H3ABioNet
Pan African Bioinformatics Network for H3Africa

WELLCOME GENOME CAMPUS
CONNECTING SCIENCE
ADVANCED COURSES+ SCIENTIFIC CONFERENCES

Next Generation Sequencing Bioinformatics Course 2021

Module 2: Introduction to NGS Technologies

Experimental Design

Fatma Guerfali

Learning Objectives

- Describe the essential steps to conduct a NGS experiment (from Biological question to Biological interpretation)
- Recognize the importance of experimental design and its influence on each of these step
- Summarize the essential elements of eperimental design

Session Plan

01

The NGS Experiment

From Biological question to Biological interpretation

02

The Experimental Design (DNA/RNA)

Essential elements of an Experimental Design

Session Plan

01

The NGS Experiment

From Biological question to Biological interpretation

02

The Experimental Design (DNA/RNA)

Essential elements of an Experimental Design

The NGS experiment: Overview of key steps

Overview

01
The NGS Experiment

From Biological question to Biological interpretation

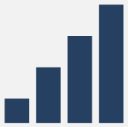
02
The Experimental Design (DNA/RNA)

Essential elements of an Experimental Design

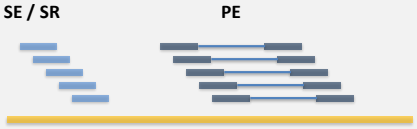
Key Considerations for Experimental Design

SAMPLE TYPE

SEQUENCING DEPTH



SEQUENCING MODE

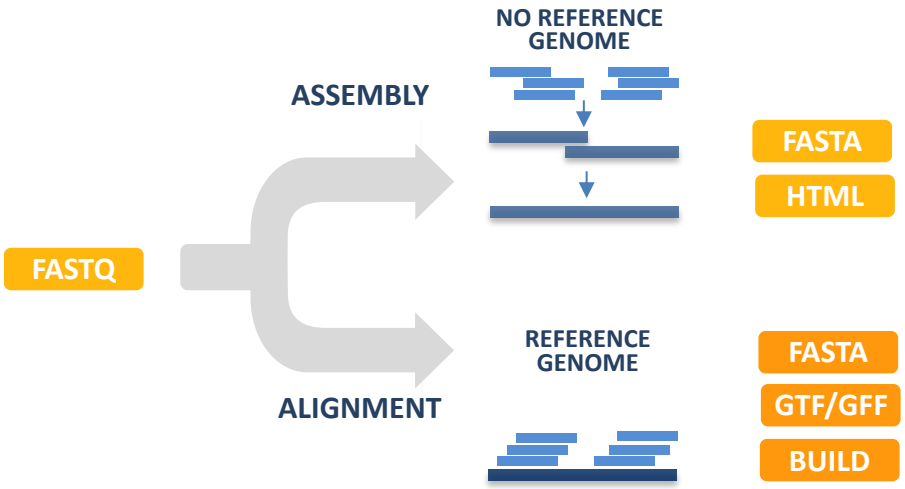


SEQUENCING STRATEGY

SAMPLE SIZE

Sample Type

- *Low input? consider quality & quantity checks (specific kits)*
- *Reference Genome ? consider the status of finishing (build, version)*



Sequencing Depth

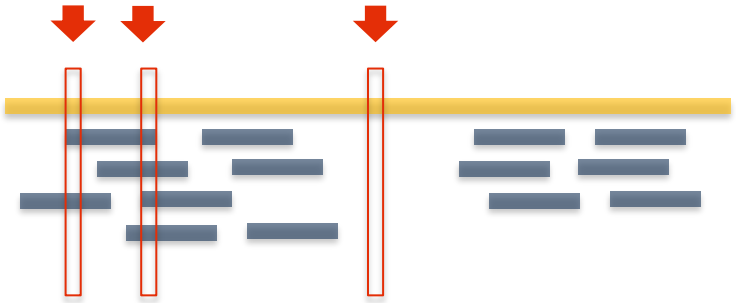
Coverage

High-pass sequencing design

Low-pass sequencing design

► Coverage

The (average) number of times each nucleotide is « read »
→ Fold Coverage (number + X)



► Detection of low frequency mutations within a mixed cell population

Somatic mutations may only exist within a small proportion of cells in a given tissue sample
→ region of DNA having the mutation must be sequenced at extremely **high coverage**, >1000×

► Genome-wide variant discovery

Study design involves sequencing many samples (hundreds to thousands) at **low coverage**
→ allows to achieve greater statistical power within a given population.

https://informatics.fas.harvard.edu/whole-genome-resequencing-for-population-genomics-fastq-to-vcf.html#design/

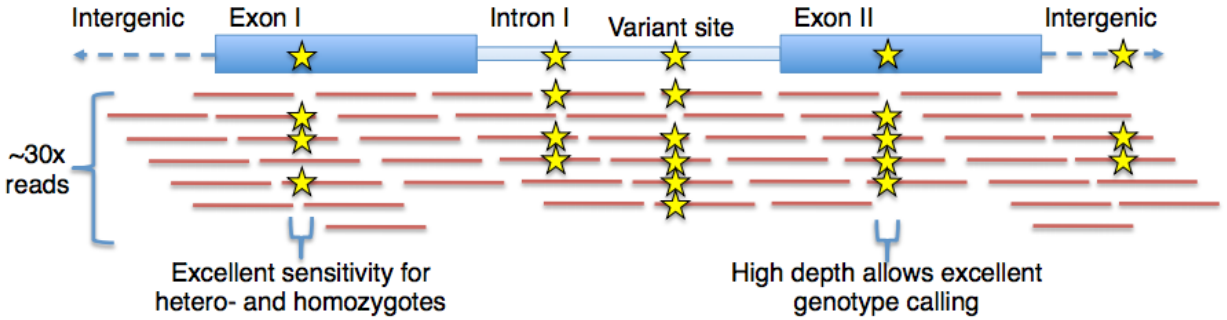
Sequencing Depth

Coverage

High-pass sequencing design

Low-pass sequencing design

Low-coverage WGS vs High-Coverage WGS
→ important to confidently call variants



Data requirements per sample

Targeted bases	~3 Gb
Coverage	Avg. 30x
# sequenced bases	100 Gb
# lanes of HiSeq	~8 lanes

Variant detection among multiple samples

Variants found per sample	~3-5M
Percent of variation in genome	>99%
Pr{singleton discovery}	>99%
Pr{common allele discovery}	>99%

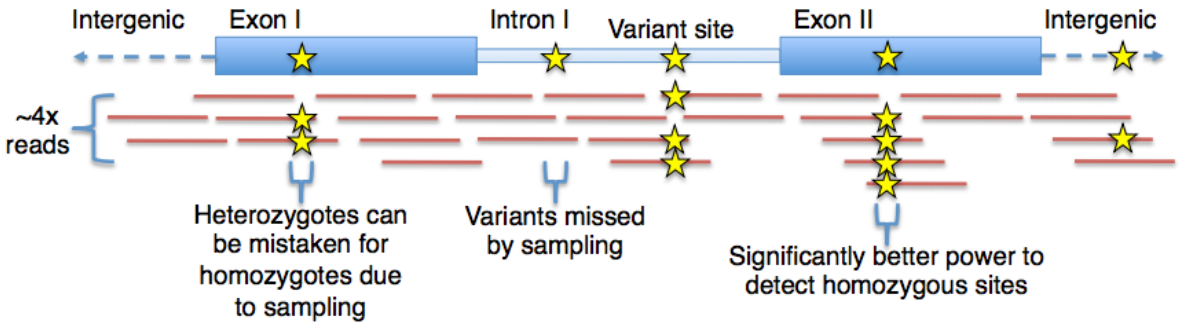
Chris Fields, 2019
https://slideplayer.com/slide/17061224/

Sequencing Depth

Coverage

High-pass sequencing design

Low-pass sequencing design



Data requirements per sample	
Targeted bases	~3 Gb
Coverage	Avg. 4x
# sequenced bases	20 Gb
# lanes of HiSeq	~1.25

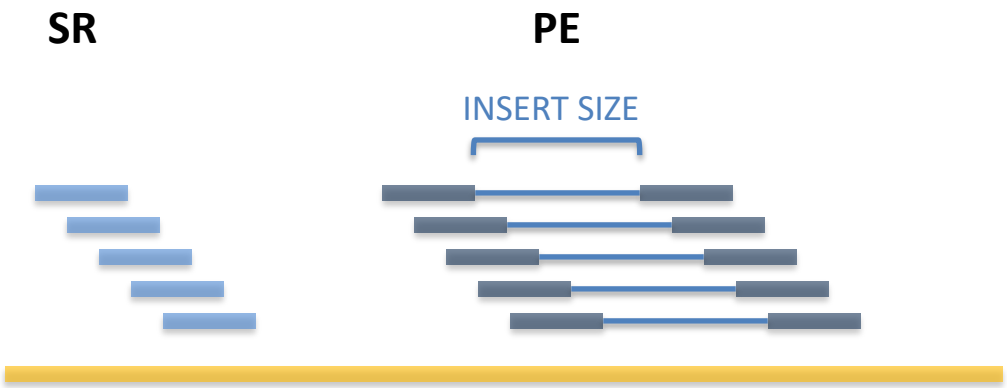
Variant detection among multiple samples	
Variants found per sample	~3M
Percent of variation in genome	~90%
Pr{singleton discovery}	<50%
Pr{common allele discovery}	~99%

Chris Fields, 2019
https://slideplayer.com/slide/17061224/

Sequencing Mode

Single-End (SE/SR) vs Paired-end (PE)

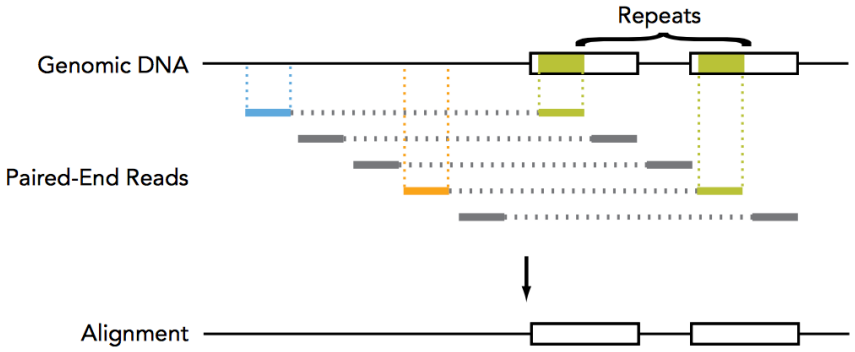
- *SE (Single-End Reads)*
- *PE (Paired-end Reads):* PE involves sequencing both ends of the DNA fragments and aligning the forward and reverse reads as read pairs



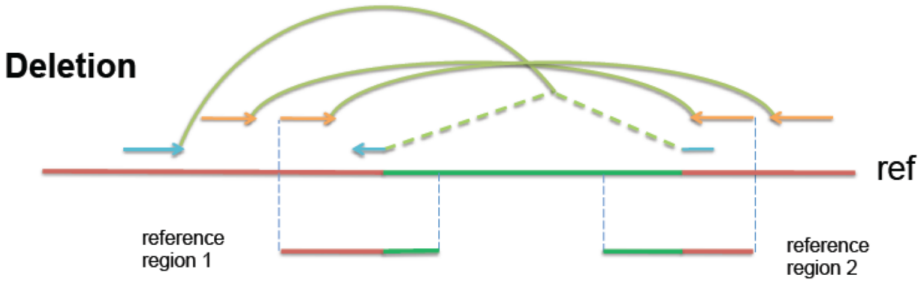
Sequencing Mode

PE Advantages

- *More accurate reads alignment*
- *Less ambiguous mapping of repeats*
- *Detection of even small deletions*
- *Estimation of InDels sizes*
- *Allows removal of PCR duplicates (common artifact resulting from PCR amplification during library preparation: via Analysis of differential read-pair spacing)*



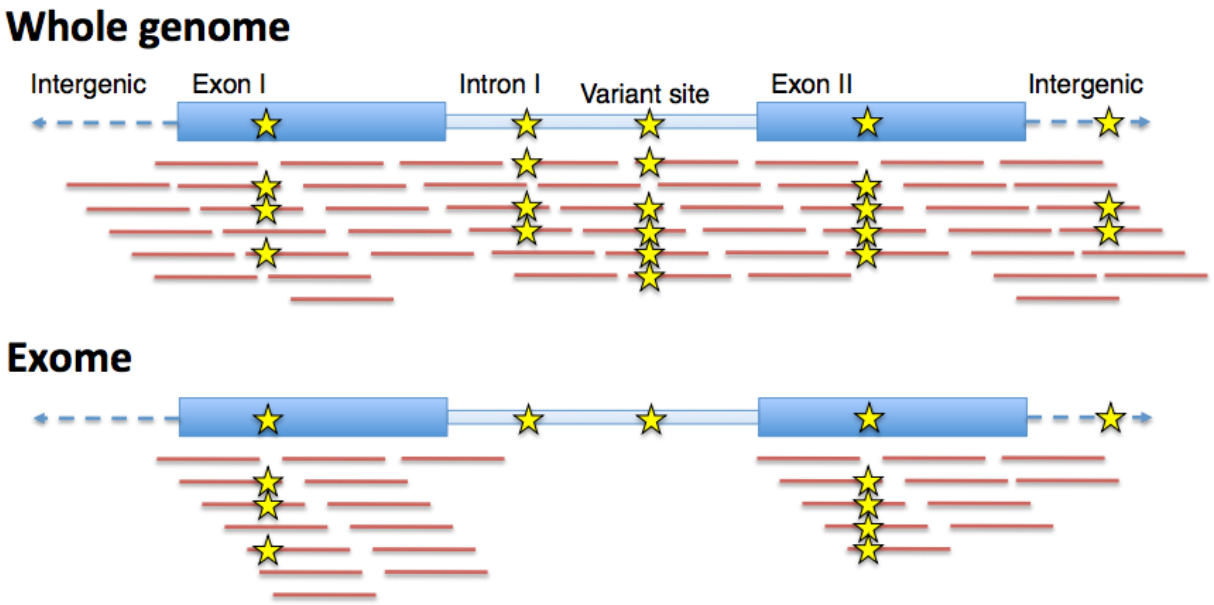
Reads in repeats (green) can be unambiguously aligned in complex genomes. Each read is associated with a paired read (blue or orange) and the separation between read pairs is known from the fragment size of the input DNA.



http://www.illumina.com/

Sequencing Strategy: WGS vs WES

- ***WGS**: Covers all, but higher cost if deep sequencing required (30X - 50X - 100X)*
- ***WES**: Covers exons only, but higher coverage of transcribed sequences (**targeted**)*
- ***Targeted**: gene panels, etc*



Chris Fields, 2019
https://slideplayer.com/slide/17061224/

Sample Size

Number

Single vs Multiple samples

Replicates

How many individuals to sequence?
Depends on the types of analysis to conduct !

- Describe population structure
  → few individuals
- Detailed demographic inference
  → small (old events, testing models)
  → Large (recent events)
- Identify allele frequency shifts or GWAS
  → Large (power to detect significant differences)

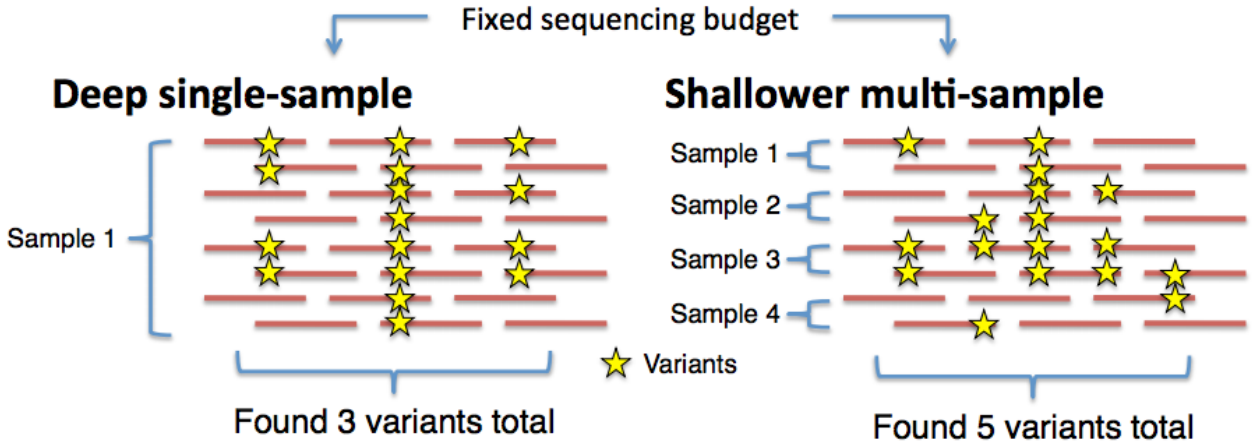
https://informatics.fas.harvard.edu/whole-genome-resequencing-for-population-genomics-fastq-to-vcf.html#design/

Sample Size

Number

Single vs Multiple samples

Replicates



- Higher sensitivity for variants in the sample
- More accurate genotyping per sample
- Cost: no information about other samples

- Sensitivity dependent on frequency of variation
- Worse genotyping
- More total variants discovered

Chris Fields, 2019
https://slideplayer.com/slide/17061224/

Sample Size

Number

Single vs Multiple samples

Replicates

- Multiplexing or not ?

- **multiplexing** = attach samples to a specific **barcode** sequence to identify later the sample from which it originates.
- Libraries pooled and sequenced in parallel
- Reads from each library are differenciated by using barcode to de-multiplex
- Each set is aligned to the reference genome

Sample Size

Number

Single vs Multiple samples

Replicates

► Multiplexing or not ?

Multiplexing (Pooled seq) vs individual barcoded sequencing

- **Multiplexing** : cost saving in library prep & have estimates of allele frequencies, but risk of unequal library representation & poor haplotype information

- **Individual** : variants can be called from individuals with high coverage, but higher cost

https://informatics.fas.harvard.edu/whole-genome-resequencing-for-population-genomics-fastq-to-vcf.html#design/

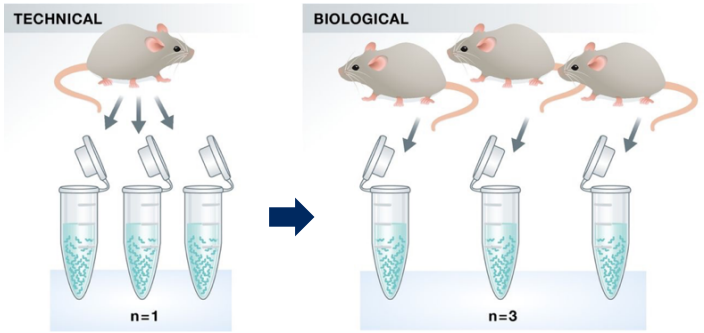
Sample Size

Number

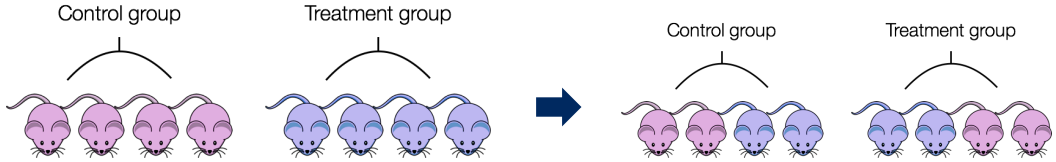
Single vs Multiple samples

Replicates

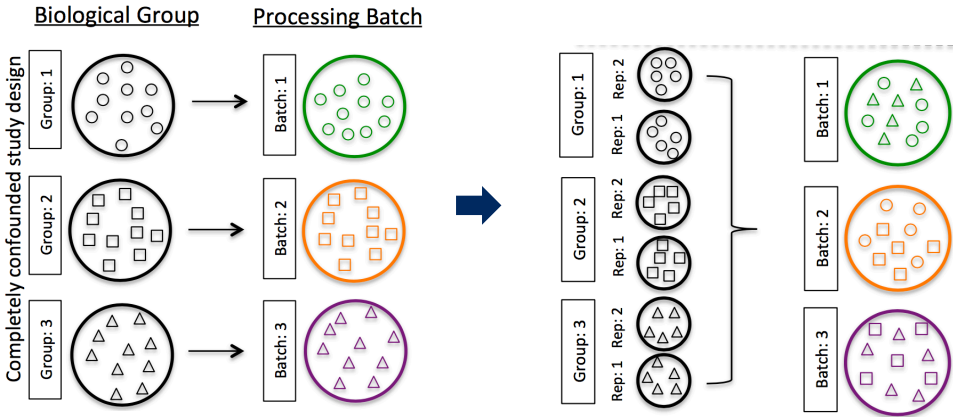
- Number and type of **replicates**



- Avoid **confounding effects**



- Avoid **batch effects**



https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/experimental_planning_considerations.html

Take-home message

- **Understanding each step of an NGS experiment is essential to properly design your NGS experiment (all connected, 1 step can bias the others)**

- **Because each step can be a potential source of bias → greatly affect the quality of the analysis and biological interpretation**

- **A proper Experimental Design takes into account all these special considerations and should be discussed with different actors of the analysis before performing the experiment (Biologists, Bioinformaticians, Biostatisticians)**