3. Gene prediction

- All genes end on a stop codon
- A simple algorithm for gene prediction
- Searching for start and stop codons
- Predicting all the genes in a sequence
- Making the predictions more reliable
- Boyer-Moore algorithm
- Index and suffix trees
- Probabilistic methods
- **Benchmarking the prediction methods**
- Gene prediction in eukaryotic genomes

The need for a reference

- A well annotated genome
  - for which predictions have been confirmed by experimental results
  - for instance, an *E. coli* genome

In practice, very few genomes, if any, have been fully experimentally confirmed

Comparing the two gene lists

- Apply the gene predictor on the reference sequence
- Compare the position for start and stop codons

- True positives (TP): the genes which are predicted and are confirmed on the reference genome
- False positives (FP): genes which are predicted, but are not found on the reference genome
- False negatives (FN): genes which are not predicted, but do exist on the reference genome

Sensitivity and precision

Sensitivity = TP / (TP + FN)

Precision = TP / (TP + FP)

Example

- Our gene predictor on the well studied *B. subtilis* genome

- Predicts correctly 3,500 genes out of 4,100 expected
- But produces also 1,200 false positives