As far as I have understood, when developing a pipeline, there are many different things you can work with:

-different platforms

- different assembly, mapping, and basepair calling

Understanding pros and cons of each platform

Example: illumina struggles to read repetitive regions in sample sequences, whereas nanopore platforms can read repetitive regions better

- different primers (V3 or V4) - amplicon or whole genome sequencing

- different read length (short reads and long reads, look up each platform to see how much reads they can seq)

- different paired vs single reads (R1, R2, 1 vs 2 fasta file)

Paired end reads = R1 and R2 files, you have to make sure that in the pipeline, you can get R1 to properly pair with R2 reads

Single reads = you only have one fastq files to work with, so you don’t have to worry about proper alignment with R2.

* Different errors (indel errors)
  + Phred scores – nanopore – longer reads, lose quality, use Qscore of 12 in pipelines for higher coverage and allow the reads to correct themselves. Illumina – use qscore of 30 and that would give you a base calling of 99.9%
* Different file types from different platforms
  + Illumina produces fastq files
  + Nanopore produces fast5 files
  + Knowing how to convert these file types to workable files (fastq) files

-different sample types

- DNA

- RNA (which you will need to convert to cDNA)

- protein

- virus vs bacteria (virus 🡪 ssDNA, dsDNA, ssRNA, dsRNA….etc, bacteria 🡪 DNA)

- ancestry 🡪 tracing DNA, extracting human DNA

- depends on what information you want to extract out of these samples

Essentially, the output of the file or the goal is to end up with a fastq file (workable file)

Reading lots of literature and following to see what people are saying on social media (slack) and what they use