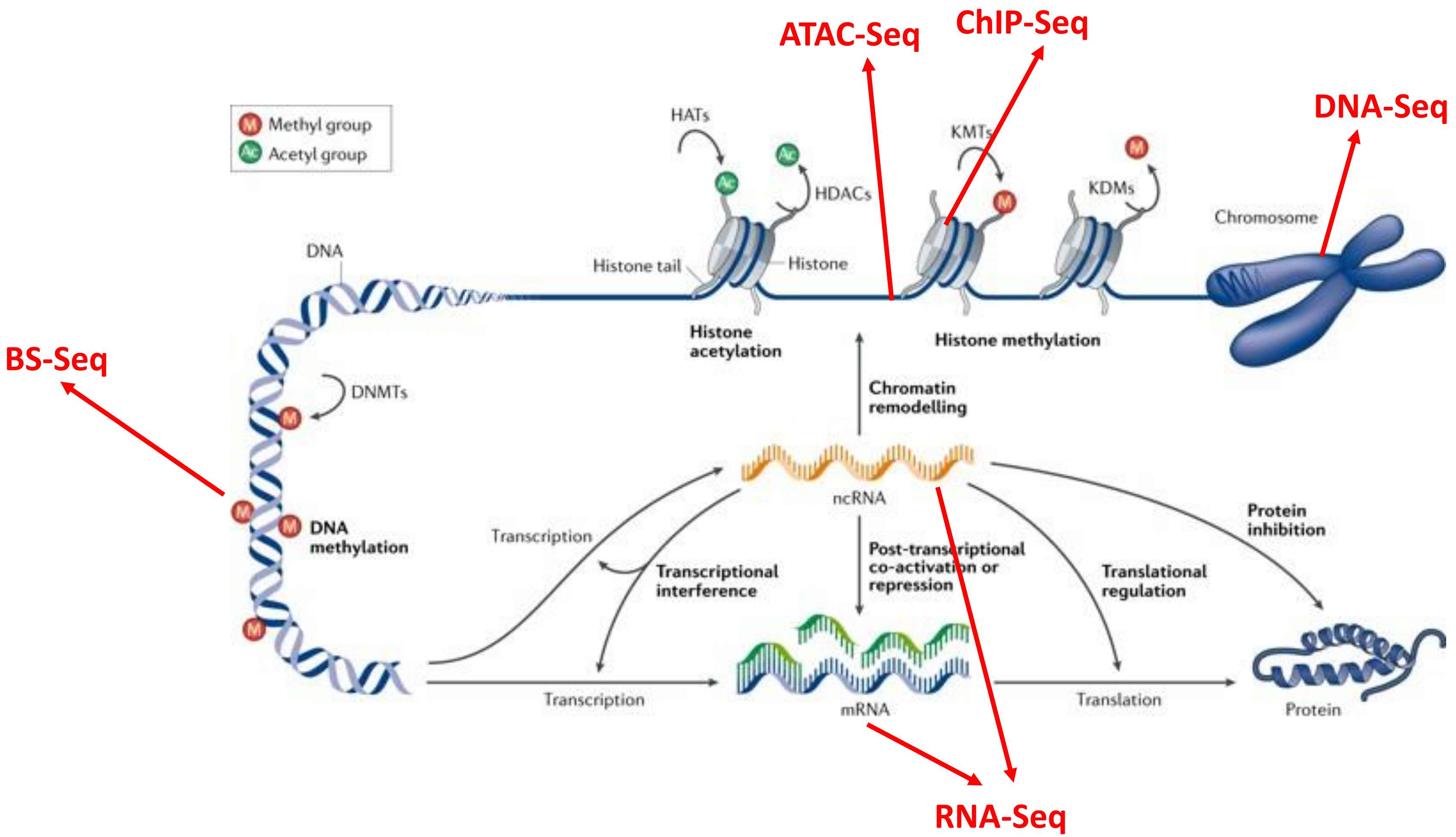
# Experimental design

Dr Gustavo A. Silva-Arias

Technische Universität München

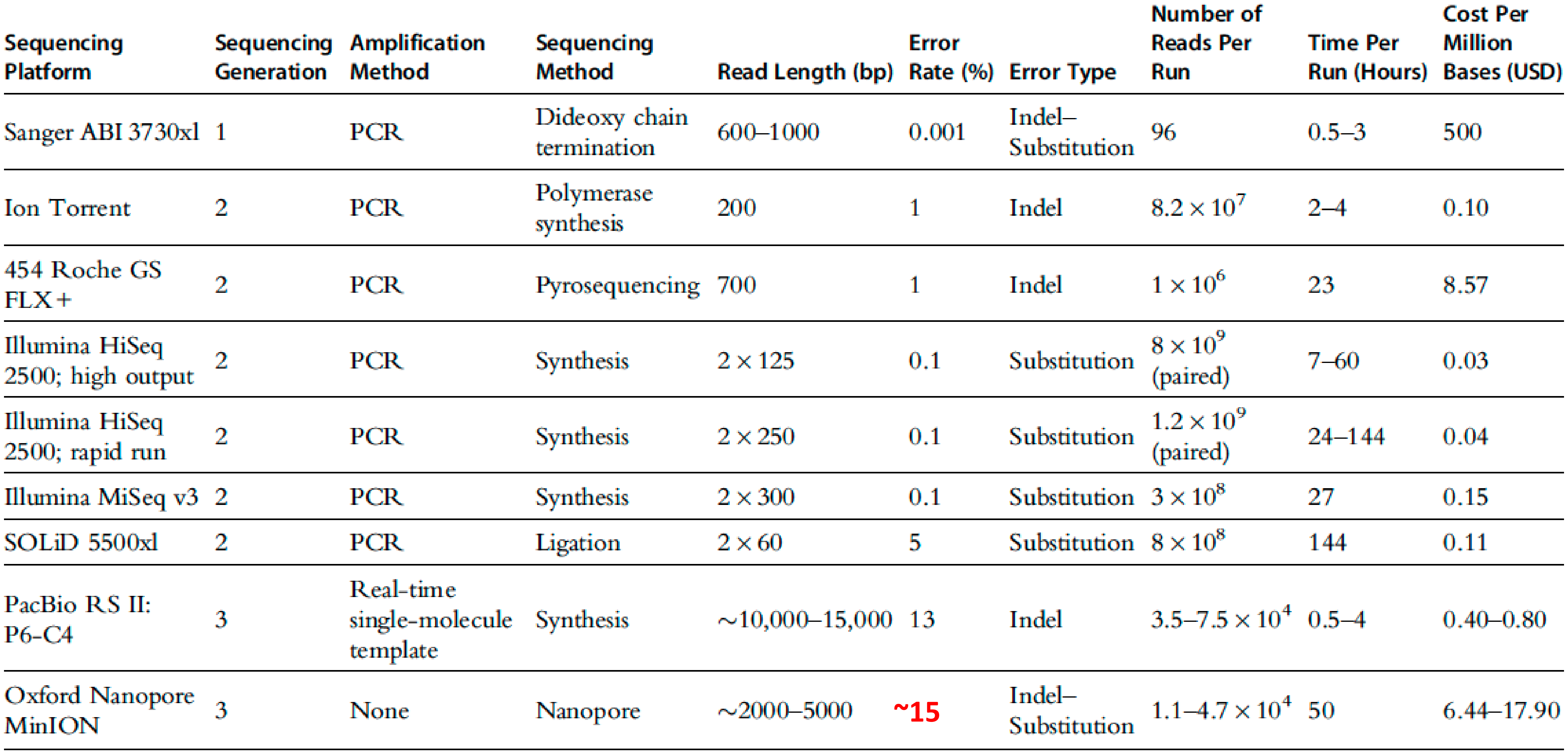
Bogotá, 23 de Agosto 2021

# NGS Applications



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

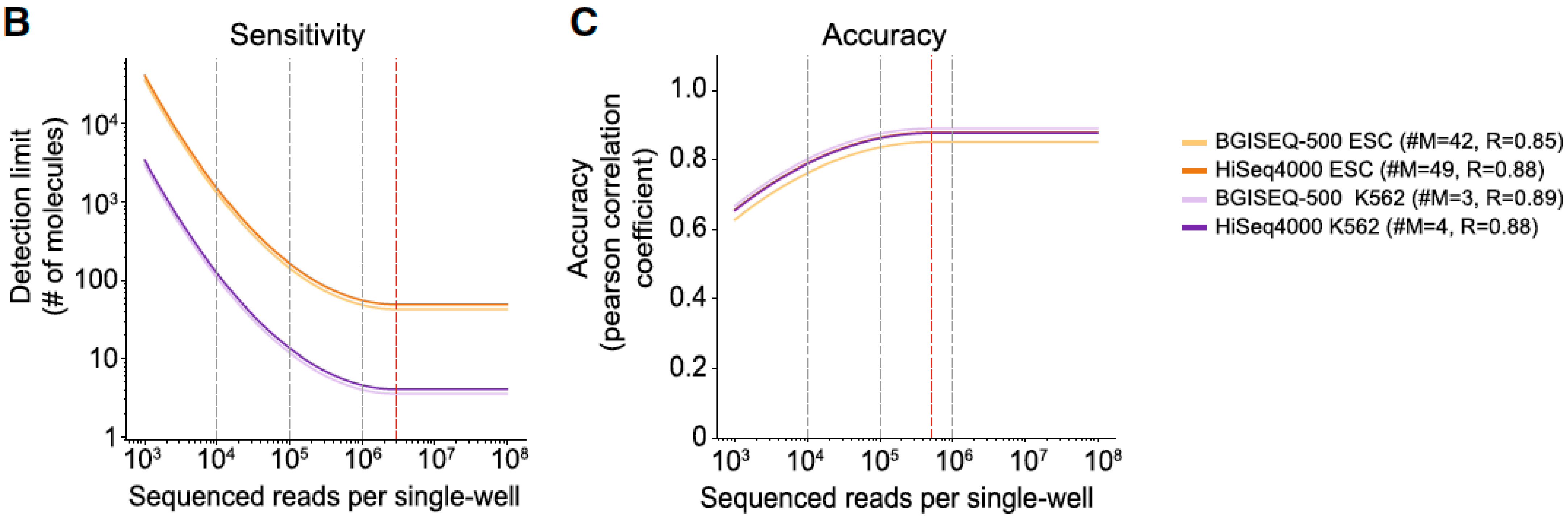
# NGS Platform Comparison



| Sequencing Platform | Sequencing Generation | Amplification Method | Sequencing Method | Read Length (bp) | Error Rate (%) | Error Type | Number of Reads Per Run | Time Per Run (Hours) | Cost Per Million Bases (USD) |
|---|---|---|---|---|---|---|---|---|---|
| Sanger ABI 3730xl | 1 | PCR | Dideoxy chain termination | 600–1000 | 0.001 | Indel–Substitution | 96 | 0.5–3 | 500 |
| Ion Torrent | 2 | PCR | Polymerase synthesis | 200 | 1 | Indel | $8.2 \times 10^7$ | 2–4 | 0.10 |
| 454 Roche GS FLX+ | 2 | PCR | Pyrosequencing | 700 | 1 | Indel | $1 \times 10^6$ | 23 | 8.57 |
| Illumina HiSeq 2500; high output | 2 | PCR | Synthesis | 2 × 125 | 0.1 | Substitution | $8 \times 10^9$ (paired) | 7–60 | 0.03 |
| Illumina HiSeq 2500; rapid run | 2 | PCR | Synthesis | 2 × 250 | 0.1 | Substitution | $1.2 \times 10^9$ (paired) | 24–144 | 0.04 |
| Illumina MiSeq v3 | 2 | PCR | Synthesis | 2 × 300 | 0.1 | Substitution | $3 \times 10^8$ | 27 | 0.15 |
| SOLiD 5500xl | 2 | PCR | Ligation | 2 × 60 | 5 | Substitution | $8 \times 10^8$ | 144 | 0.11 |
| PacBio RS II: P6-C4 | 3 | Real-time single-molecule template | Synthesis | ~10,000–15,000 | 13 | Indel | $3.5–7.5 \times 10^4$ | 0.5–4 | 0.40–0.80 |
| Oxford Nanopore MinION | 3 | None | Nanopore | ~2000–5000 | ~15 | Indel–Substitution | $1.1–4.7 \times 10^4$ | 50 | 6.44–17.90 |

Source: https://doi.org/10.1016/bs.aecr.2017.12.001

# NGS Platform Comparison



Source: https://doi.org/10.1186/s13059-019-1676-5

# 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

# 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

# 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 „loblolly pine“ *Pinus taeda*

***Aim:***

- Very little is known about conifer genomes, except that they are large and complex.
- The genome sequence is needed as a basis for other genetic/genomic studies in conifers.

Source: https://www.utahpeoplespost.com/2014/03/scientists-sequence-largest-loblolly-pine-tree-genome-ever/

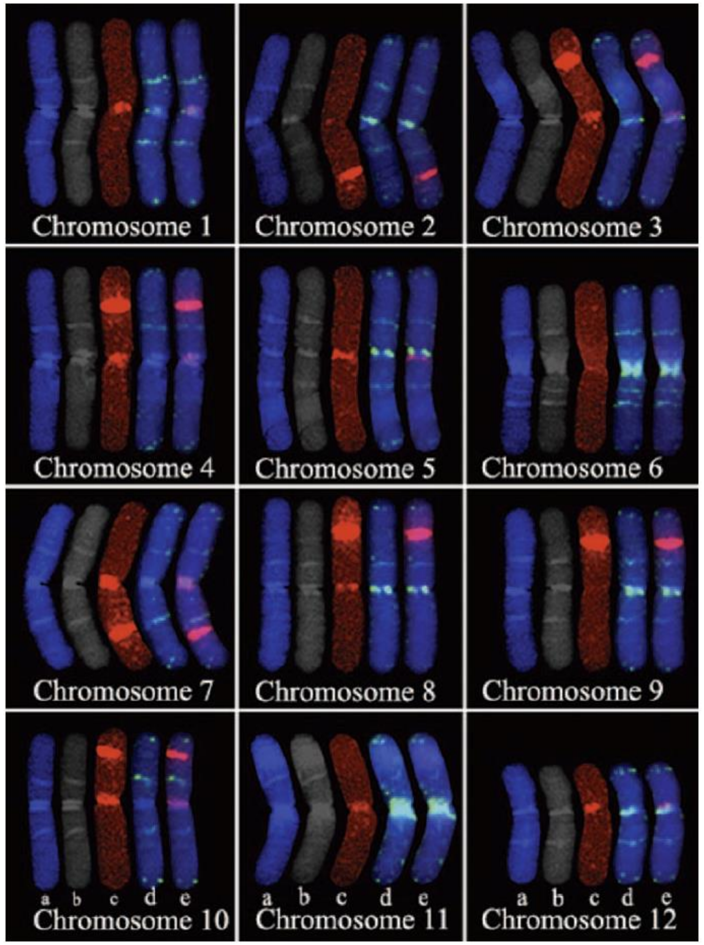
# Example 1: Genome Assembly of the „loblolly pine“ *Pinus taeda*

***Genomic resources available:***

- Essentially none.
- Flow cytometry estimates indicate a genome size of ~22Gbp (6.6x the size of the human genome!).
- Also the karyotype had been determined, it has 12 chromosomes.

***Multiplexing:***

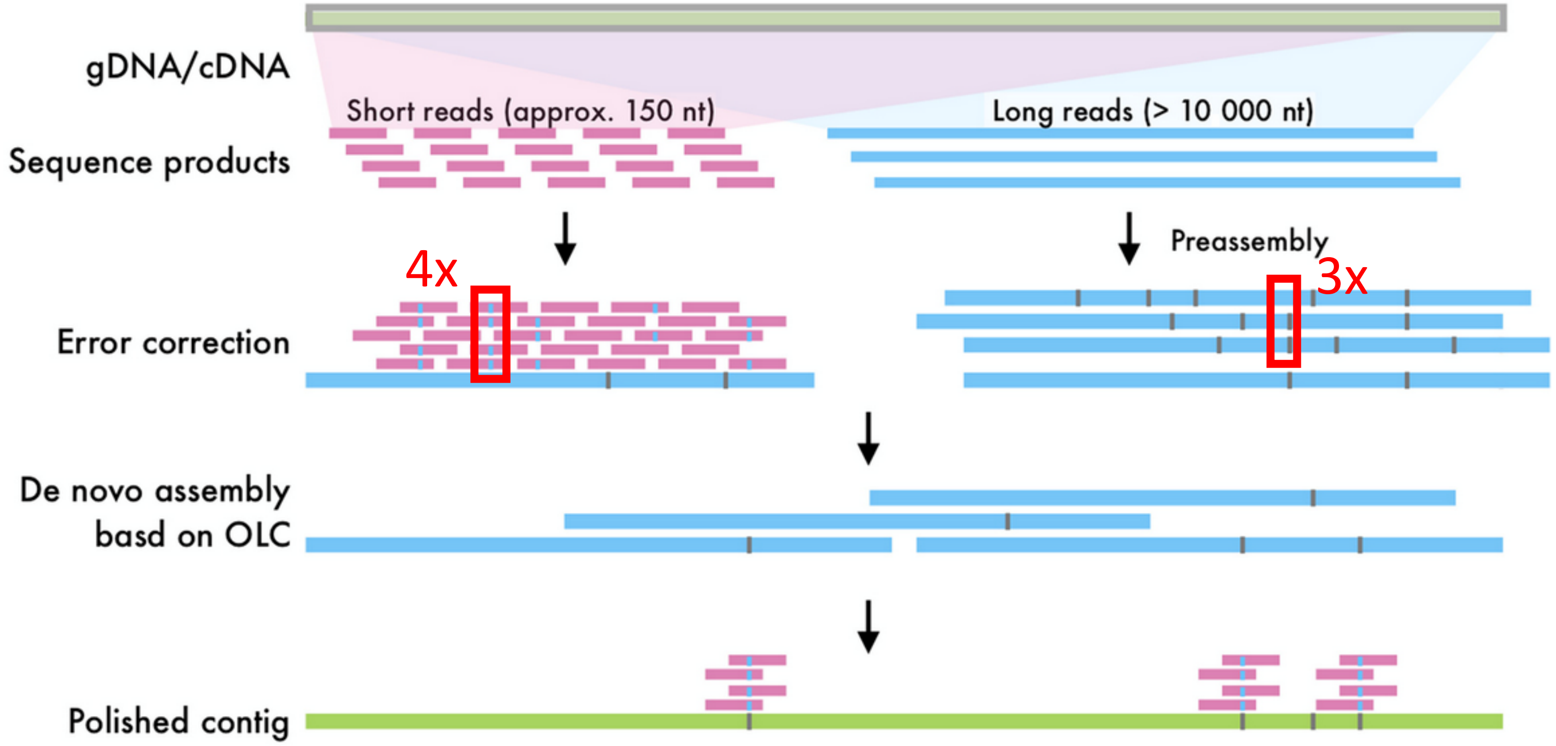
- Not needed, we need to generate a genome from a single sample.



Source: https://doi.org/10.1139/G06-153

# Example 1: Genome Assembly of the „loblolly pine“ *Pinus taeda*

- ***Sequencing: Read length and error tolerance***



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

# Example 1: Genome Assembly of the „loblolly pine“ *Pinus taeda*

- ***Sequencing: Coverage, how many reads needed to achieve it?***
- Number of reads = (Genome Size * Coverage) / Read Length

- Genome Size = 22Gbp or 22,000,000,000 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:*
- 66 million PacBio reads and ~3,600 million Illumina reads!

# Example 1: Genome Assembly of the „loblolly pine“ *Pinus taeda*

- ***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 nodes we use in the cluster for practice have 64Gb of RAM and 24 CPUs, not enough for this project!

# Example 2: Annotation of the „loblolly pine“ genome

***Research question:***

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

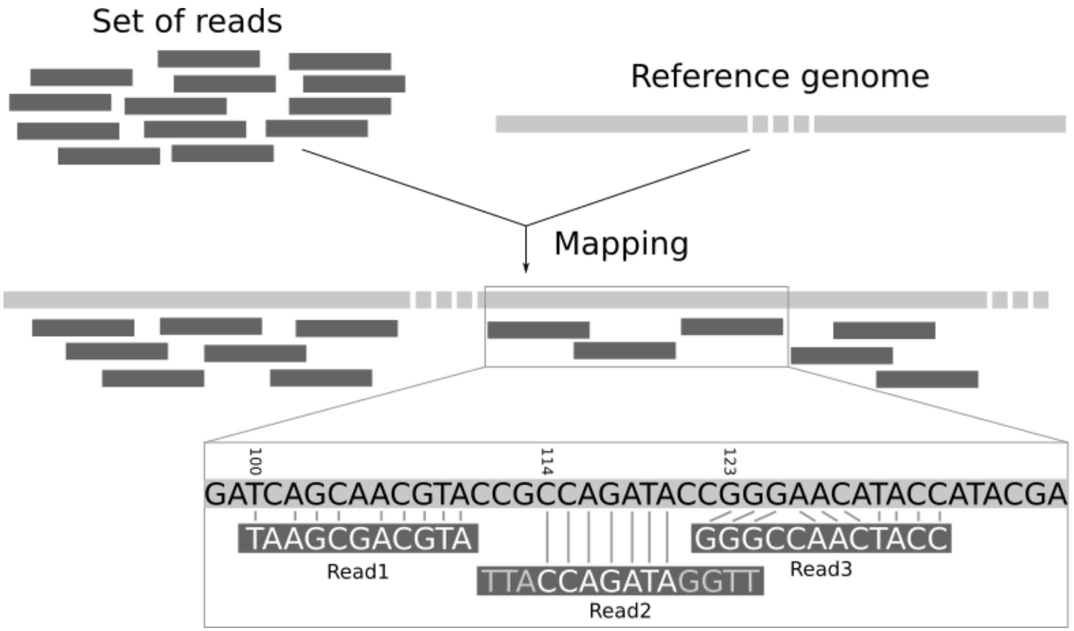
**Procedure**

- Use TE databases to predict mobile elements in the genome.
- Use software to predict where are the likely positions of the genes, but perhaps some are pseudogenes or they are not expressed.
- Transfer the annotation ‘conserved’ genes from close related species.
- Sequence the DNA that gets 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 „loblolly pine“ 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](https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html)

# Example 2: Annotation of the „loblolly pine“ 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 „loblolly pine“ 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 = 22 Gbp * 0.01 = 220,000,000 bp
- Coverage = 100x
- Illumina read length = 2x150 bp = 300 bp

- We need:
- ~73 million reads, we need accuracy, Illumina in this case good enough

# Example 2: Annotation of the „loblolly pine“ 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 nodes are sufficient for this task, your laptop probably not.

# Example 3: What genes are involved in cone sex differentiation in the „loblolly pine“?

***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 male and female cones differentiation of the loblolly pine?
- What time and conditions determine de differentiation of reproductive organs in the loblolly pine?

https://www.britannica.com/science/apical-meristem

Source: https://www.sciencephoto.com/media/17715/view/male-and-female-pine-cones

# Example 3: What genes are involved in cone sex differentiation in the „loblolly pine“?

- ***Genomic resources available:***

✓ The genome reference sequence

✓ Reference annotation

✓ Transcriptome

➢ We can now sequence to estimate expression levels by “counting” how many reads are assigned to each transcript in the transcriptome.

# Example 3: What genes are involved in cone sex differentiation in the „loblolly pine“?

- ***Specific experimental considerations:***

- We are interested in expression level between tissues, we need to sequence each tissue separately.

- Repeatability becomes an issue, we can not draw meaningful conclusions from a single male and a single female cone. **Several replicates of each type 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 cone sex differentiation in the „loblolly pine“?

- ***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 / tissue.

- 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 ~22 million reads per replicate.

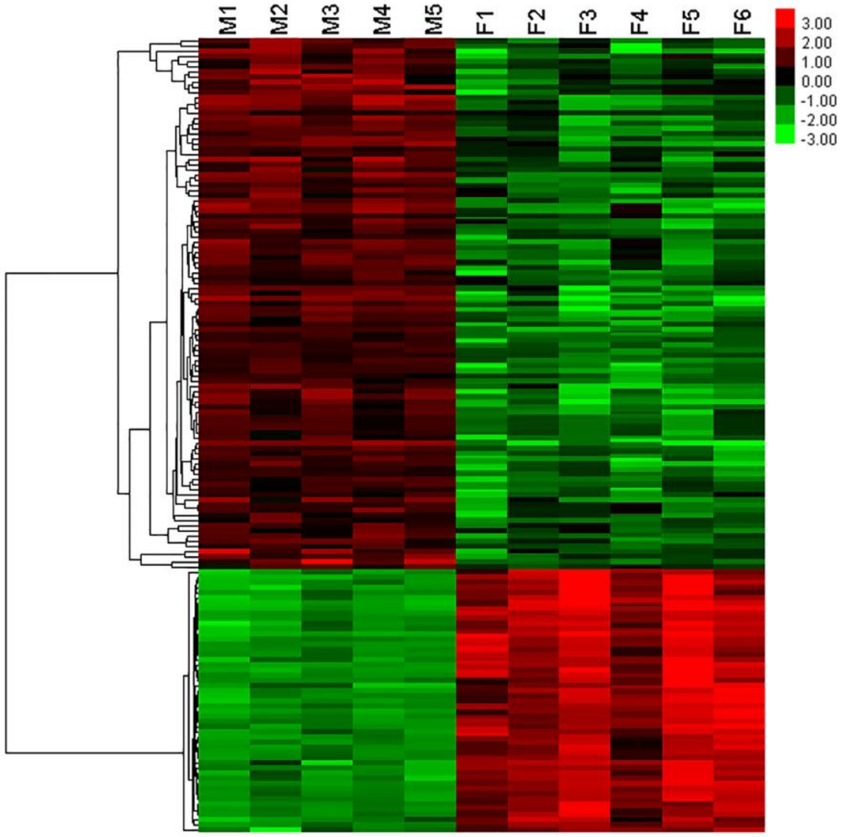
- ***Multiplexing:*** Definitely!

# Example 3: What genes are involved in cone sex differentiation in the „loblolly pine“?

- ***Analysis:***
- We assign reads to transcripts. This can be done by aligning (a.k.a. mapping).
- The reference transcriptome is ~220Mb. 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 cone sex differentiation in the „loblolly pine“?

Genes:
Grouped by
expression
profile



Over-expressed

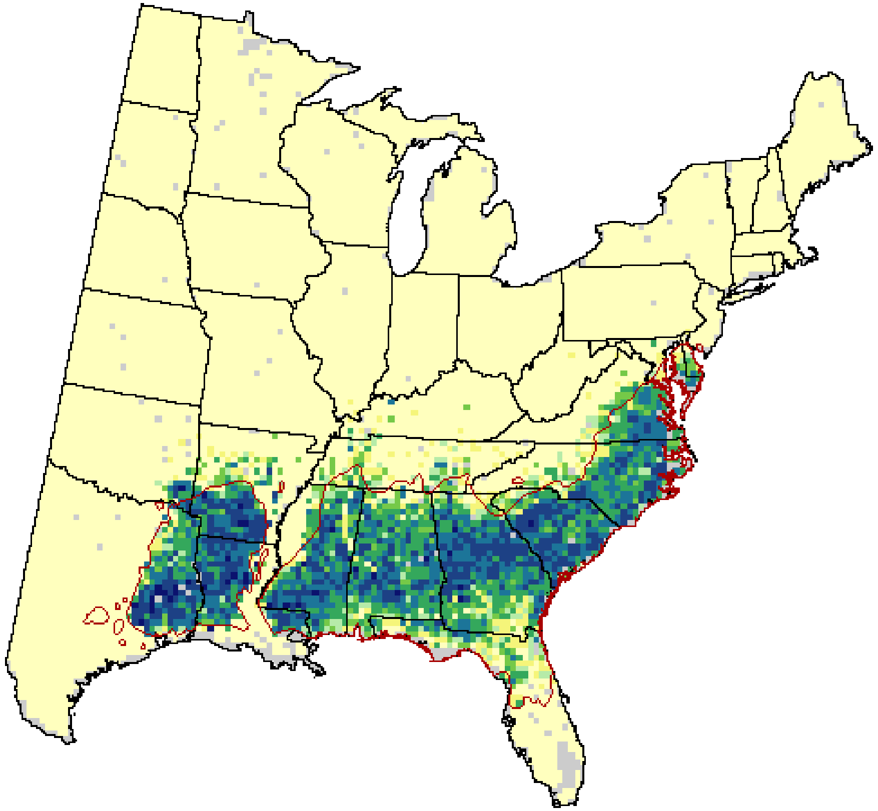
Under-expressed

MALE FEMALE

# Example 4: Genetic differences between populations of „loblolly pine“

- ***Research question:***

- The natural distribution of the species is separated in two main areas by the Mississippi river.
- Are the populations coming from each side of the river genetically distinct? Why?



Source: https://www.fs.fed.us/nrs/atlas/tree/131

# Example 4: Genetic differences between populations of „loblolly pine“

- ***Genomic resources available:***

✓ The genome reference sequence

✓ Reference annotation

# Example 4: Genetic differences between populations of „loblolly pine“

- ***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 impossible (and impractical).

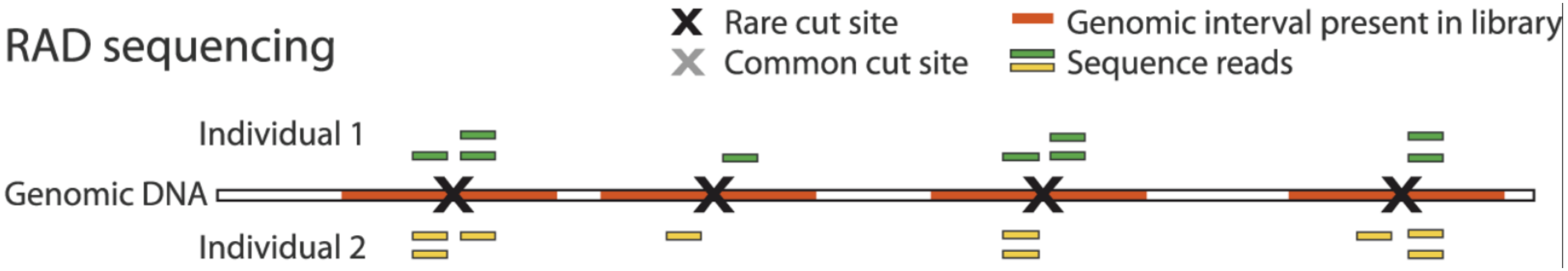
- Most of the protein coding genes are conserved within a single species (not enough genetic information to distinguish groups), so transcriptome sequencing is not optimal in this case.

- We need a way to interrogate ***non-coding*** (more variable) regions of the genome in an efficient way.
- And target potential ‘genes’ related with local adaptation processes.

# Example 4: Genetic differences between populations of „loblolly pine“

- ***Reduced Representation Libraries***

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



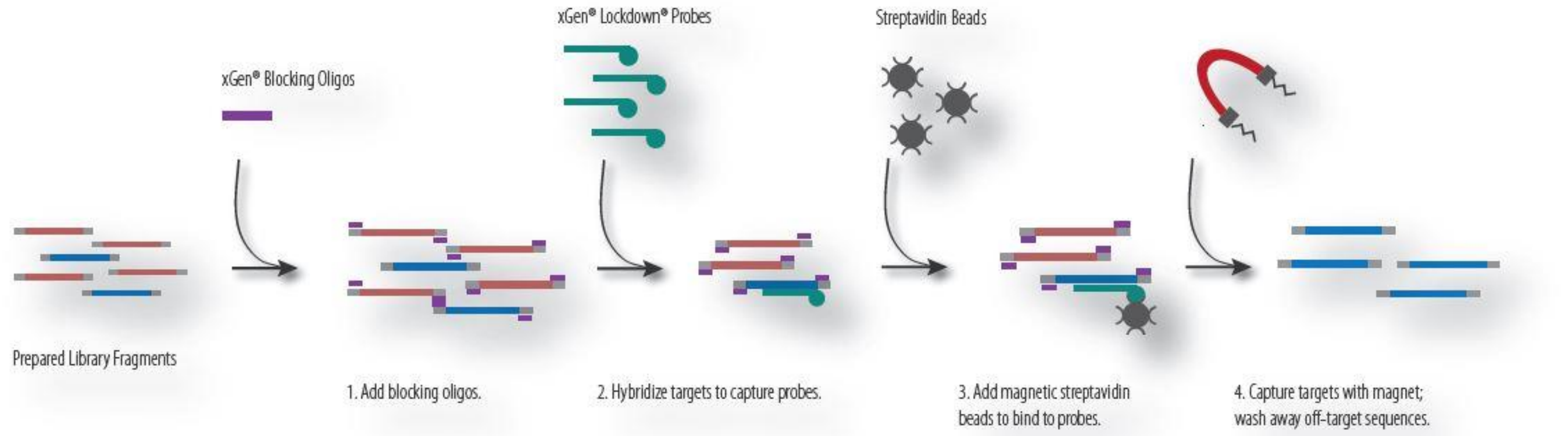
- The genome can be used to predict which enzyme will work better.

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

# Example 4: Genetic differences between populations of „loblolly pine“

- ***Reduced Representation Libraries***

- Sequencing target genes
- **Sequence capture**


xGen® Lockdown® Probes
Streptavidin Beads
xGen® Blocking Oligos
Prepared Library Fragments
1. Add blocking oligos.
2. Hybridize targets to capture probes.
3. Add magnetic streptavidin beads to bind to probes.
4. Capture targets with magnet; wash away off-target sequences.

- 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 populations of „loblolly pine“

- ***Sequencing: Coverage, error rate, read length***

- Analyzing the genome sequence we found EcoRI cuts the genome in 20,000 pieces, we just need to sequence the first and last ~300 bp of each fragment. We need to sequence: 20,000 * 2 * 300 = <u>12,000,000 bp</u>
- Coverage: accurate genotypes with at least 10x, let's sequence <u>20x</u>
- Read length: Illumina 2x150: <u>300 bp</u>

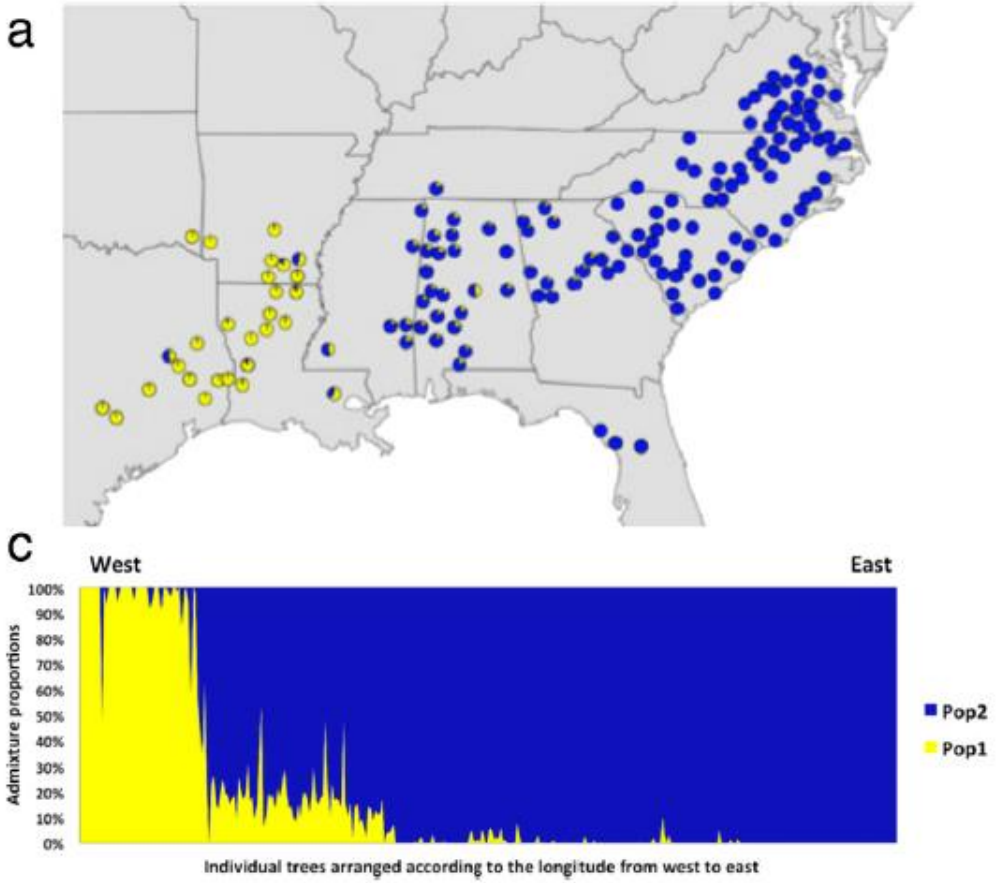
- We need ~800,000 reads per sample (good, because we will need to sequence hundreds of samples).

- ***Mutiplexing:*** Definitely necessary.

# Example 4: Genetic differences between populations of „loblolly pine“

- ***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 20Gbp. We need at least 20 Gigabytes of RAM and enough storage for all the reads coming from all samples.
- A node in the cluster can easily accommodate this analysis (remember, 24 CPUs and 64 Gigabytes of RAM).

# Example 4: Genetic differences between populations of „loblolly pine“



Source: https://doi.org/10.1186/s12864-016-3081-8

# Experimental design

- NGS course 2021
- https://gitlab.lrz.de/gustavo/ngscourse2021-tum/-/wikis/00.-home