QC metrics for DNA sequencing

Outline

- Metrics
- Tools
- Correlations
  - Plots
- Quiz

Learning outcomes and course content

- Learning outcomes:
  - Understand how to apply technology to obtain relevant information from the cancer genome.
  - Perform quality control on DNA (and RNA) sequencing data for cancer sequencing purposes.

- Course content:
  - QC of both DNA (and RNA) sequencing data

- Focus on DNA here, RNA QC is covered in RNA lecture

Quality control metrics

- Was the sequencing successful or not?
- Many steps can go wrong
  - Extraction of DNA, DNA input amount
  - Library prep, e.g. PCR amplification
  - Capture
  - Sequencing
  - Demultiplexing
- Important with quality control metrics
- Is data quality “good enough”?
  - Requirements can vary a lot depending on the experiment

Quality control metrics

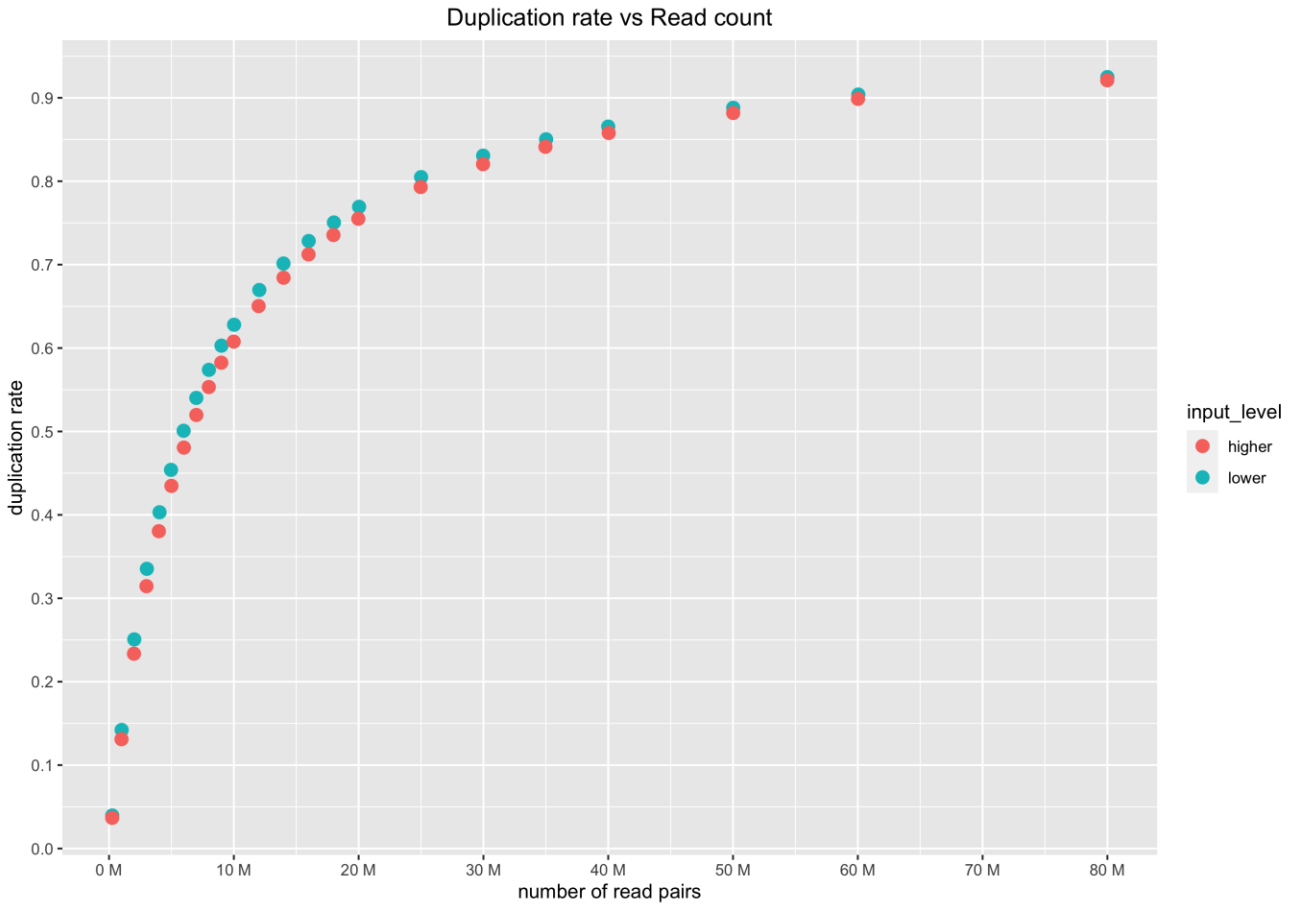
- The most important metrics
  - **Coverage** (after deduplication) – the average number of (unique) reads covering the targeted regions, also known as depth
  - **Read count**– total number of reads for a sample
  - **Duplication rate** – what fraction of all reads where duplicates (not unique)
  - **Fold enrichment** – how much more the targeted regions are amplified compared to non-targeted regions, x-fold
  - **Contamination** – DNA from another source

Quality control metrics

- Example of tools for QC
  - Picard CollectHsMetrics
    - Coverage, fold enrichment
  - Picard MarkDuplicates
    - Read count, duplication rate
  - GATK (v ≤3) ContEst
    - Contamination
- Input: bam files
- Output: txt tables,
    - can be parsed in e.g. R for plotting, summary tables etc.

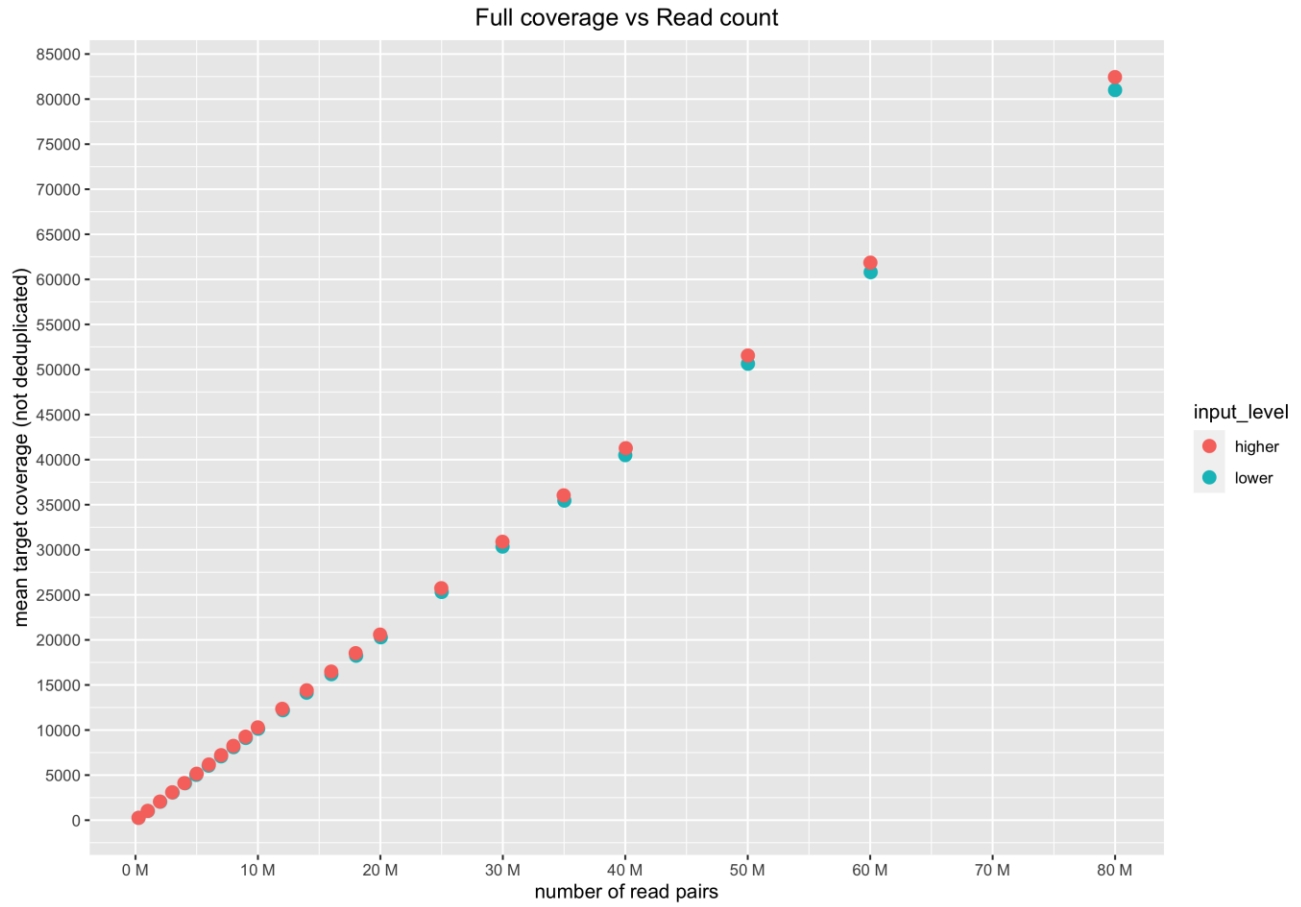
Duplication rate vs read count

- Increasing number of reads give increasing duplication rate
- Small difference in DNA input amount (few ng)
- Higher input amount gives lower duplication rate



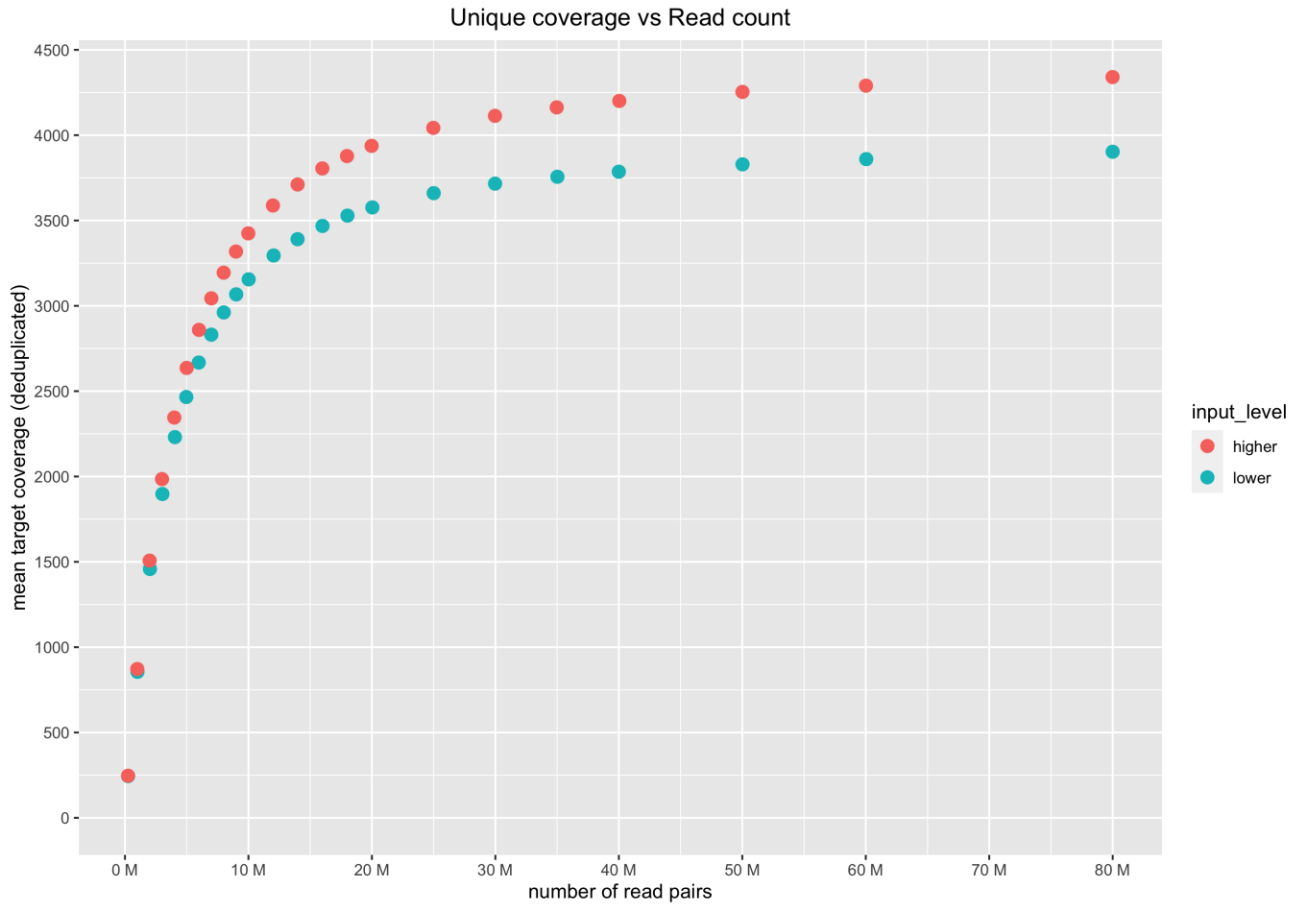
Coverage vs read count

- Including duplicates
  - Not de-duplicated
- Increasing number of reads give higher coverage
- No difference by input amount



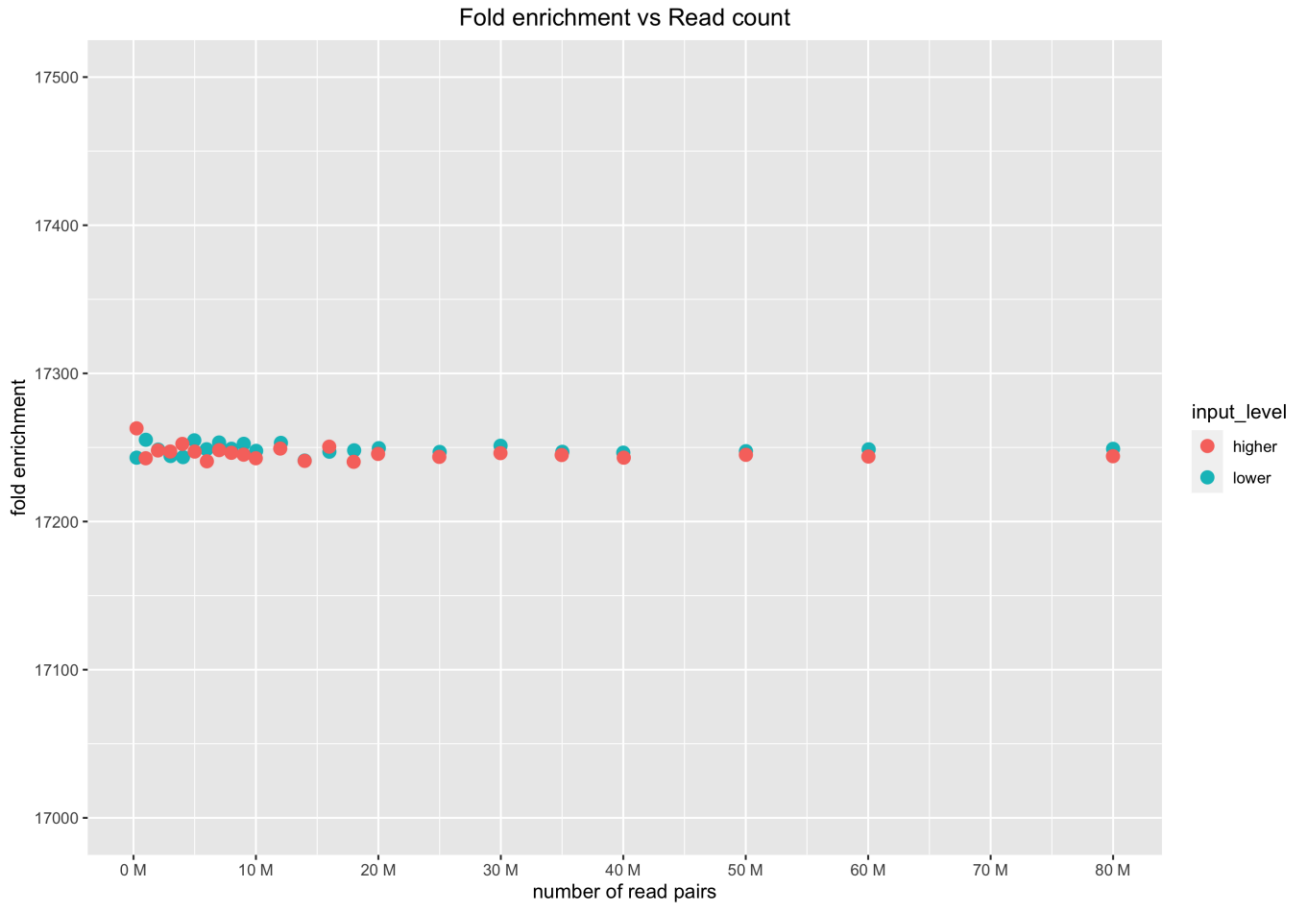
Coverage vs read count

- Unique
  - De-duplicated
- Increasing number of reads give higher coverage
  - Up until a certain level
- Higher input amount gives higher unique coverage
- The max coverage possible depends on input amount
- When max coverage is reached, all unique DNA molecules have been sequenced and all additional reads will be duplicates



Fold enrichment vs read count

- Before de-duplication
- Independent of read count and input amount
- Related to the size of targeted regions
- What could be the cause of a deviating number?



Quality control metrics – quiz

1. How does the DNA input amount relate to the max coverage possible?
   a) Increased max coverage with decreased input amount
   b) Decreased max coverage with decreased input amount
   c) No correlation between them
2. How does this relate to duplication rate?
   a) Increased duplication rate with decreased input amount
   b) Decreased duplication rate with decreased input amount
   c) No correlation between them
3. What could be the cause of a deviating fold enrichment?
   a) Too low DNA input amount
   b) Failed capture
   c) Failure on the sequencer
4. How can these metrics help in answering “is it good enough”?
   a) Was the min required coverage reached?
   b) Did capture work, so we got data for the targeted regions?
   c) Was the DNA input amount high enough?

Quality control metrics – quiz answers

1. How does the input amount relate to the max coverage possible?
   b) Lower input → fewer unique molecules → lower max unique coverage
2. How does this relate to duplication rate?
   a) Lower input → fewer unique molecules → higher duplication rate for same number of reads
3. What could be the cause of a deviating fold enrichment?
   b) Failure in capture step
      - Lower fold enrichment than expected → more reads in non-targeted regions
4. How can these metrics help in answering "is it good enough"?
   a) Certain min coverage required, if less → not good
   b) Much lower fold enrichment → capture failure → not good
   c) Low coverage and low duplication rate (~ <60%) → not all unique molecules sequenced →sequencing more will give higher coverage

Credits

- Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia
  - Lecture notes from the course “Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)”, 29 October – 2 November, 2018
  - McDonell Genome Institute, Washington University of St Louis School of Medicine

Questions?