Experimental design

Dr Gustavo A. Silva-Arias

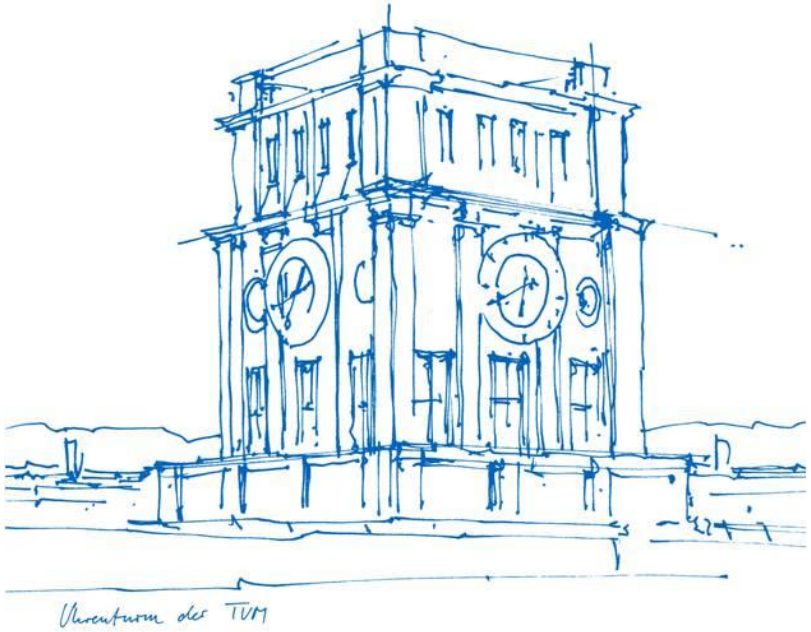
Professorship for Population Genetics

Dept of Life Science Systems

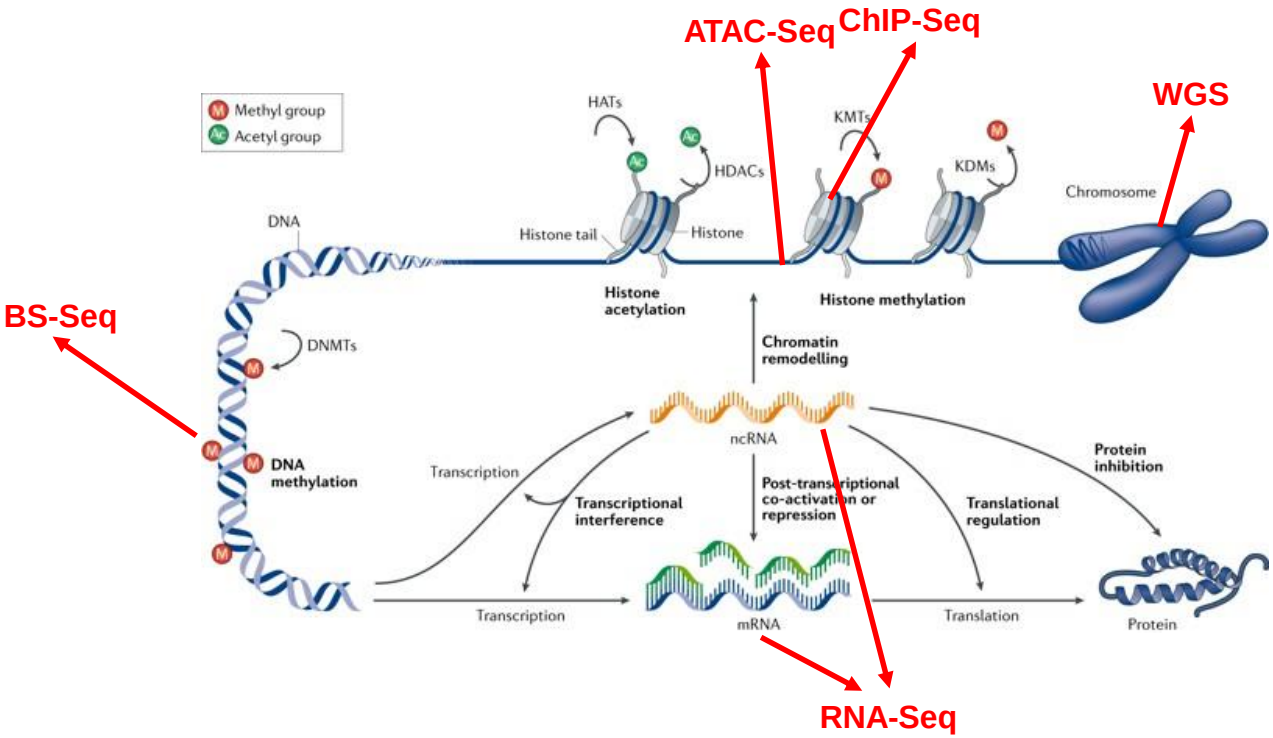
Technical University of Munich

La Paz, Cesar

1 de Agosto de 2022



NGS Applications



Source: https://doi.org/10.1038/s41585-018-0023-z

NGS Project Checklist

1. What is the research question?

2. What genomic resources have been developed for the species?

3. Sequencing decisions:

    1. What sequencing coverage do we need?
    2. How much error rate can we tolerate?
    3. What read length is the most appropriate?
    4. Should we perform Single-End or Paired-End sequencing?
    5. Can we use multiplexing?
    6. Analysis requirements

NGS Project Checklist

1. What is the research question?

2. What genomic resources have been developed for the species?

3. Sequencing decisions:

    1. What sequencing coverage do we need?
    2. How much error rate can we tolerate?
    3. What read length is the most appropriate?
    4. Should we perform Single-End or Paired-End sequencing?
    5. Can we use multiplexing?
    6. Analysis requirements

NGS Project Checklist

1. What is the research question?

2. What genomic resources have been developed for the species?

3. Sequencing decisions:

   - What sequencing coverage do we need?
   - How much error rate can we tolerate?
   - What read length is the most appropriate?
   - Should we perform Single-End or Paired-End sequencing?
   - Can we use multiplexing?
   - Analysis requirements

Example 1: Genome assembly of the „Lima bean“ *Phaseolus lunatus* L.

Aim:
Agronomically and economically significant species within the *Phaseolus* genus. Provides a vital source of nutrients globally.

The genome sequence is needed as a basis for other genetic/genomic studies.

- Study the molecular basis of convergent phenotypic adaptation
- Shows adaptation to a wide range of ecological conditions, especially to heat and drought stresses, traits that are key in scenarios of adaptation to climate change

Example 1: Genome assembly of the „Lima bean“ *Phaseolus lunatus* L.

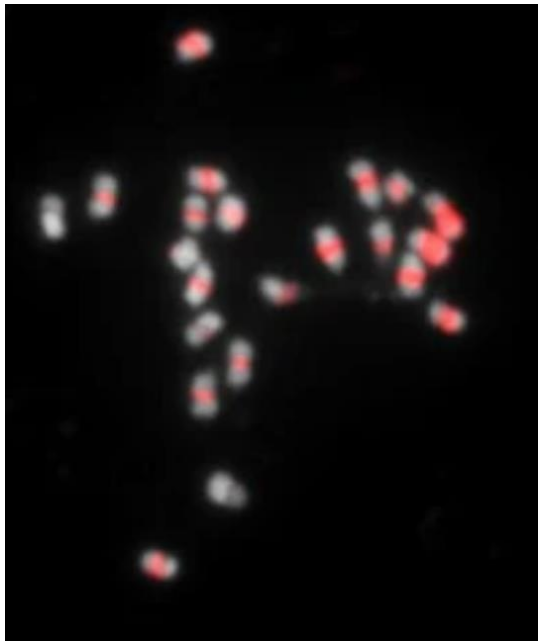
Genomic resources available:

Essentially none.

- Flow cytometry estimates indicate a c-value of ~1.41 pg/nucleus indicating a genome size of ~1410 Mbp.
- Also the karyotype had been determined, it has 22 chromosomes.

Multiplexing:

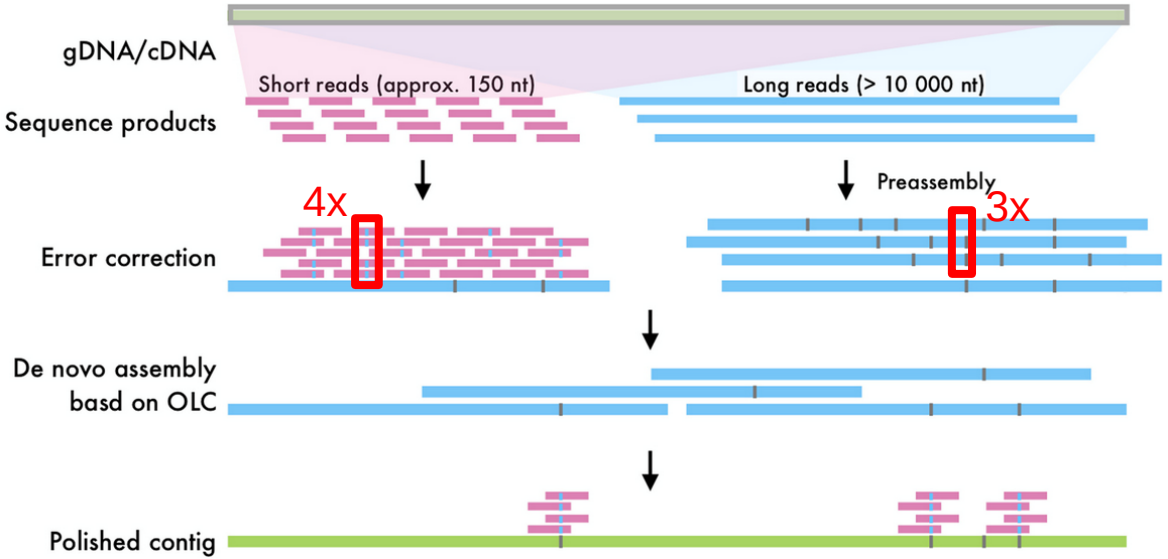
Not needed, we need to generate a genome from a single sample.



Source: https://doi.org/10.1007/s00122-012-1806-x

Example 1: Genome assembly of the „Lima bean“ *Phaseolus lunatus* L.

Sequencing: Read length and error tolerance



Source: https://doi.org/10.1111/dgd.12608

Example 1: Genome assembly of the „Lima bean“ *Phaseolus lunatus* L.

Sequencing: Coverage and error tolerance

Genome Size = 1.41Gbp or 1,410,000,000 bp

PacBio recommended coverage = 30x
PacBio average read length = 10Kbp or 10,000 bp

Illumina recommended coverage = 50x (100x if only Illumina)
Illumina read length = 2 x 150 bp (Paired-End) or 300 bp

We need:
~4.5 million **PacBio** reads and ~435 million **Illumina** reads

Sequencing: Coverage, how many reads needed to achieve it?

Number of reads =
(Genome Size * Coverage) / Read Length

Example 1: Genome assembly of the „Lima bean“ *Phaseolus lunatus* L.

Analysis:

De novo assembly is storage and RAM-demanding, process is currently parallel so it benefits from having many CPUs.

We will need a computer with hundreds of Gigabytes of RAM (512Gb – 1024Gb), most commercial laptops have 8Gb.

A hard drive with several Terabytes of space is needed. A laptop usually has between 0.5-1 Tb.

The node we use in the cluster for practice have 24Gb of RAM and 16 CPUs, not enough for this project!

Example 2: Annotation of the „Lima bean“ genome

Research question:

How many / Which genes, transposable elements and any other genomic features are present in the *Phaseolus lunatus* genome?

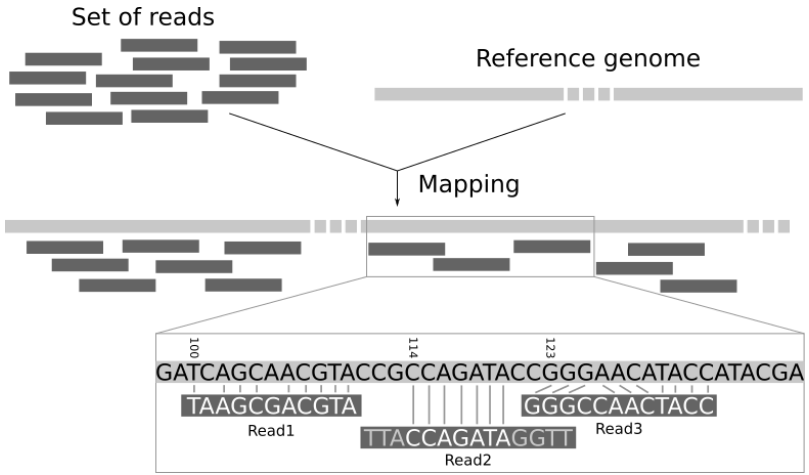
Procedure

- Use TE databases to predict mobile elements in the genome.
- We can use software to predict where are the likely positions of the genes, but perhaps some are pseudogenes or are not expressed.
- We can transfer the annotation of ‘conserved’ genes from close related species (e.g. common bean).
- We need to sequence the DNA that is transcribed (some of it will be expressed as proteins) to find the exact regions spanned by genes in the genome.

Example 2: Annotation of the „Lima bean“ genome

Genomic resources available:

The previous project obtained the genomic sequence that can be used as a reference.



Source: https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Example 2: Annotation of the „Lima bean“ genome

Specific experimental considerations:

We need to extract RNA. Remember that Illumina can only sequence DNA, so we need to **retro-transcribe RNA** into cDNA for sequencing (different library preparation).

Different organs transcribe genes in varying amounts (differential expression), we need to extract **RNA from different organs** and use the mixture to have a good representation of all the genes.

Example 2: Annotation of the „Lima bean“ genome

Sequencing: Coverage and error tolerance

Software gene prediction indicates ~0.7% of the total genomic sequence corresponds to genes. Let’s be generous and assume 1%

Transcriptome size = 1.41 Gbp * 0.01 = 14,100,000 bp
Coverage = 30x
Illumina read length = 2x100 bp = 200 bp

We need:
~2-3 million reads, we need accuracy, Illumina in this case good enough

Sequencing: Coverage, how many reads needed to achieve it?

Number of reads = (Genome Size * Coverage) / Read Length

But check
RNA sequencing with the Iso-Seq method

PacBio

Example 2: Annotation of the „Lima bean“ genome

Analysis:

Aligning short reads to a reference genome is not very RAM or CPU demanding. Storage can be a concern.

You need enough RAM to at least hold the reference genome in memory, so at least 22Gb of RAM. 73 million Illumina reads occupy ~12 Gigabytes of disk, so you need that for the input data and twice as much for the output data.

The cluster node could be sufficient for this task, a laptop probably not.

But check
RNA sequencing with the Iso-Seq method

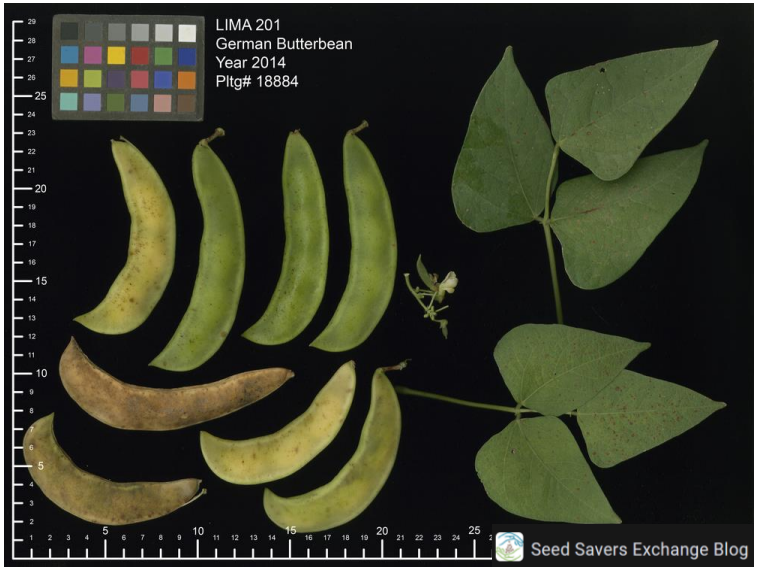
PacBio

Example 3: What genes are involved in pod development in the „Lima bean“?

Research question:

Organ differentiation depends, among many factors, on the levels of specific proteins being expressed in specific tissues and times.

- What are the genes involved in pod differentiation of the Lima bean?
- What time and conditions determine de differentiation of pods in the lima bean?



Example 3: What genes are involved in pod development in the „Lima bean“?

Specific experimental considerations:

We are interested in **expression level** between developmental stages, we need to sequence tissue of each stage separately.

Repeatability becomes an issue, we can not draw meaningful conclusions from a single pod per stage. **Several replicates of each stage have to be sequenced.**

Conditions of tissue collection are important. Expression changes with time of day, temperature, etc.

Example 3: What genes are involved in pod development in the „Lima bean“?

Sequencing: Coverage, read length, error tolerance

In this case we want to estimate gene expression just by counting how many transcripts of each gene are produced in each sample / stage.

Single-End Illumina would be enough to address the question.

Coverage is NOT SO critical and we need to sequence **multiple samples**. 30x is reasonable, this amounts to ~2-3 million reads per replicate.

Multiplexing: Definitely!

Example 3: What genes are involved in pod development in the „Lima bean“?

Analysis:

We assign reads to transcripts. This can be done by aligning (a.k.a. mapping).

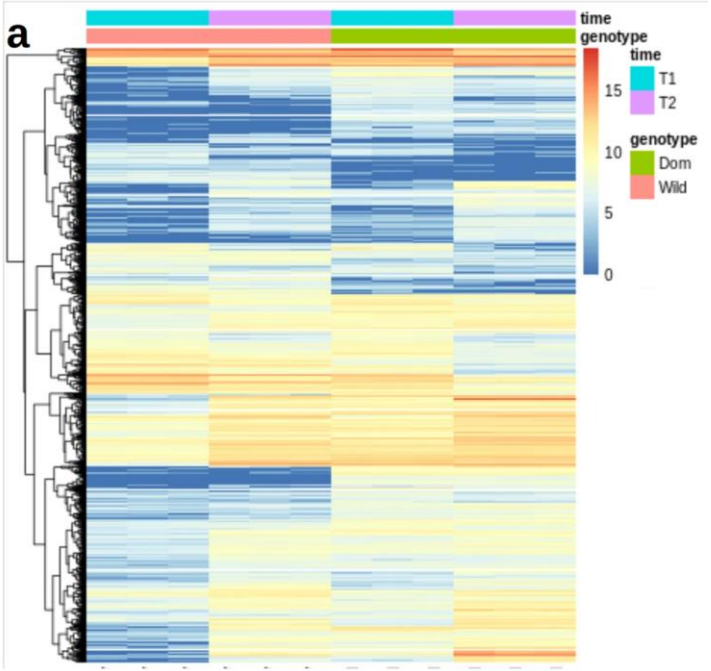
The reference transcriptome is ~1.41Gb. Minimal RAM requirements, just need storage to hold all input reads from all the replicates.

You can probably run this in your laptop.

Analyze and visualize results in R.

Example 3: What genes are involved in pod development in the „Lima bean“?

Normalized expression values within genes with differential expression



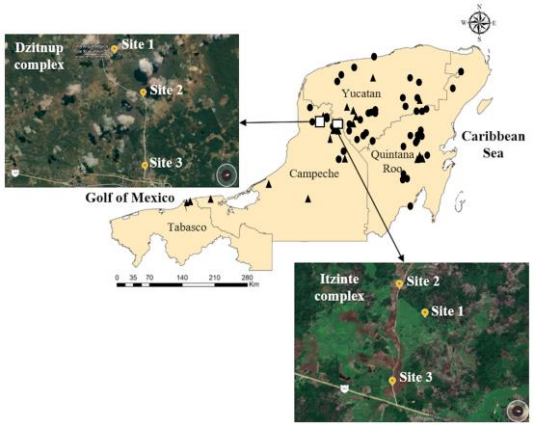
https://doi.org/10.1038/s41467-021-20921-1

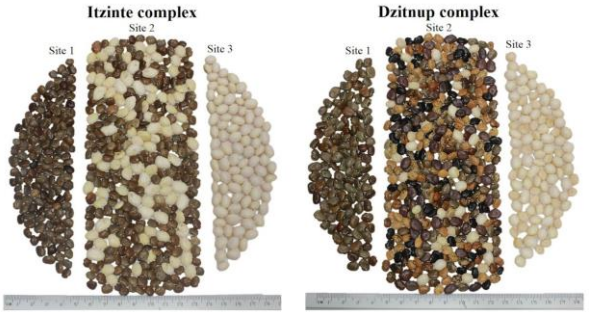
Example 4: Genetic differences between gene pools of „Lima bean“

Research question:

Wild and domesticated plants of Lima bean co-exist in the Yucatan peninsula.

There is gene-flow between wild and cultivates populations? What is the impact of gene-flow in agronomic traits?





https://doi.org/10.7717/peerj.13690

Example 4: Genetic differences between gene pools of „Lima bean“

Genomic resources available:

- ✓ The genome reference sequence
- ✓ Reference annotation

Example 4: Genetic differences between gene pools of „Lima bean“

Specific experimental considerations:

Assess natural variation is more important, we need to sample multiple individuals per population. Resequencing the entire genome of all individuals is impractical.

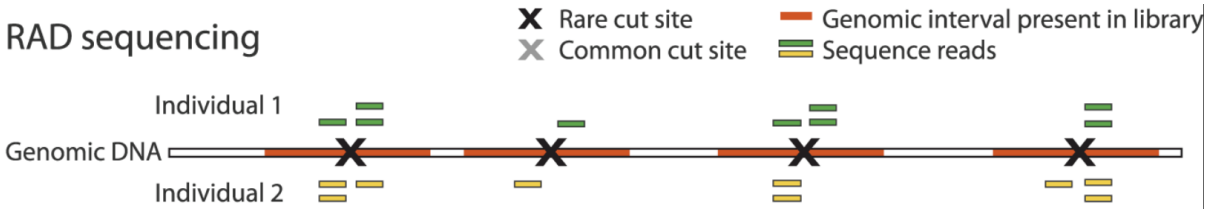
Most of the protein coding genes are conserved within a single species (not enough genetic information to distinguish groups), so transcriptome sequencing is not optimal in this case.

- We need a way to interrogate ***non-coding*** (more variable) regions of the genome in an efficient way.
- And target ***genetic variation related to domestication process***.

Example 4: Genetic differences between gene pools of „Lima bean“

Reduced Representation Libraries

Sequencing small portions of the genome, but consistently across individuals
RAD-Seq : Restriction Associated DNA sequencing



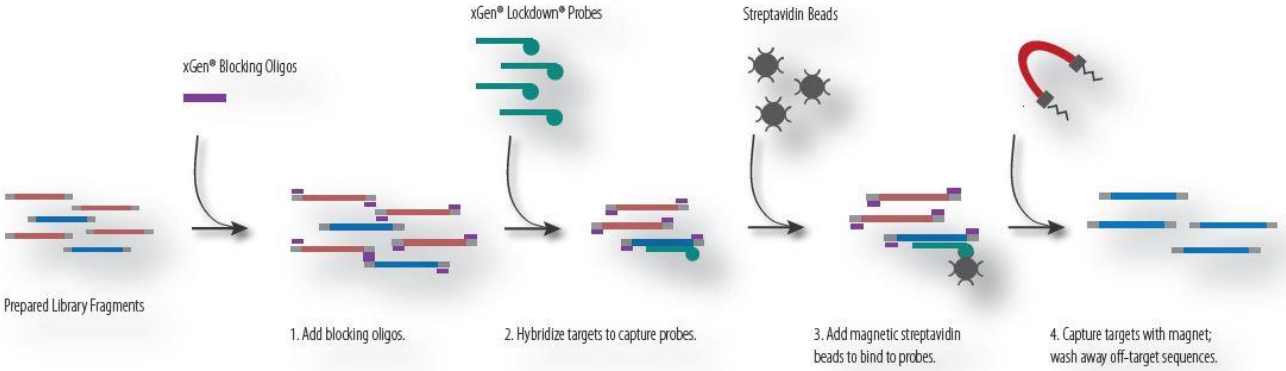
The genome can be used to predict which enzyme will work better.

Source: https://doi.org/10.1371/journal.pone.0037135

Example 4: Genetic differences between gene pools of „Lima bean“

Reduced Representation Libraries

Sequencing target genes
Sequence capture



The annotated genome or transcriptome can be used to generate the probes.

Source: https://eu.idtdna.com/pages/education/decoded/article/target-enrichment-facilitates-focused-next-generation-sequencing

Example 4: Genetic differences between gene pools of „Lima bean“

Sequencing: Coverage, error rate, read length

Analyzing the genome sequence we found *Ape* KI give us 20,000 fragments of ~300 bp. Then, we need to sequence: 20,000 * 2 * 300 = 12,000,000 bp per sample.
Coverage: accurate genotypes with 20x
Read length: Illumina 2x150: 300 bp

We need ~800,000 reads per sample (good, because we will need to sequence hundreds of samples).

Mutiplexing: Definitely necessary!
We can sequence up to 500 samples per lane

Sequencing: Coverage, how many reads needed to achieve it?

Number of reads =
(Genome Size * Coverage) / Read Length

NovaSeq SP flowcell

300–400 million read-pairs per lane

Example 4: Genetic differences between gene pools of „Lima bean“

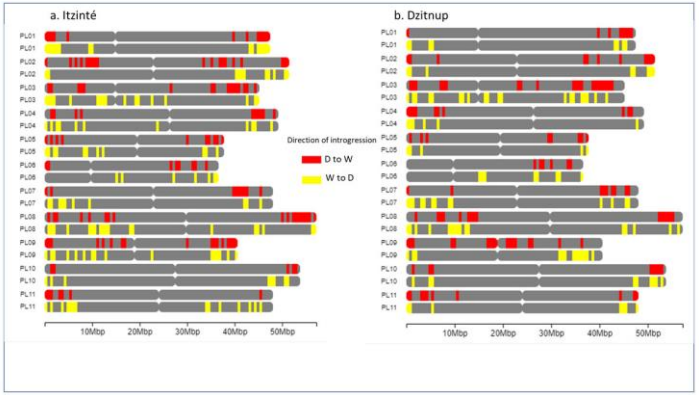
Analysis:

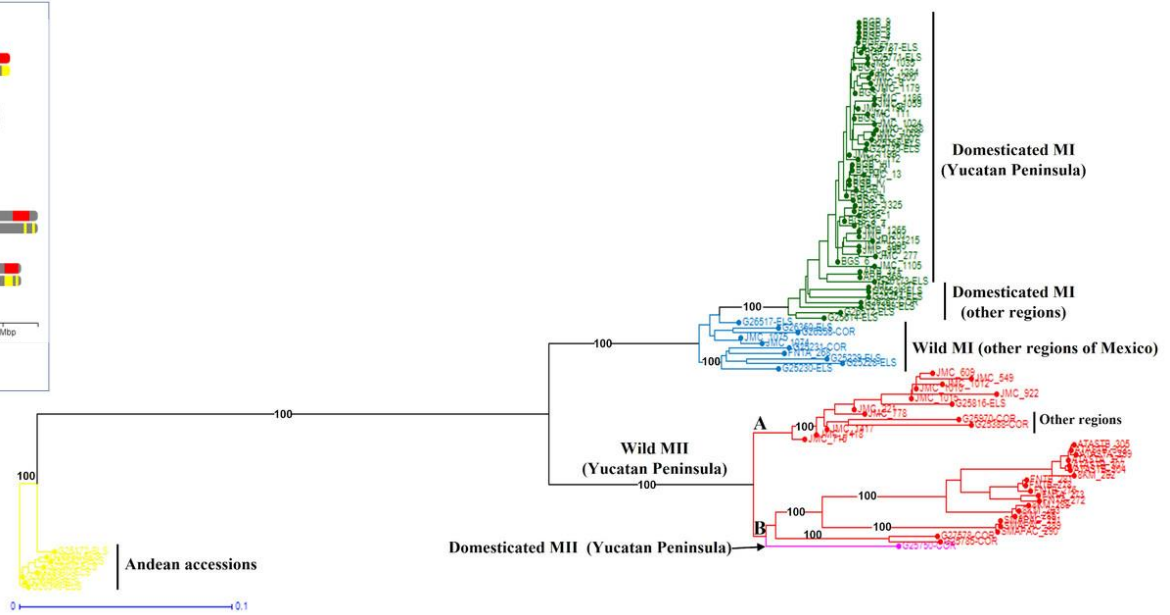
We need to align (map) the reads from all samples to the reference genome to discover SNPs (Single Nucleotide Polymorphisms).

This genome is around 21,4Gbp. We need at least 2 Gigabytes of RAM and enough storage for all the reads coming from all samples.

A node in the cluster can easily accommodate this analysis (16 CPUs and 24 Gigabytes of RAM).

Example 4: Genetic differences between gene pools of „Lima bean“





https://doi.org/10.7717/peerj.13690

Experimental design