Sequencing coverage a single uniform population

Sequencing Coverage (Depth)

Lander/Waterman model (random reads from random genome)

Coverage \( C = \frac{N \times L}{G} \)

- \( N \) = number of reads, \( L \) = length of reads, \( G \) = size of genome
- Probability of a base not being sequenced
  \( P = \exp(-C) \)
- Example: \( N=35\) million, \( L=35 \), \( G=250 \) million
  \( C = 5 \rightarrow 0.6\% \) genome not sequenced \( \rightarrow 15 \) million bases not covered

Coverage for an even mixture of two bacteria
Coverage for known abundances of two bacteria

Coverage for unknown abundance of unknown number of bacteria

Coverage for a transcriptome

Coverage for populations that contain variation
Realistic coverage measures
Neither of the models assumptions are correct
Empirical observation \(\rightarrow\) raise the required coverage at least 10 fold
What part of the genome is coverable to begin with?
What part of the genome is uniquely coverable with a give read size?
Nomenclature: “accessible”, “mappable”, “effective” genome sizes

Sequence duplication
Type: Natural Artificial (PCR, Optical)
Source: Sample Contaminant
Detection: Sequence identity Alignment identity
How to deal with duplicates \(\rightarrow\) no easy or universal answer
SNP calling \(\rightarrow\) typically remove if coverage is not too high
All other methodologies \(\rightarrow\) need careful evaluation, natural duplicates my be important

Sequence duplication
Type: Natural Artificial (PCR, Optical)
Source: Sample Contaminant
Detection: Sequence identity Alignment identity
How to deal with duplicates \(\rightarrow\) no easy or universal answer
SNP calling \(\rightarrow\) typically remove if coverage is not too high
All other methodologies \(\rightarrow\) need careful evaluation, natural duplicates my be important

samtools rmdup
alignment based read de-duplication

Dealing with paired data
• Make sure to understand which parts of the DNA fragments have been sequenced. Consult your sequencing operator for details on the library preparation.
• When in doubt you can always map them in single end mode, then visualize the results (covered in later lectures) and see how the pairs are located relative to one another. (sanity check)
• Consult vendor materials \(\rightarrow\) comprehensive but will also contain a lot of details that are not relevant – not easy to make sense of these
Sequencing data simulators
Generating sequencing data with known properties – then try to detect the known features – some generators include error models for sequencing platforms

- *wgsim* (whole genome simulator) by Heng Li
- *dwgsim* (an expanded version of wgsim) by Niels Homer
- *simNGS* from EBI incorporates statistical models to generate errors
- *ART*: a next-generation sequencing read simulator

Some tools have evaluation scripts that generate reports on mapping quality.
Find simulators for the type of data you are studying.

samtools faidx
Another way to index and access fasta files

![Sample command for samtools faidx](image)

Make sure to index files for samtools as well

Pileups, variant calling
• *Pileup* → show all bases at a given index

Homework 19
Simulate a dataset that covers your genome with the theoretical coverage of 4.

- What parameters did you choose?
- What percent of your genome do you expect to remain unsequenced?

Align the data against the reference.

- What percent of your genome has zero coverage?
- What is the average coverage?
- What is the maximal coverage of your simulated data?