

## Understanding DNA sequence notation

When writing DNA sequences, there are a few universal rules and shortcuts that scientists use to keep data clear and consistent. Whether it's working on PCR primers, analyzing gene sequences, or reading bioinformatics data, this summary breaks down and guides a better understanding of the language of DNA.

### Always write 5' to 3'

DNA sequences are ALWAYS written from the 5' (five prime) end to the 3' (three prime). That's the direction DNA polymerase reads and builds from, so keeping this standard helps everyone stay on the same page—literally!

For example: Instead of saying “tagcatcg,” you mean “5'-TAGCATCG-3”.

Forward vs. Reverse Strand

DNA has two strands, each of which is the reverse complement of the other.

The forward strand (sense or “+” strand) is the one that matches the RNA transcript.

The reverse strand (antisense or “-“strand) is complementary to the forward strand.

### IUPAC codes

When you are unsure of which base is present, or multiple versions exist, that is when IUPAC codes come in. These are single letters that represent ambiguous positions in a DNA sequence.

For example,

R = A or G

Y = C or T

N = Any base (A, T, C, or G)

W = A or T

S = G or C

A sequence like ACRYGNT provides flexibility and is beneficial when designing primers or aligning variable regions.

These codes pop up all the time, especially when:

- Designing primers
- Describing restrictions on enzyme recognition sites
- Showing codon families

## Complementing Ambiguity

Each IUPAC code also has a complement, which is what you'd get if you reverse the strand:

- R (A/G)  $\leftrightarrow$  Y (T/C)
- S (G/C)  $\leftrightarrow$  S (G/C)
- N  $\leftrightarrow$  N (no change)

This makes it easier to work with both DNA strands and stay consistent.