualitative and quantitative information about gene presence and abundance.
9. **Specific Variations**:
    - **RT-PCR (Reverse Transcription PCR)**: A specialized form of **qPCR** that **reverse transcribes RNA into complementary DNA (cDNA)** before quantifying it. This is used specifically for **gene expression studies**.

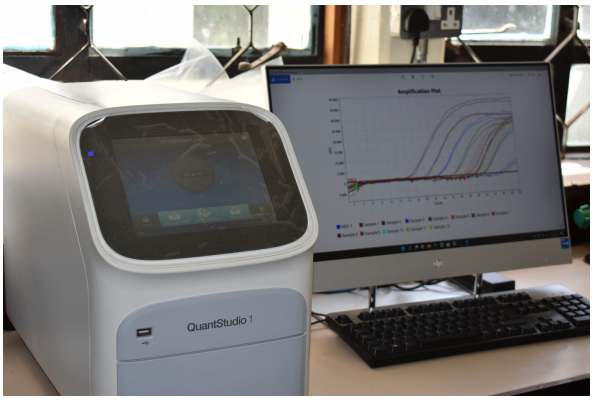
10. **Limitations and Advantages**:

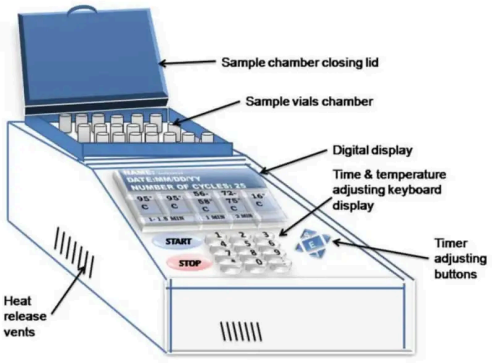
    - **PCR**: Primarily limited to qualitative detection of **monogenic mutations**, and used to prepare samples for **DNA sequencing** or **microarray analysis**.
    - **qPCR/RT-PCR**: More versatile and can be applied to **gene expression**, **mutation quantification**, **microbial identification**, and various diagnostics in medical and research settings.

In conclusion, while **PCR** is a foundational technique that provides qualitative data for gene detection, **qPCR** expands on PCR's capabilities by providing **quantitative** data, enabling it to be used in more complex analyses like **gene expression profiling** and **pathogen quantification**.

Here's the simplified comparison in a table format:Here's the updated table with DNA-based techniques and RNA-based techniques grouped together:

Technique	Process
Normal PCR	DNA is isolated → PCR amplifies the target area
RT-PCR	RNA is isolated → cDNA is made using reverse transcription (RT) → PCR amplifies the target area





Post-PCR Analysis:

- PCR reactions often generate specific bands that correspond to the amplicon of interest. The size of the band can confirm whether the PCR amplified the correct target sequence.
- PCR products, or **amplicons**, can be visualized using **agarose gel electrophoresis**, a method that separates DNA fragments based on their **size** and **charge**.
- However, **primer dimers**—undesired by-products of PCR—may appear as **smudgy bands** near the bottom of the gel. These result from primers binding

to each other rather than the target DNA, potentially interfering with the desired PCR product.

qPCR: https://www.youtube.com/watch?v=1kvy17ugI4w

qPCR vs. Gel Electrophoresis:
Unlike traditional PCR, **quantitative PCR (qPCR)** or **real-time PCR** does not require post-PCR analysis like gel electrophoresis. Instead, qPCR analyzes the DNA product **in real-time** as the amplification occurs, allowing for the monitoring of the quantity of DNA during each cycle of amplification.

	PCR	qPCR
Full form	Polymerase chain reaction	Quantitative polymerase chain reaction
Principle	Primer amplification	Either probe hydrolysis or fluorescence through intercalating dye
Chemistry	Non-fluorescence	Fluorescence
Ingredients	PCR primers, Taq DNA polymerase, PCR buffer and template DNA	Set of probes, dye, primer set, PCR buffer, template DNA, taq or reverse transcriptase enzyme
Assay set up	Reaction preparation, amplification and agarose gel electrophoresis.	Reaction preparation, amplification and real time detection.
End results	DNA bands on gel	Peak or graph of amplicons
Resolution	Low resolution amplification	High resolution
Applications	Amplification, detection of mutation	Amplification and quantification

qPCR and Fluorescent Chemistries

1. **Fluorescent Dyes**:
    - qPCR uses fluorescent chemistries to measure PCR product concentration.
    - **SYBR Green I** is the most common dye, emitting fluorescence when bound to double-stranded DNA. The fluorescence intensity increases with the PCR product concentration.
2. **Challenges with SYBR Green I**:
    - SYBR Green I binds to all double-stranded DNA, including nonspecific products like **primer dimers**.
    - Careful primer design is necessary to avoid nonspecific binding.

3. **Melt Curve**:
    - To ensure specificity, a **melt curve** is used after PCR.
    - The reaction is exposed to a temperature gradient (60°C to 95°C) to melt the DNA.
    - Fluorescence decreases as the double-stranded DNA dissociates. Specific products melt at a unique temperature, while nonspecific products like primer dimers melt at lower temperatures.
4. **Amplification Graph**:
    - qPCR data can be plotted on an **amplification graph** with cycle number (X-axis) and fluorescence (Y-axis).
    - The **threshold cycle (CT value)** is the cycle when fluorescence exceeds the background.
    - The higher the target DNA amount, the lower the CT value (detected earlier).
5. **Quantification Methods**:
    - **Absolute Quantification** measures the exact number of target molecules (e.g., viral particles in blood).
    - **Relative Quantification** compares gene expression between samples and calculates fold changes.

In short, **SYBR Green I** helps quantify PCR products, and the **melt curve** ensures specificity. **Absolute** and **relative quantification** methods are used based on the experimental need.

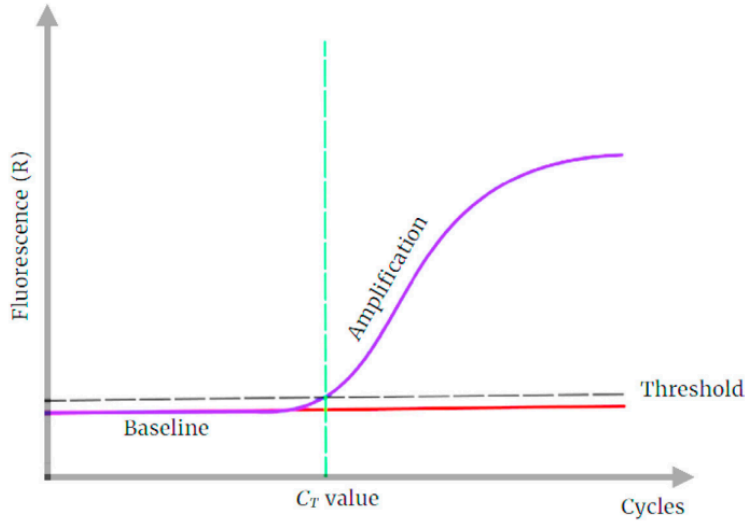
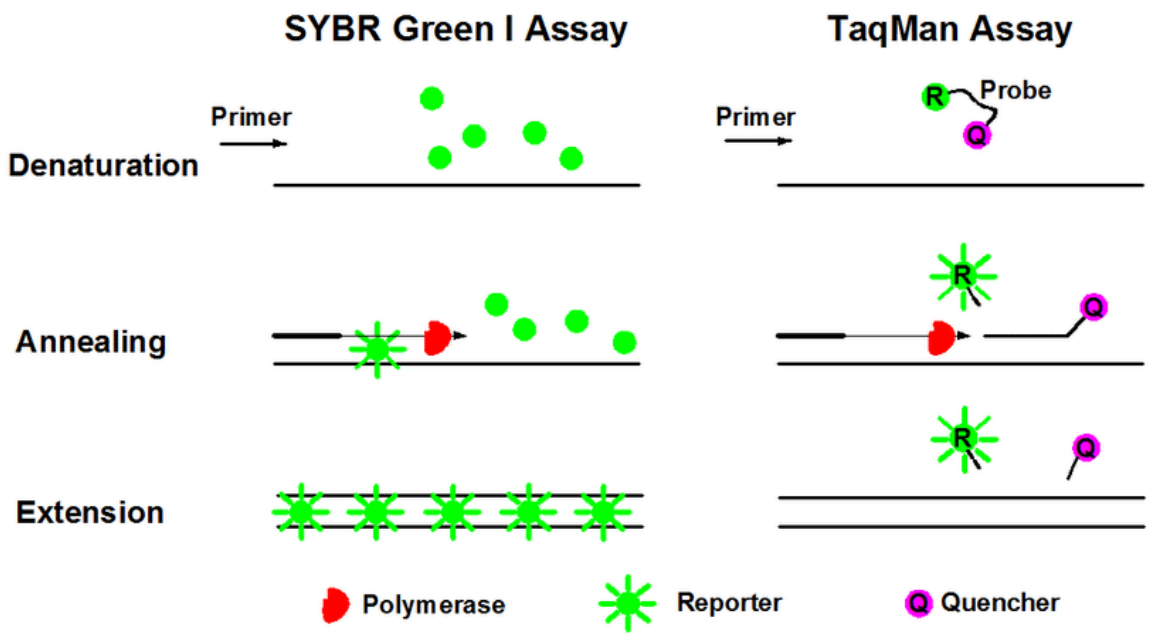


FIGURE 1.7 qPCR amplification plot. Baseline-subtracted fluorescence versus number of PCR cycles. The threshold cycle (C_T) is the cycle number at which the fluorescent signal of the reaction crosses the established threshold line.

https://www.youtube.com/watch?v=n14TYAO5N3g



PRIMER DESIGN

https://youtu.be/mcOwlFVEino?si=vTQwnC7bMluACRQw

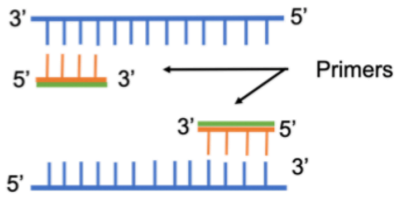
1. **Importance of Primer Design**: Proper primer design is essential for a successful PCR experiment, especially for real-time PCR (qPCR), RT-qPCR, bisulfite PCR, and methylation-specific PCR (MSP). Key factors include primer specificity, size, and amplicon length.

2. **PCR Process**: PCR amplifies a target DNA sequence by cycling through heating and cooling steps. Primers bind to the DNA and help DNA polymerase replicate the target region, doubling the DNA amount in each cycle.

3. **Challenges in PCR**: Low primer efficiency can hinder PCR. To ensure good results, the PCR product must be concentrated and pure.

4. **Solution for Purity**: Zymo Research's DNA Clean & Concentrator Kits help concentrate and clean PCR products, making them suitable for sensitive downstream applications.

5. **Primer Design Tips**: Proper primer design is crucial for effective PCR, qPCR, RT-qPCR, bisulfite PCR, and MSP.



	Normal Target Sequence
Normal Primer	3' CGTTAGACGAT. ..5' .....G CAATCTGCTA...

PCR workflows are sensitive to various factors that can affect the results. Some variables, like the sample source or the need for reverse transcription, are unavoidable. Assay design is crucial and can determine PCR success, reproducibility, and sensitivity. Here's how the process flows:

1. **Target Location**: First, choose the target sequence. This may depend on the application, like SNP detection or gene copy number.
2. **Primer Selection:** Choose the most suitable primers, and make necessary modifications. When multiplexing, consider potential primer interactions and target abundance.
3. **Difficult Cases:** For detecting low-copy targets or small differences, test multiple primer combinations with appropriate probes.

Using software like OligoArchitect simplifies assay design. The online tool offers a range of options, and more specialized designs can be requested from expert molecular biologists.

Amplicon Selection:

- The amplicon is the DNA region analyzed, defined by the forward and reverse primers.
- The size of the amplicon depends on the analysis method:
    - Gel electrophoresis: The fragment should be large enough to stain and fit within size markers.
    - Capillary electrophoresis: PCR products should range from 100 bp to 2 kb.
    - qPCR: Smaller amplicons (75–200 bp) provide accurate quantification.

Key Primer Design Guidelines:

1. **GC Content**: Aim for 40-60% GC content. Include a GC-rich 3' end (GC Clamp) to improve binding stability. Avoid excessive repeating G or C to prevent primer-dimer formation.
2. **High-Quality DNA**: Use pure, high-quality DNA to ensure reliable PCR results.
3. **Primer Length**: Keep primers between 18-30 bases long for effective binding and specificity. Shorter primers generally bind more efficiently.
4. **Melting Temperature (Tm)**: Target a Tm between 65°C and 75°C, and ensure both primers (forward and reverse) have Tms within 5°C of each other. More GC content increases Tm.
5. **Annealing Temperature (Ta)**: Set the Ta 3-5°C below the Tm for efficient primer binding.
6. **Secondary Structure**: Avoid regions that could form secondary structures (like hairpins) and ensure a balanced distribution of GC and AT regions.
7. **No Repeats**: Avoid long runs of the same base (e.g., AAAA or GGGG) and dinucleotide repeats (e.g., ATATAT). Avoid G/C repeats at the 3' end of the primer to prevent off-target binding.
8. **Avoid Homology**: Do not have complementary sequences within the primer (intra-primer homology) or between the forward and reverse primers (inter-primer homology) to prevent primer-dimers.
9. **Restriction Enzyme Sites**: Add 3-4 nucleotides 5' to the restriction enzyme site in the primer for efficient cutting.
10. **Purification for Cloning**: For cloning purposes, cartridge purification is recommended for primers.
11. **Mutagenesis Primers**: Place mismatched bases toward the middle of the primer for optimal mutagenesis.
12. **TOPO Cloning**: For TOPO cloning, avoid phosphate modifications on the primers.
13. **Amplicon Length**: Keep the amplicon length between 70-140 bp for efficient primers and a probe (for qPCR).
14. **Exon/Exon Junctions:** Users can specify that primers span exon-exon junctions with adjustable bases on either side.
15. **Intron Spanning:** Primers can also span introns, with options to specify intron sizes.

16. **RefSeq Requirement:** These features require a RefSeq accession since it offers accurate exon/intron boundary annotations.

These guidelines will help ensure your primers are effective, specific, and free from issues that can affect your PCR results.

TaqMan® Probe Design Tips:

1. **Probe Tm**: Keep probe Tm 4-8°C higher than the primers for the best specificity.
2. **Probe Length**: Make probes 20-25 base pairs long for stability.
3. **No Overlap**: Ensure the primer and probe binding sites don't overlap.
4. **Avoid Guanine at 5' End**: Guanine can reduce probe signal.

General Primer/Probe Considerations:

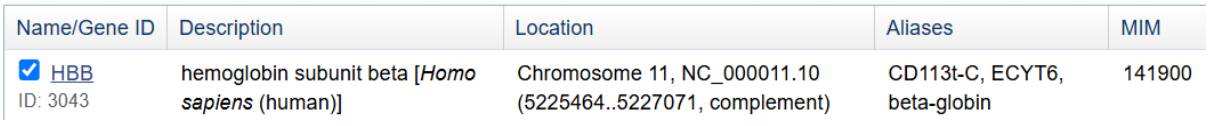
1. **Avoid SNPs**: Check for common SNPs that could interfere with primer or probe binding.
2. **Avoid Hairpins/Dimers**: Ensure primers and probes don't form hairpins or dimers, especially at the 3' end.
3. **Specificity Check**: Use tools like NCBI BLAST to check primer specificity against the genome.

RT-qPCR Primer Tips:

1. **Design Over Exon-Exon Junctions**: To avoid genomic DNA interference, design primers over exon-exon junctions to ensure specificity for mRNA.

How to design a primer online

Primers are often designed to span exon-exon junctions in RT-qPCR (quantitative reverse transcription PCR) to specifically target cDNA, reducing the risk of amplifying contaminating genomic DNA, which contains introns.

1. https://www.ncbi.nlm.nih.gov/
2. **The HBB gene in humans** codes for the beta-globin protein, which is a key component of hemoglobin, the protein responsible for carrying oxygen in red blood cells; essentially, mutations in the HBB gene can lead to genetic disorders like sickle cell anemia and beta-thalassemia as it affects the structure of the hemoglobin molecule by altering the beta-globin chain.
3. 

Name/Gene ID	Description	Location	Aliases	MIM
☑ HBB ID: 3043	hemoglobin subunit beta [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5225464..5227071, complement)	CD113t-C, ECYT6, beta-globin	141900

4. https://www.ncbi.nlm.nih.gov/nuccore/NC_000011.10?report=fasta&from=5225464&to=5227071&strand=true

```
>NC_000011.10:c5227071-5225464 Homo sapiens chromosome 11, GRCh38.p14 Primary Assembly
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAG
ACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGG
TGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT
TCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAG
GAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCT
CAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
```

TTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA
TCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAAT
ATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAAT
CATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAA
TTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATA
CTTTCCCTAATCTCTTTCTTTCAGGGCAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAG
AATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCTCTGCATATAAATATTTCTGCATATAAAT
TGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTGCTTTTATTTT
ATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTT
ATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAA

5. Pick Primers

Download primer pairs

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CTGTCTCCACATGCCCAGTT	Plus	20	5226867	5226886	59.96	55.00	4.00	0.00
Reverse primer	AGAAGTCTGCCGTTACTGCC	Minus	20	5226999	5226980	60.04	55.00	5.00	3.00
Product length	133								

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CCAGGCCATCACTAAAGGCA	Plus	20	5226663	5226682	60.03	55.00	5.00	3.00
Reverse primer	CTGGGCATGTGGAGACAGAG	Minus	20	5226884	5226865	60.11	60.00	4.00	0.00
Product length	222								

Primer pair 3

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CTCTGTCTCCACATGCCCAG	Plus	20	5226865	5226884	60.11	60.00	4.00	2.00
Reverse primer	ACCATGGTGCATCTGACTCC	Minus	20	5227024	5227005	59.75	55.00	8.00	2.00
Product length	160								

Primer pair #	Forward primer Sequence (5'->3')	Reverse primer Sequence (5'->3')
1	CTGTCTCCACATGCCCAGTT	AGAAGTCTGCCGTTACTGCC
2	CCAGGCCATCACTAAAGGCA	CTGGGCATGTGGAGACAGAG
3	CTCTGTCTCCACATGCCCAG	ACCATGGTGCATCTGACTCC
4	CAAGGGTAGACCACCAGCAG	TGGGCAGGTTGGTATCAAGG
5	AGCCTTCACCTTAGGGTTGC	AAACTGGGCATGTGGAGACA
6	CCTCTGGGTCCAAGGGTAGATGGGCATGTGGAGACAGAGA	
7	CCTTGATACCAACCTGCCCA	CCATGGTGCATCTGACTCCT
8	ACCTTGATACCAACCTGCCC	CTGAGGAGAAGTCTGCCGTT
9	GCACTTTCTTGCCATGAGCC	ACTGGGCATGTGGAGACAGA
10	TGTCTCCACATGCCCAGTTT	GGCAAGGTGAACGTGGATGA

Tool for Forward/ Reverse compliments:
https://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html

6. https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi
7. https://www.bioinformatics.org/sms2/pcr_primer_stats.html **(To verify the primers)**

```
General properties:
-------------------
                Primer name: forward
            Primer sequence: CTGTCTCCACATGCCCAGTT
            Sequence length: 20
                Base counts: G=3; A=3; T=6; C=8; Other=0;
             GC content (%): 55.00
  Molecular weight (Daltons): 6003.94
                  nmol/A260: 5.69
            micrograms/A260: 34.19
       Basic Tm (degrees C): 54
 Salt adjusted Tm (degrees C): 49
Nearest neighbor Tm (degrees C): 65.03
PCR suitability tests (Pass / Warning):
-----------------------------------
           Single base runs: Pass
      Dinucleotide base runs: Pass
                     Length: Pass
                 Percent GC: Pass
       Tm (Nearest neighbor): Warning: Tm is greater than 58;
                   GC clamp: Pass
              Self-annealing: Pass
          Hairpin formation: Pass
------------------------------------------------------------
General properties:
-------------------
                Primer name: reverse
            Primer sequence: AGAAGTCTGCCGTTACTGCC
            Sequence length: 20
                Base counts: G=5; A=4; T=5; C=6; Other=0;
             GC content (%): 55.00
  Molecular weight (Daltons): 6093.01
                  nmol/A260: 5.35
            micrograms/A260: 32.62
       Basic Tm (degrees C): 54
 Salt adjusted Tm (degrees C): 49
Nearest neighbor Tm (degrees C): 64.71
PCR suitability tests (Pass / Warning):
-----------------------------------
           Single base runs: Pass
      Dinucleotide base runs: Pass
                     Length: Pass
                 Percent GC: Pass
       Tm (Nearest neighbor): Warning: Tm is greater than 58;
                   GC clamp: Warning: There are more than 3 G's or C's in the last 5 bases;
              Self-annealing: Pass
          Hairpin formation: Pass
```

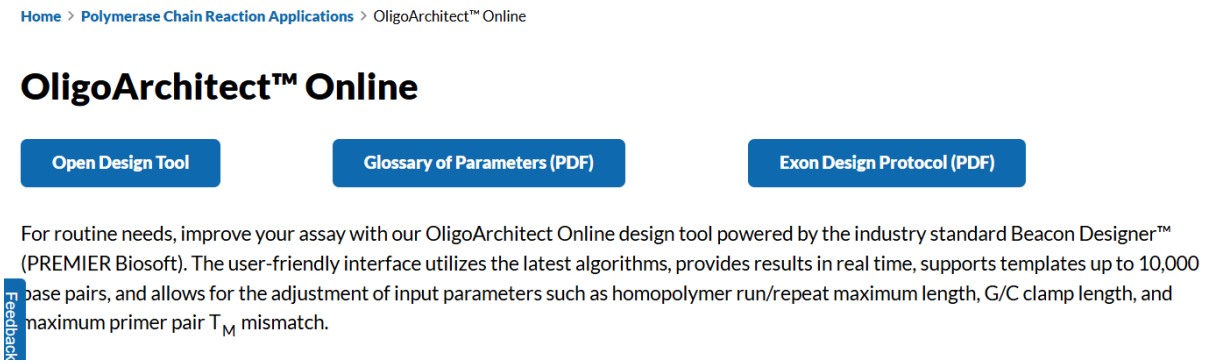
In primer design, GC content refers to the percentage of guanine (G) and cytosine (C) bases in the primer sequence, and it's generally recommended to aim for a GC content between 40% and 60% for optimal primer performance.

In primer design, "Tm" refers to the melting temperature, the temperature at which half of a DNA duplex dissociates into single strands, and it's crucial for PCR efficiency and specificity. Primers with Tm values between 52-58°C generally produce better results, and the Tm of forward and reverse primers should ideally be within 5°C of each other.

A GC clamp refers to the presence of guanine (G) or cytosine (C) bases within the last few nucleotides (typically 3-5) at the 3' end of a primer.

https://www.sigmaaldrich.com/IN/en/technical-documents/technical-article/genomics/pcr/oligoarchitect-online

To design Probes
https://www.idtdna.com/PrimerQuest/Home/Details/0_0



PCR products can be sequenced to identify the DNA sequence.

I. Simplified FASTA Format for Nucleotide Sequences

1. FASTA Definition Line:

    - Starts with ">" (greater-than symbol). Example: >SeqABCD
    - Followed by a unique sequence identifier (SeqID) (max 25 characters, no spaces). Required: Scientific name in `[organism=Name]` format (Modifiers).
    - Optional: Additional details like strain, chromosome, or isolate (e.g., `[strain=C57BL/6]`). No spaces around the "=" sign.
    - Allowed characters: letters, digits, hyphens (-), underscores (_), periods (.), colons (:), asterisks (*), and number signs (#).

      >SeqABCD [organism=Mus musculus] [strain=C57BL/6]

    - The database will replace SeqID with an Accession number upon submission.

      A brief description of the sequence (e.g., gene name, mRNA type).
      >SeqABCD [organism=Mus musculus] [strain=C57BL/6] Mus musculus neuropilin 1 (Nrp1) mRNA, complete cds.

2. Nucleotide Sequence Rules:

    - Starts on the next line after the definition line.
    - Can contain returns (line breaks), but each line should be ≤80 characters.
    - Use IUPAC nucleotide symbols (A, T, G, C, and N for unknown bases).
    - Do not use "?" or "-" (except in alignments).

>DNA.new [organism=Homo sapiens] [chromosome=17] [map=17q21] [moltype=mRNA] Homo sapiens BRCA1 mRNA, complete cds.

ATGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGC

GCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTA

This format ensures your sequence is correctly processed in databases.

BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGETYPE=BLASTHome

BLAST QuickStart: Simplified Guide

What is BLAST?

BLAST (**Basic Local Alignment Search Tool**) finds similar regions between nucleotide or protein sequences. It compares a query sequence to a database and calculates the significance of matches. This chapter introduces BLAST and explains how to use different BLAST programs with step-by-step tutorials.

---

1. Introduction to BLAST

BLAST aligns a **query** sequence (your input) with **subject** sequences (in a database). It **originally worked for proteins but later expanded to nucleotides and even cross-comparisons between them.**

- **Web & Standalone Versions**: Available at NCBI.
- **Genome Searches**: BLAST can scan entire genomes, including human, mouse, rat, and plants like *Arabidopsis thaliana*.

1.1 Query and Database Formats

- Sequences use **FASTA format** (starts with > followed by an identifier).
- BLAST databases are built from FASTA sequences using **formatdb**.

1.2 Scoring Alignments

- Alignments pair letters (nucleotides or amino acids) between sequences.
- **Scoring:**
  - **Protein matches** use **substitution matrices** (e.g., *BLOSUM62*, *PAM*).
  - **Nucleotide matches**:
    - Identical letters: +2 points
    - Mismatches: -3 points
  - **Gaps**: Have penalties, with new gaps costing more than extensions.

1.3 How BLAST Works

- **Indexing:** BLAST breaks the query sequence into small words (default: **3 for proteins, 11 for nucleotides**).
- **Search:** It scans the database for exact (nucleotides) or high-scoring (proteins) matches.

- **Extension:** Matches are extended until scores stop increasing or drop too much (*dropoff* value).

1.4 Statistical Significance

- BLAST assigns an **Expect Value (E-value)** to matches:
    - **Higher E-value**: More matches, but lower accuracy.
    - **Lower E-value** (0.001 to 0.0000001): More reliable alignments.
    - **Default = 10** (ensures no important match is missed).

The E-value is not a direct probability, but it's closely related to the probability of finding a match with a given score by chance. A lower E-value corresponds to a lower probability of a random match.

A good p-value is generally considered to be less than 0.05, often written as $p < 0.05$. This means that there's a less than 5% chance that the observed results are due to random chance alone if the null hypothesis is true.

BLAST (**Basic Local Alignment Search Tool**) is a widely used bioinformatics tool for **comparing biological sequences.** Since its release in 1990, it has been continuously updated to improve speed and accuracy. It plays a crucial role in research and has inspired other sequence comparison tools.

Types of BLAST

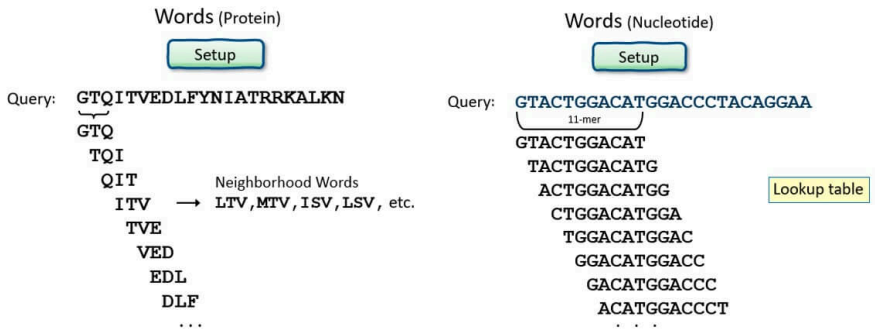
BLAST has five main types, depending on whether the query and database sequences are DNA or protein:

1. **BLASTN** – Compares a **nucleotide** query to a **nucleotide** database.
2. **BLASTP** – Compares a **protein** query to a **protein** database.
3. **BLASTX** – Translates a **nucleotide** query into six protein reading frames and compares it to a **protein** database.
4. **TBLASTN** – Translates a **nucleotide** database into six protein reading frames and compares it to a **protein** query.
5. **TBLASTX** – Translates both the **nucleotide** query and **nucleotide** database into six protein reading frames and compares them.

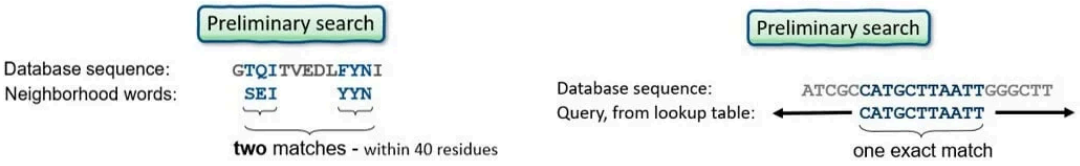
BLAST compares a query sequence to a database to find similarities using a fast, heuristic approach.

Steps in BLAST Alignment

1. **Seeding** – The query sequence is broken into short segments (**words**) for searching:
    - **Protein:** 3 amino acids per word
    - **DNA:** 11 nucleotides per word



2. **Searching** – The database is scanned for sequences containing matching words.



3. **Scoring** – Matches are scored using **substitution matrices**:
    - **Protein:** Uses **PAM or BLOSUM matrices.**
    - **Nucleotide:** Uses match-mismatch scoring.
    - Matches above a certain threshold are considered significant.
4. **Alignment Extension** – Matching words are extended in both directions while tracking alignment scores.
    - If the score drops below a threshold, extension stops.
    - The highest-scoring segment pair (HSP) is recorded.
5. **Statistical Significance (E-value)** – BLAST calculates the **Expect Value (E-value)**, which represents the probability of a random match:
    - **Lower E-value** = More significant match.
    - **Higher E-value** = More likely a random match.

Key Features of BLAST

- **Fast & Efficient** – Handles large databases quickly.
- **Versatile** – Works with both nucleotide and protein sequences.
- **Highly Sensitive** – Detects even small sequence similarities.
- **Local Alignment** – Focuses on similar regions rather than full-sequence alignment.
- **User-Friendly** – Easy to input sequences and interpret results.

Applications of BLAST

- **Identifying Unknown Sequences** – Compares sequences to known databases to predict gene or protein function.
- **Phylogenetic Analysis** – Helps determine evolutionary relationships between species.
- **Detecting Conserved Protein Domains** – Finds functional regions within proteins.

BLAST is an essential tool for genetic research, evolution studies, and functional genomics.

ATGCGCCTCCATCCTCGCCTGCCTTCTCTCGGTCCCTCGTGATTGATTCCACCCTTGCTTCCCCTTTC

TCCGCGCCGCCTGTTCCGTGCTCGTTTTCCCCTCTTCCTCCTTAAGTCTGGTCCTTCCACCCCCTCCTCT

TCAAGCTGTGCGTGTCCCCTGATTCTAATGCTTTCTGTGTAACTCATTGAAACTGCGTTCTGGTTCCCCT

CCCGCGTCCATTCTCCATTCATGCGCGACCGCCCTTCCCGCGCCCCAGTTCCCTCTCGCCGCCCCTCCCC

CTGCTTGCTGGTCACGTCCGCTCCCCCGCATCCCCTTCCTCGCCTGGCGTGTCCGCTCCCTCCCTCCCTC

TCTGCTCTGGTCGCGCCCGCCCACTTGCTCCGGTCTCCGAGCGCGGTCCCACCCCCCTTTTCCATCACCG

CCTCCCAGCTCCCAGCAGGCTCGGGCGGTGCTGAGGCCCCGTGTCCGGGGCGGGCGGGCGGAGGGCTGGG

CTGGGTGCCCCGCGCGGCGGGCGGGATGCGGCGGCCCGGGGGCAGCTGGAGACTTACGTAACGTTGGCCT

GCCCCGCTGCCGGGAGGCAGGGCGGTTGCCCCTGCGGCGCGGTGCCGTCCCCTGTGGCCGGGATTAGATG

GGCGGCCTGCGAGGGCCTGGGAATGGCTCGGGGCCCGAGAGCTCGACCGGCCCTTGCCGGGTGGCCGCCA

GGGACCACGCTCCATCTCGCCGCGGCCGGGCTGCACGTAGCGGCCGCGCCCAGGGCCCACCCCGCTTCAC

CGGGCGATGGCCTTGGGCCTCGTAACGGGCGGGATAAACCTCTGCAGGCTTTGCTGGGGCCTCCTGGCCC

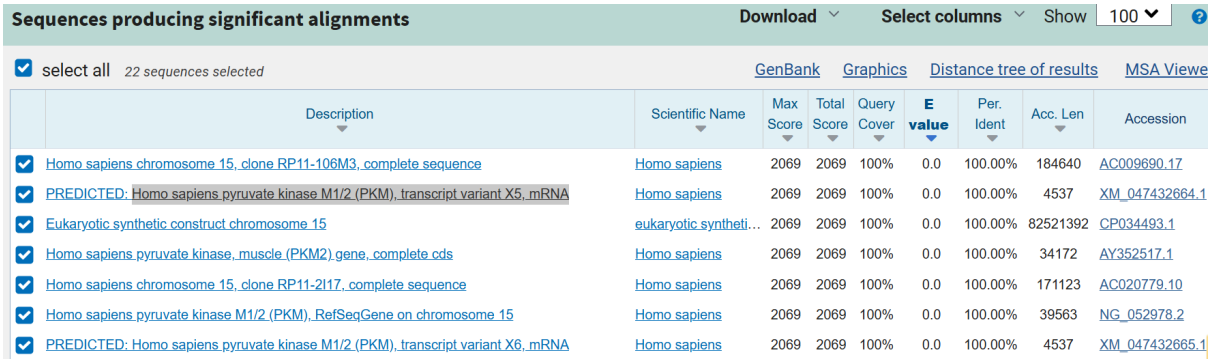
TCGCCCGTTCCGCGCCTCTCCGGCACGTTTCTTTCTTTTCCTTTCTTTTTTCTTTCTTTTTTTTTTTTTT

AATTATCCGGCACATTTTTTAACAAATGCGTCCTGATTGTGGAACGCGGAGGCCGCGGGGTGGGGTGGGG

ATCTGGTTACGGAGGGGGCAGGAAATCTGTCGCGTTCACTGAACGCAAACGGTGTGGGTCAAGGGCTGTT

TGGGGGTCAGAGTTAGAGACCAGGATGACTAGACGAGTCATAGCCCACCGAGCTCACAATCTTAAAATGT

ATCTCCTGTAATGCTGGGAGTGGGGTACGAGCTTCCTGCTGTGGGAGGGAGGGGGACAGGAAGCCTCGTA



Sequences producing significant alignments Download Select columns Show 100

select all *22 sequences selected* GenBank Graphics Distance tree of results MSA Viewe

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens chromosome 15, clone RP11-106M3, complete sequence	Homo sapiens	2069	2069	100%	0.0	100.00%	184640	AC009690.17
PREDICTED: Homo sapiens pyruvate kinase M1/2 (PKM), transcript variant X5, mRNA	Homo sapiens	2069	2069	100%	0.0	100.00%	4537	XM_047432664.1
Eukaryotic synthetic construct chromosome 15	eukaryotic syntheti…	2069	2069	100%	0.0	100.00%	82521392	CP034493.1
Homo sapiens pyruvate kinase, muscle (PKM2) gene, complete cds	Homo sapiens	2069	2069	100%	0.0	100.00%	34172	AY352517.1
Homo sapiens chromosome 15, clone RP11-2I17, complete sequence	Homo sapiens	2069	2069	100%	0.0	100.00%	171123	AC020779.10
Homo sapiens pyruvate kinase M1/2 (PKM), RefSeqGene on chromosome 15	Homo sapiens	2069	2069	100%	0.0	100.00%	39563	NG_052978.2
PREDICTED: Homo sapiens pyruvate kinase M1/2 (PKM), transcript variant X6, mRNA	Homo sapiens	2069	2069	100%	0.0	100.00%	4537	XM_047432665.1

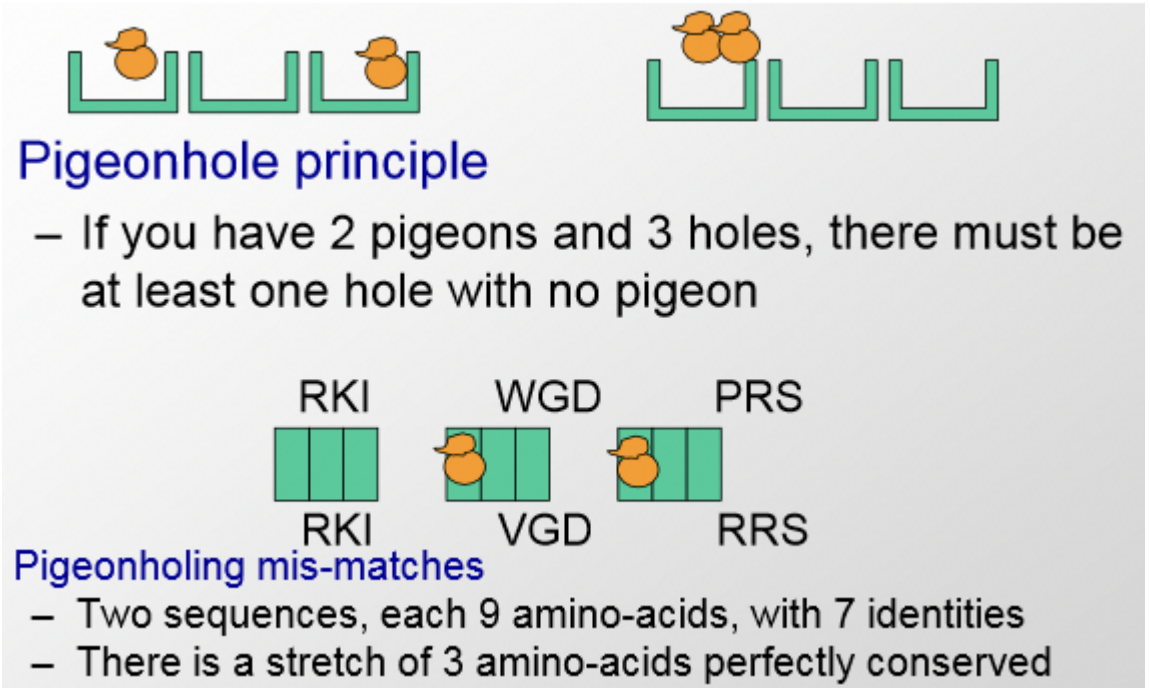
In Homo sapiens, pyruvate kinase M1/2 (PKM) is encoded by the PKM gene, and transcript variant X5, which encodes the M2 isoform, is a key enzyme in glycolysis, catalyzing the conversion of phosphoenolpyruvate to pyruvate.

Simplified Overview of the BLAST Algorithm:

BLAST (Basic Local Alignment Search Tool) is used to search a large database of biological sequences (like DNA or proteins) for similar sequences to a given query sequence. It helps identify related sequences even if they are not perfectly aligned, focusing on finding regions with significant similarity.

Key Concepts:

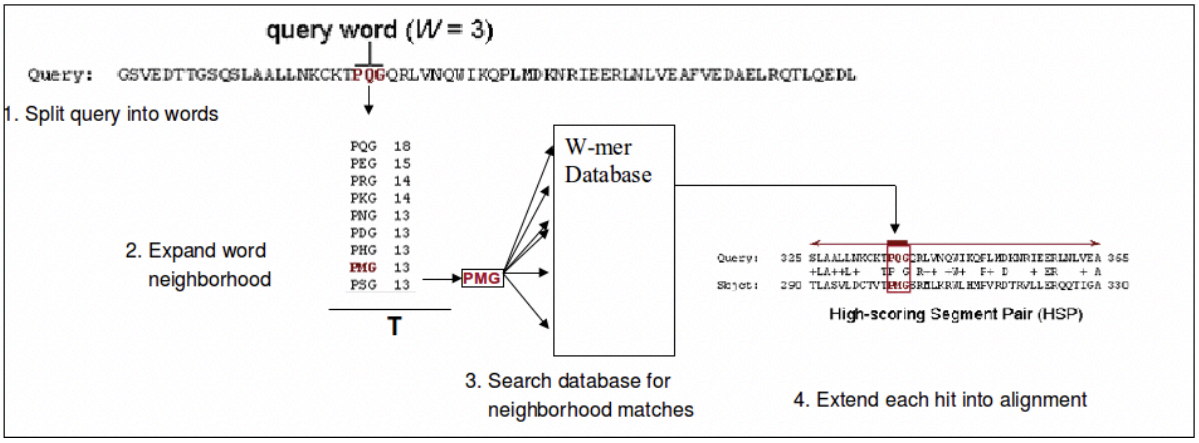
1. **Query and Target Sequences**:
    - **Query**: A new sequence you're looking for matches for.
    - **Target**: A set of older sequences that may match the query.
2. **Perfect Alignments and BLAST**:
    - Perfect alignments (matches) are not required. BLAST looks for initial "seeds" or regions of similarity and then refines the match.
    - It's more efficient because many sequences in the database won't match the query, so BLAST narrows the search by looking for stretches of matching nucleotides (DNA bases).
3. **Pigeonhole Principle**:
    - This principle helps in BLAST. If a sequence matches, it will have several consecutive nucleotides that match. This makes searching more efficient by focusing on those regions first.



4. **W-mers and Pre-screening**:
    - BLAST breaks up the query sequence into small subsequences (called **W-mers**). It searches for these W-mers in the target sequences, looking for matches or similar sequences, even if they are not exact.
5. **Hashing**:
    - BLAST uses a **hash table** to quickly look up where each W-mer occurs in the database. This speeds up the process of finding potential matches.
6. **Refining Matches**:
    - Once potential matches (seeds) are found, BLAST extends them to find better alignments. The extensions stop when the match score falls below a certain threshold.

BLAST Workflow:

1. **Break the query sequence into overlapping W-mers**.
2. **Find similar sequences (neighborhood)** by slightly modifying each W-mer and searching for these variations in the database.
3. **Look up these W-mers in a hash table** to find initial matching locations (seeds).
4. **Extend the seeds** to find the best match, stopping when the score is low enough.



Key Parameters in BLAST:

1. **W (word length)**:
    - Larger W means fewer, stronger matches, speeding up the process, but also may miss some sequences.
    - Smaller W gives more hits but can be slower because it might find irrelevant matches.
2. **T (threshold for similarity)**:
    - A higher T speeds up the search but might miss distant relationships between sequences.
3. **X (extension score threshold)**:
    - Controls the strictness of when the alignment stops. A higher X may avoid unnecessary searches, while a lower X may lead to more thorough searches.

Extensions to BLAST:

1. **Low Complexity Filtering**:
    - Repetitive or simple sequences (like many G's or A's in a row) can give too many irrelevant hits. These can be filtered out to avoid wasting time.
2. **Two-hit BLAST**:
    - Instead of looking for one long W-mer match, this method looks for two shorter W-mers, improving sensitivity without sacrificing speed.
3. **Combs**:

- This method uses non-consecutive nucleotides from the sequence for searching. Some nucleotides, like the third nucleotide in a codon, do not affect the protein outcome, so they can be ignored during the search.

4. **PSI-BLAST (Position-Specific Iterative BLAST)**:

    - This version of BLAST iteratively updates a scoring matrix based on multiple sequence alignments, allowing detection of more distantly related sequences.

Conclusion:

BLAST is an efficient algorithm for finding related sequences in a large database, leveraging techniques like hashing, neighborhood search, and parameter tuning to balance sensitivity and speed. It works by identifying small matching subsequences (W-mers) and refining them to find the best possible match while avoiding irrelevant data.