FASTA files - Answers

1. Main parts of FASTA file?

FASTA format has two main parts:
- heading line of each sequence, starting with “>” character, followed by specimen identification number and species. Fields are usually, bu not always, divided by “|” character. e.g. >gi|2328195|ref|NM_000878.2| Homo Sapiens Interlukin 2
- the nucleotide sequence

2. Difference between hg37 and hg38?

Both are human genome assembled by GRC (Genome Reference Consortium).
Hg is released in 2009, and hg38 in 2013.
New in hg38:
- updated annotations
- repair of incorrect reads
- addition of alternate loci
- inclusion of model centromere sequences (replacing around 3 million gaps)
- some misassembled areas retiled

3. What are NNNN regions?

N - any of the A, C, T or G (base identity could not be established) - parts of the reference genome where the sequence is not known yet

4. What are _alt chromosomes?

The _alt chromosomes are alternative sequences that differ from the reference genome. These are regions of the genome that exhibit sufficient variability to prevent adequate representation by a single sequence.

5. What are _random chromosomes?

_random chromosome - unlocalized sequence are on a specific chromosome but with unknown order or orientation.