ZMT LEIBNIZ CENTRE for Tropical Marine Research

Leibniz Association

Analysis of next generation sequencing data

11.03. – 15.03.2019

Workshop schedule

Date	Morning	Afternoon
Monday, 11.03.2019	Planning sequencing studies *	Linux command line *
Tuesday, 12.03.2019	Amplicon sequencing *	Tutorial I *
Wednesday, 13.03.2019	Tutorial II *	Assisted coding
Thursday, 14.03.2019	Multivariate data analysis *	Assisted coding
Friday, 15.03.2019	Introduction to git Q & A	Assisted coding

Collect questions and ideas during the course!

* Relevant for certificate

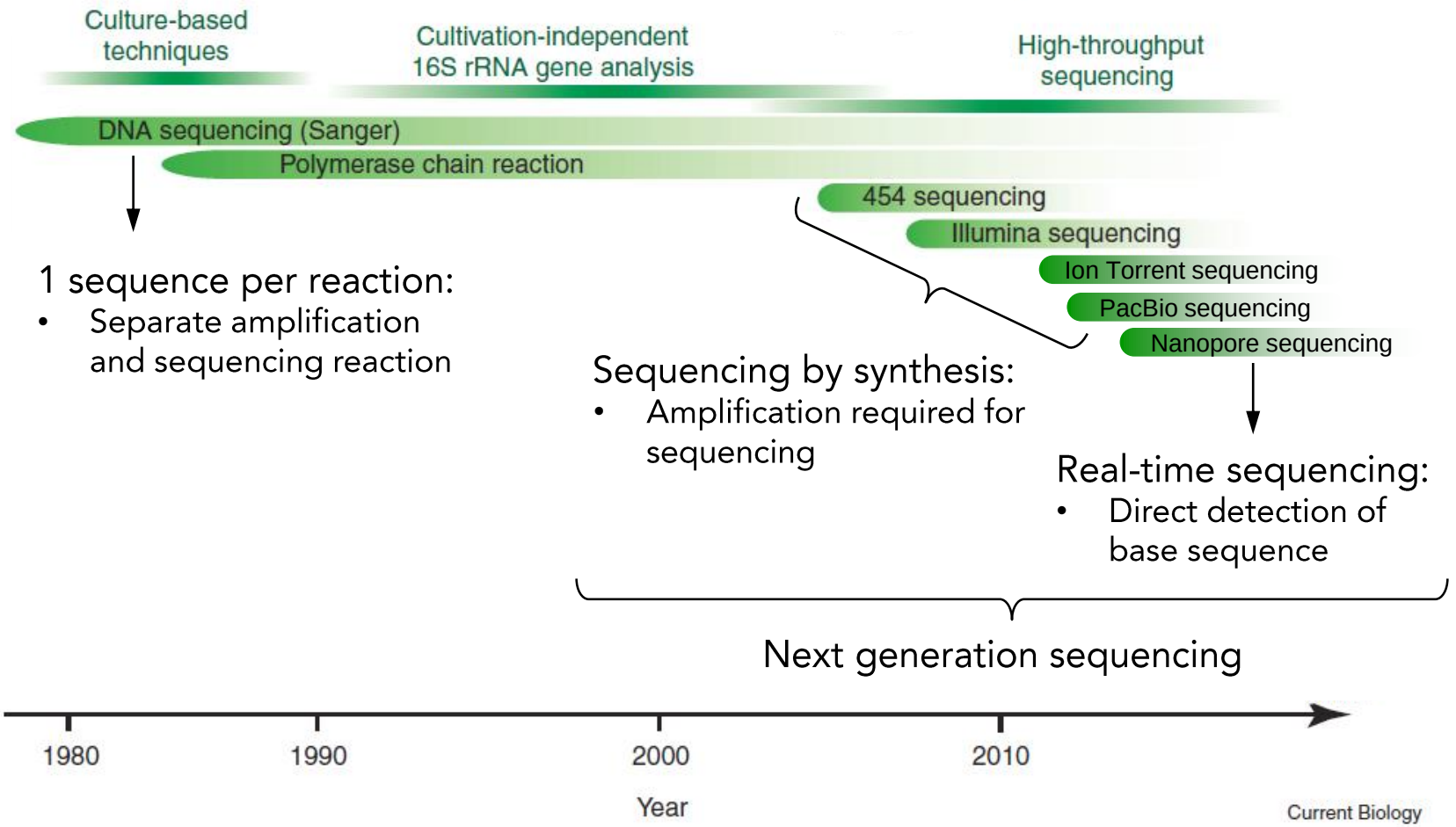
Planning sequencing studies

Content:

- Sequencing platforms
- Amplicon vs. shotgun sequencing
- Choosing your sequencing target
- Requirements for data analysis (computing facilities)
- Sampling and experimental design, power analysis
- Data archiving (11:00 - 12:00 Ivaylo Kostadinov)

https://zmtcloud.zmt-bremen.de/index.php/s/Mkgty4KxUpJ3qsi

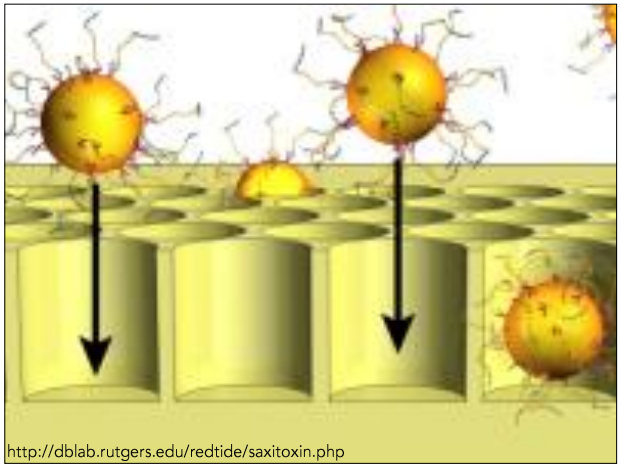
Sequencing platforms



Salazar & Sunagawa (2017)
https://doi.org/10.1016/j.cub.2017.01.017

NGS platforms

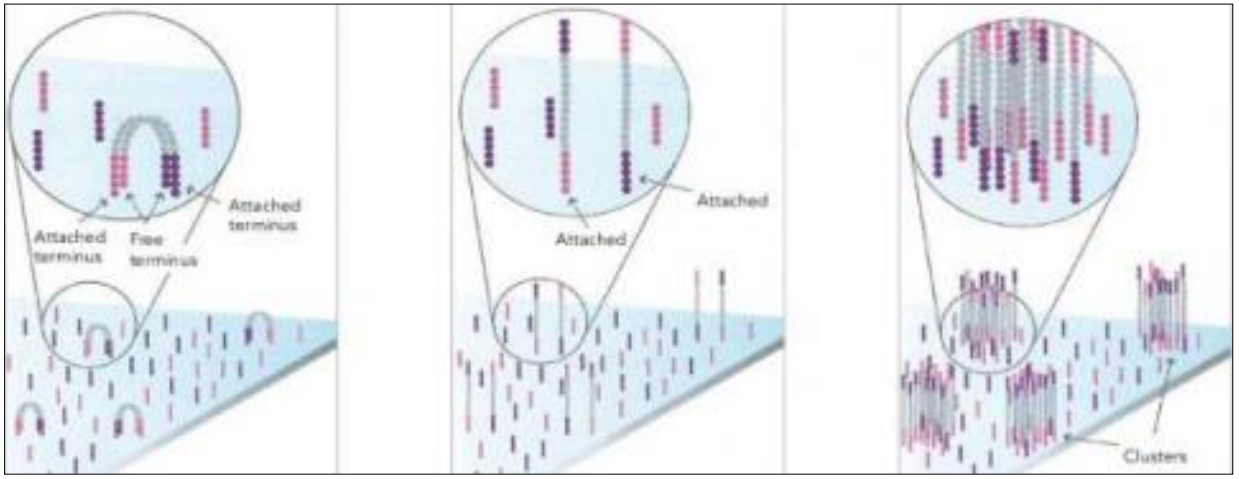
Sequencing technology	Principle	Read length	Errors	Comments
454	Light intensity ~ number of bases	~ 450 bp SE	homopolymers	discontinued



- Emulsion PCR
- Each nucleotide supplied separately in specified flow order
- Intensity of light signal ~ number of bases

NGS platforms

Sequencing technology	Principle	Read length	Errors	Comments
Illumina	Color ~ base	< 300 bp SE < 550 bp PE	substitutions	Most popular sequencing platform

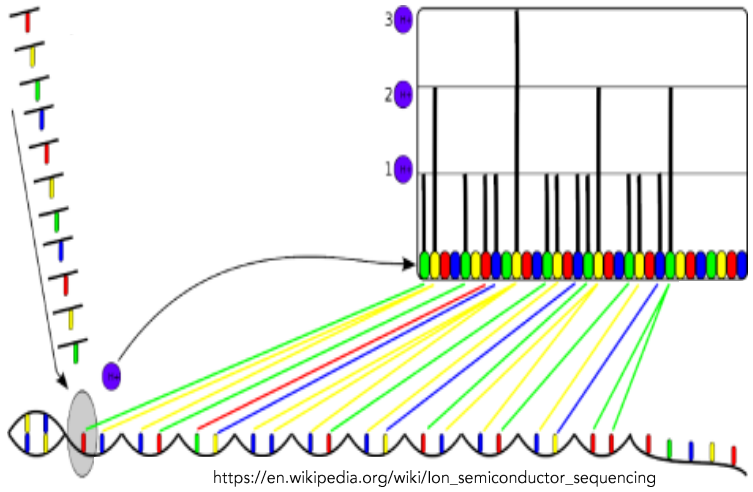


https://www.uppmax.uu.se/illumina-sequencing

- Sequencing one nucleotide at a time
- All nucleotides supplied at once, but with different 'color'
- https://www.youtube.com/watch?v=fCd6B5HRaZ8

NGS platforms

Sequencing technology	Principle	Read length	Errors	Comments
Ion Torrent	Current strength ~ number of bases	< 400 bp SE	homopolymers	

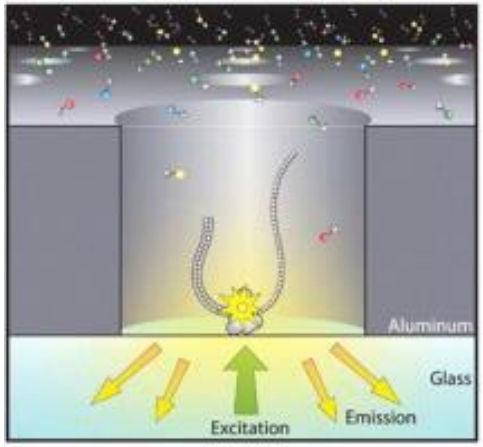


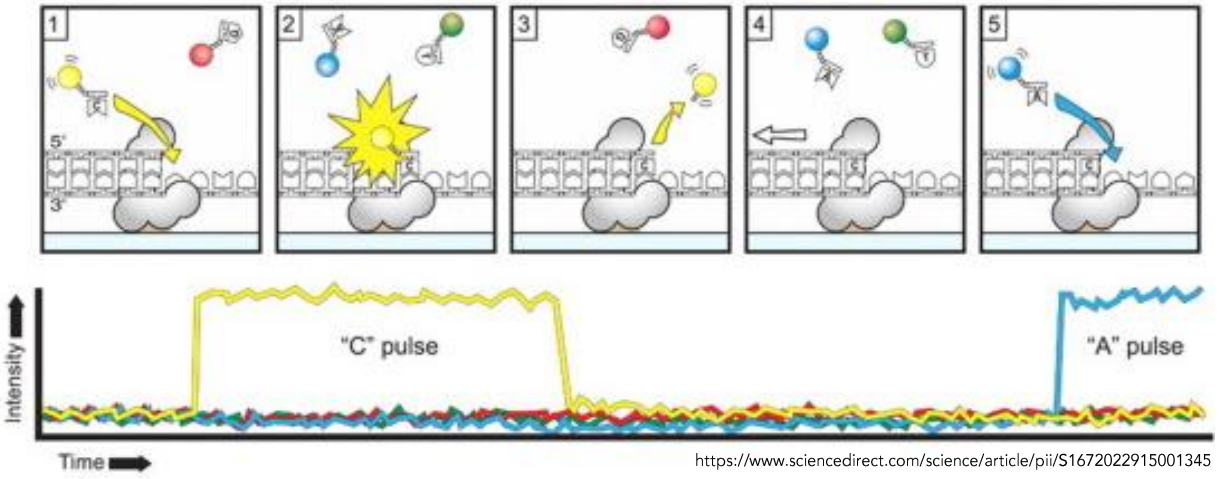
https://en.wikipedia.org/wiki/Ion_semiconductor_sequencing

- Electrical current ~ number of bases

NGS platforms

Sequencing technology	Principle	Read length	Errors	Comments
PacBio	Color ~ base	> 10 kb	~ 10% error rate (single pass)	



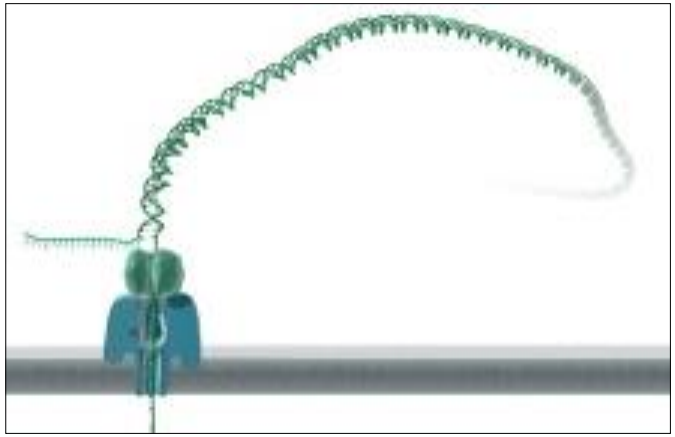


https://www.sciencedirect.com/science/article/pii/S1672022915001345

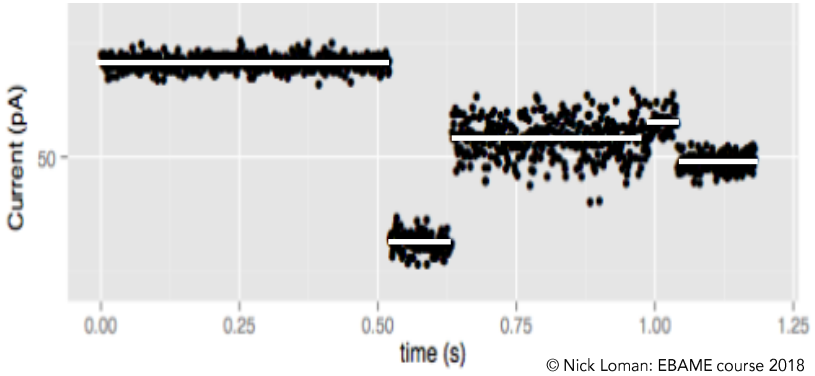
- DNA pulled through attached polymerase → base-dependent light signal

NGS platforms

Sequencing technology	Principle	Read length	Errors	Comments
Nanopore	Current ~ base Length ~ number	> 10 kb	< 4% error rate	In development, but very promising Epigenetics



https://www.nature.com/news/nanopore-genome-sequencer-makes-its-debut-1.10051

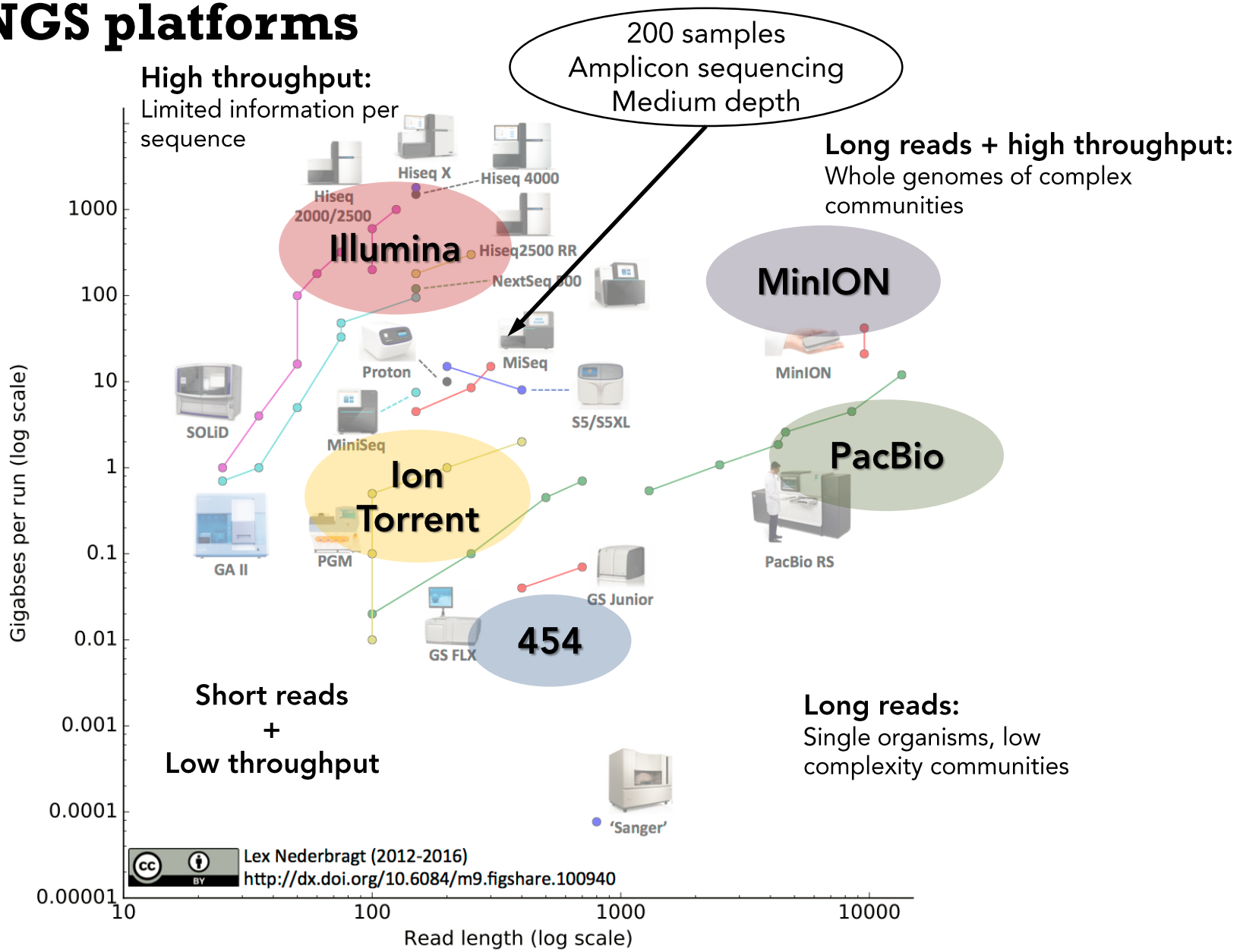


- DNA pulled through membrane-bound pore → electrical current

NGS platforms

Sequencing technology	Principle	Read length	Errors	Comments
454	Light intensity ~ number of bases	~ 450 bp SE	homopolymers	discontinued
Illumina	Color ~ base	< 300 bp SE < 550 bp PE	substitutions	Most popular sequencing platform
Ion Torrent	Current strength ~ number of bases	< 400 bp SE	homopolymers	
PacBio	Color ~ base	> 10 kb	~ 10% error rate (single pass)	
Nanopore	Current ~ base Length ~ number	> 10 kb	< 4% error rate	In development, but very promising Epigenetics

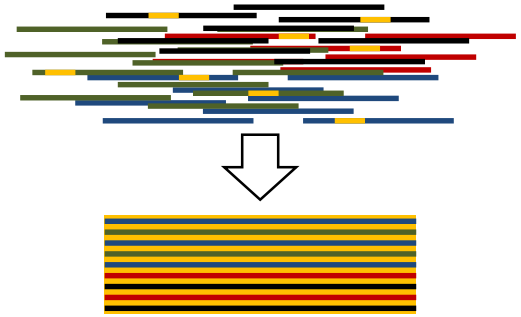
NGS platforms



Sequencing approaches

PCR-based (amplicon) sequencing:

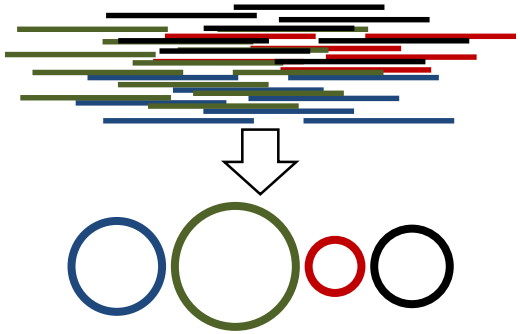
- Synonyms: metabarcoding, tag sequencing
- Marker gene
- PCR bias
- E.g. 16S screening



- Short reads
- High sample throughput
- Low sequencing depth
- 2x300bp paired-end Illumina

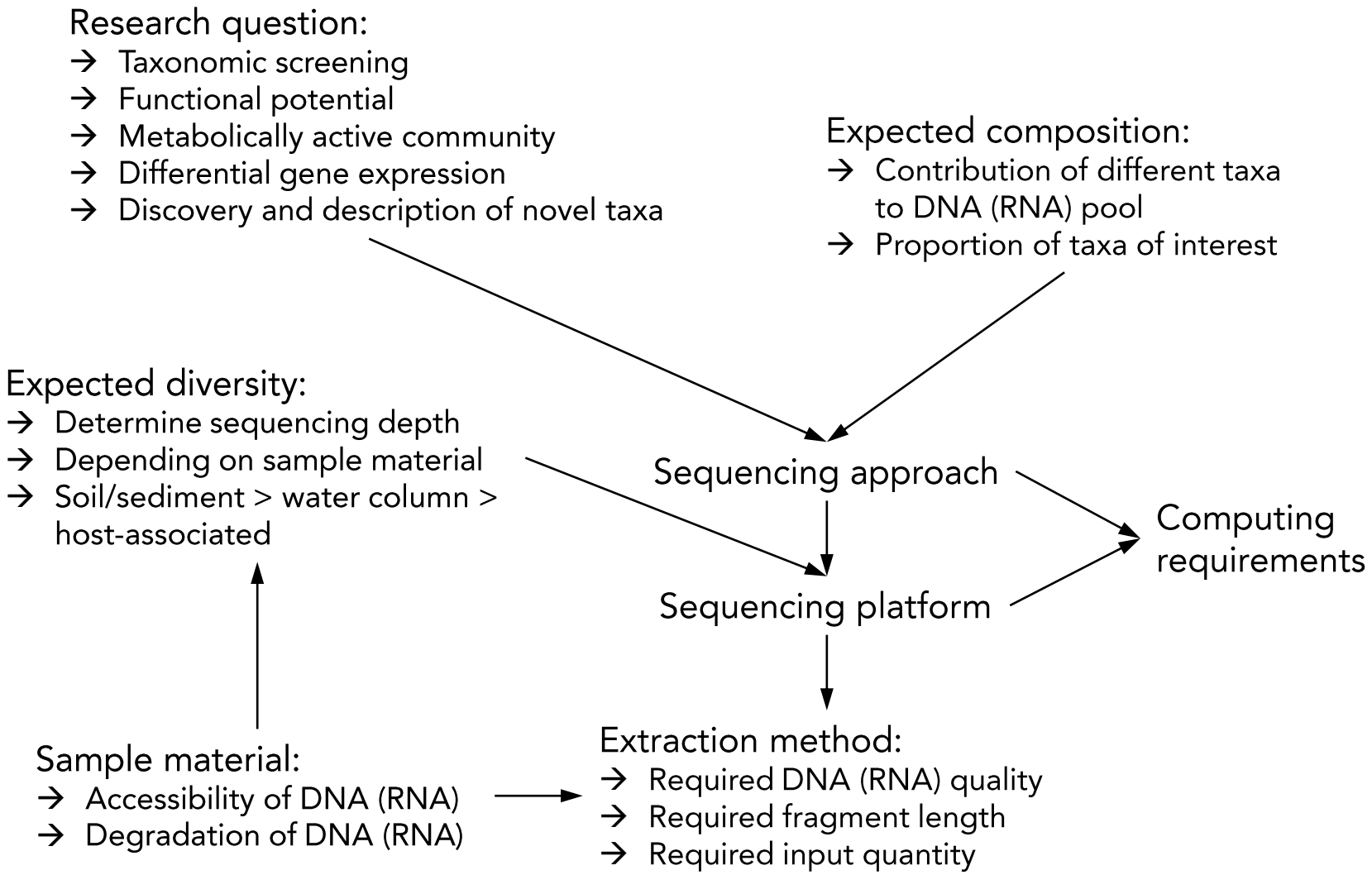
PCR-free (shotgun) sequencing:

- Not targeted
- No PCR bias
- E.g. metagenomics, metatranscriptomics

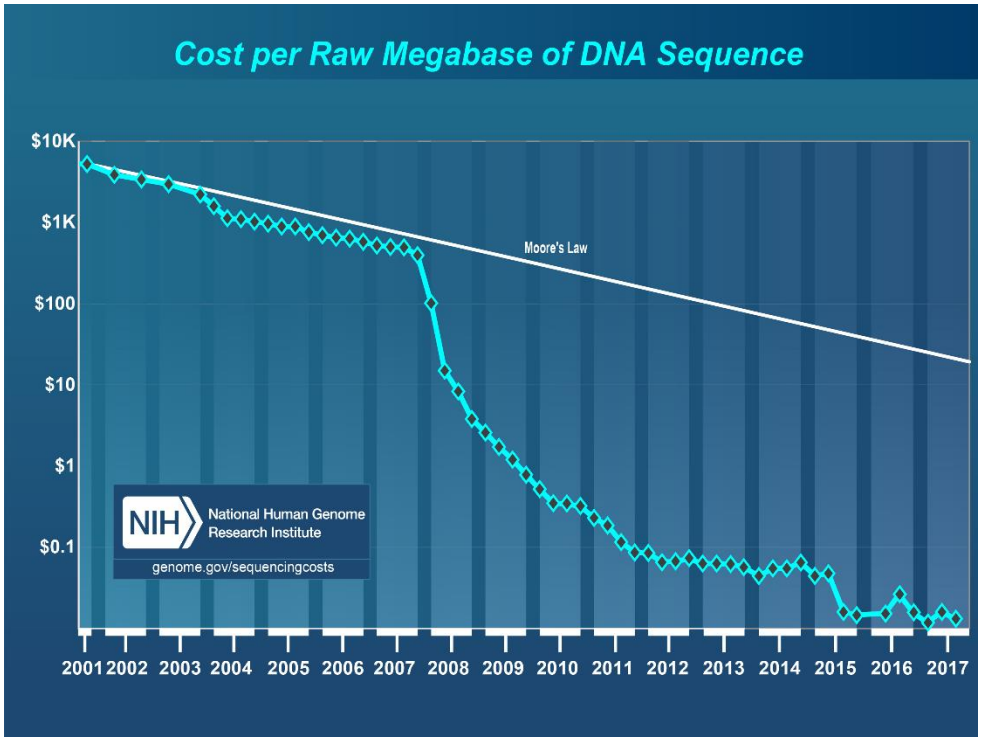


- Long reads
- Lower sample throughput
- Deep sequencing
- Illumina, PacBio, Nanopore

Choosing your sequencing target



Computing requirements



Are we producing more data than what we can handle?

200 samples
Amplicon sequencing
Medium depth

→

Memory: >32GB
CPU: >4
Disk space: ~100GB

+

Bioinformatics training

→

Higher requirements for shotgun data analysis

Sampling and experimental design

- No study without sampling design
- No analysis without appropriate sampling design
  - ‘post mortem’ of your data set
- Planning is more important than execution
  - Consider time, financial, legal, and ethical expenses
- Simpler designs are usually better than complicated ones

Access and benefit sharing:
→ Convention on biodiversity
→ Nagoya protocol

- Be aware what kind of data you are collecting: continuous, discrete, percentages (compositions), binary, etc.
- Be aware of the assumptions of the statistical tests suitable for your kind of data
- Be aware of the limitations of field, laboratory, and statistical techniques

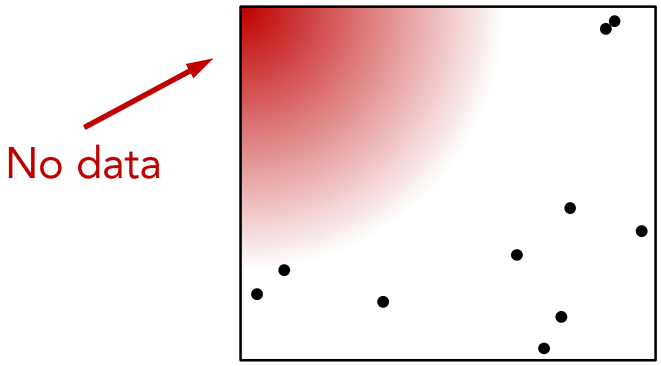
Know how to analyze your data before collecting it!

Study types

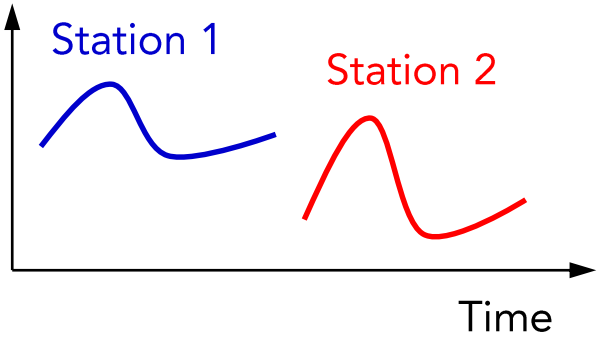
	Correlative study	Manipulative study
Pro	• Observations in natural system • Biologically relevant variation	• Controlled environment
Con	• Unknown, confounding factors • Accessability • Covariates • Correlation ≠ causation	• Bias through manipulation • Generalization to natural conditions
Examples	• Field studies	• Lab experiments

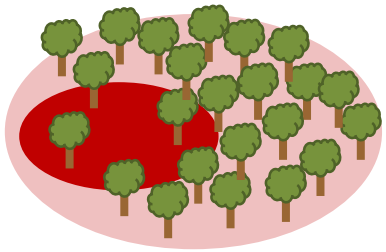
Randomization and bias

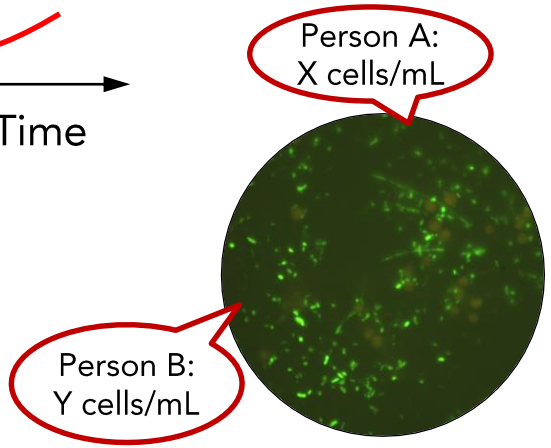
- Random sample = representative sample
- Mathematically random is not always representative
- Haphazard sampling



- Bias = Failure to obtain a representative sample
- Sample of convenience
- Time of sampling
- Observer bias
- Expectation bias

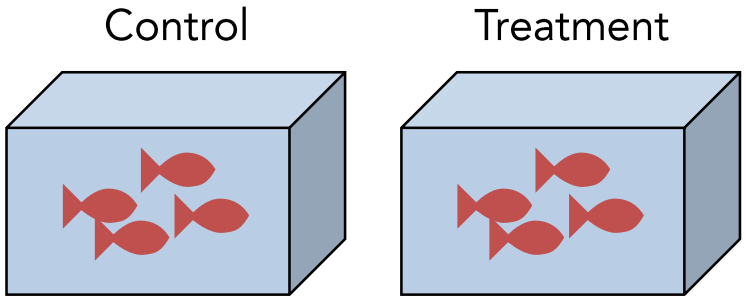


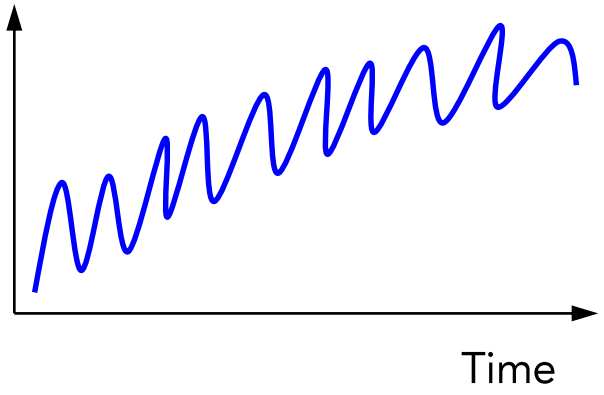


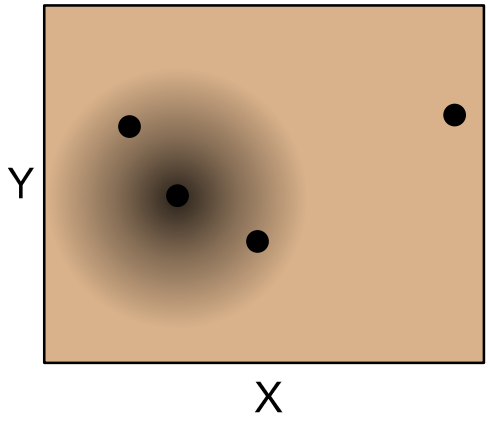


Pseudoreplication

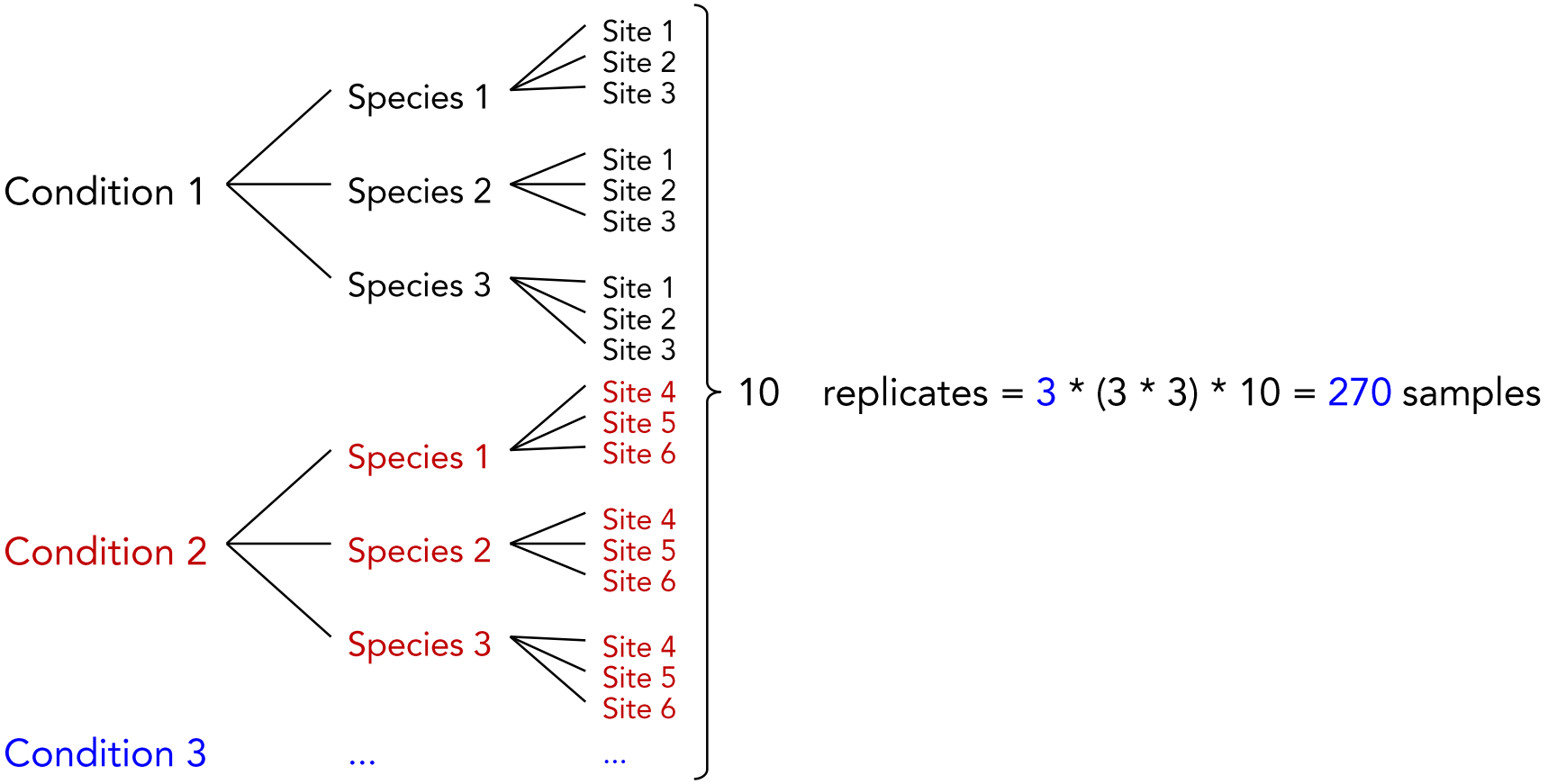
- Individual observations are **not independent**
- Other examples:
  - Common environment
  - Temporal and spatial autocorrelation
  - Duplicate measurements and technical replicates



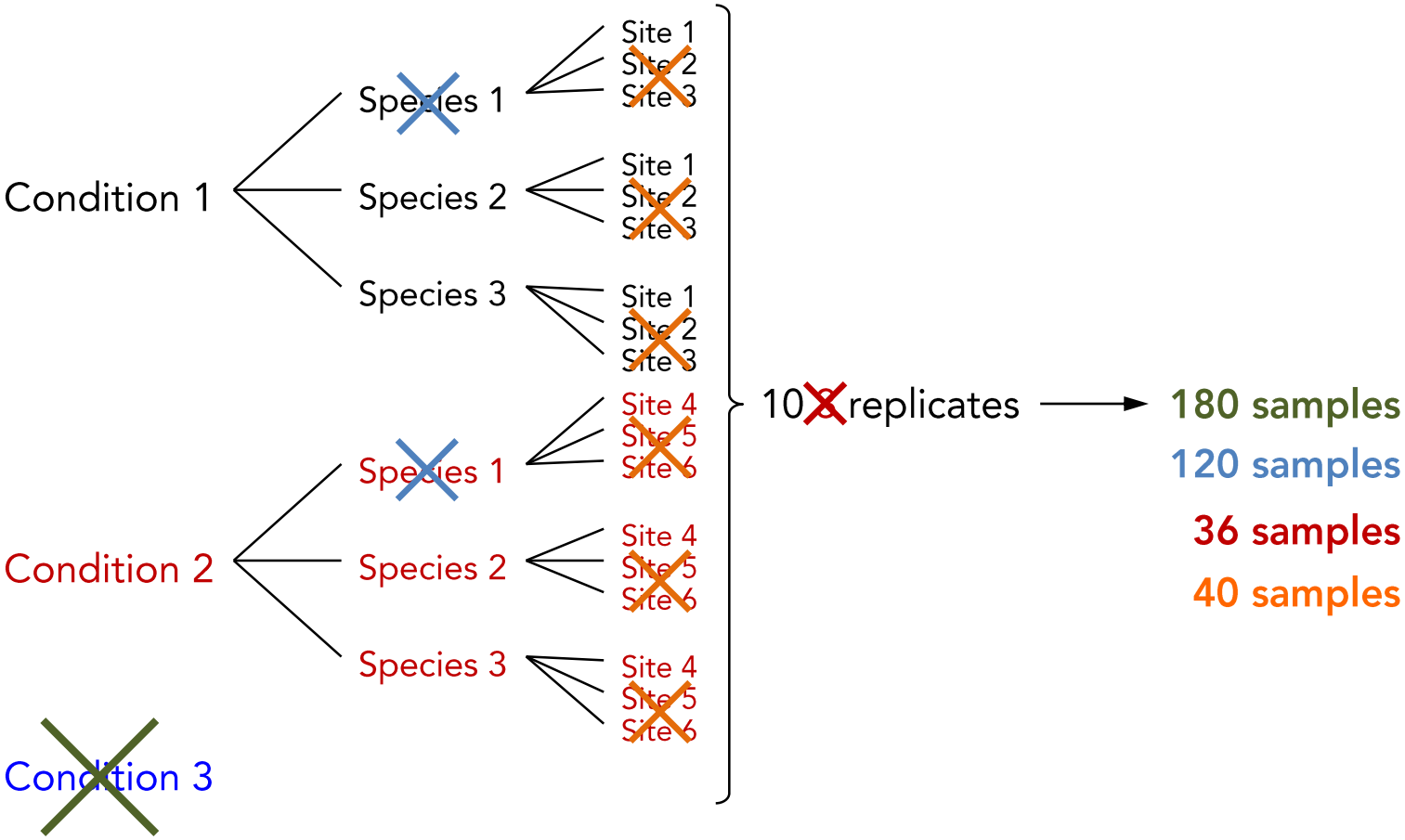




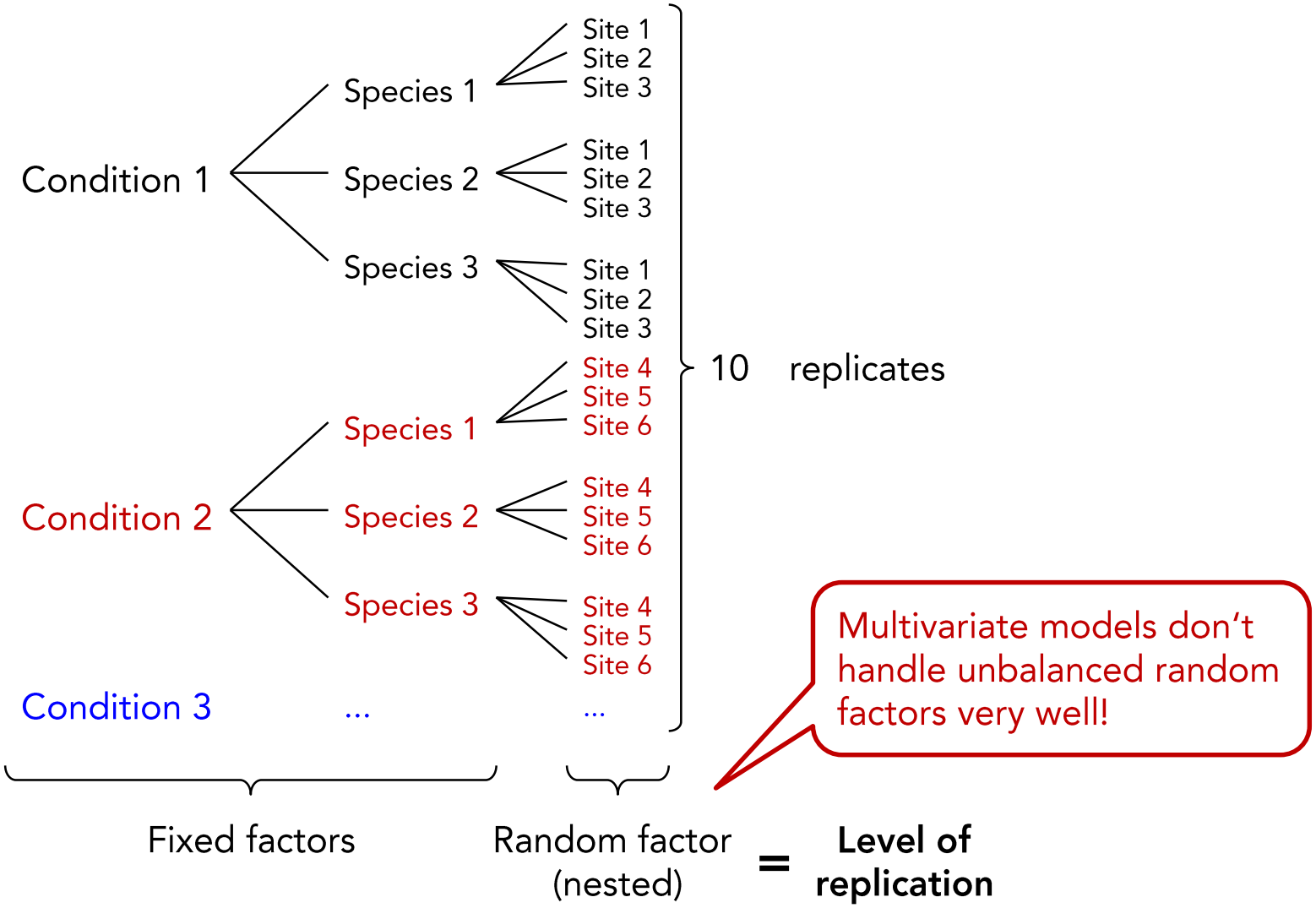
Sampling design trees



Sampling design trees



Sampling design trees

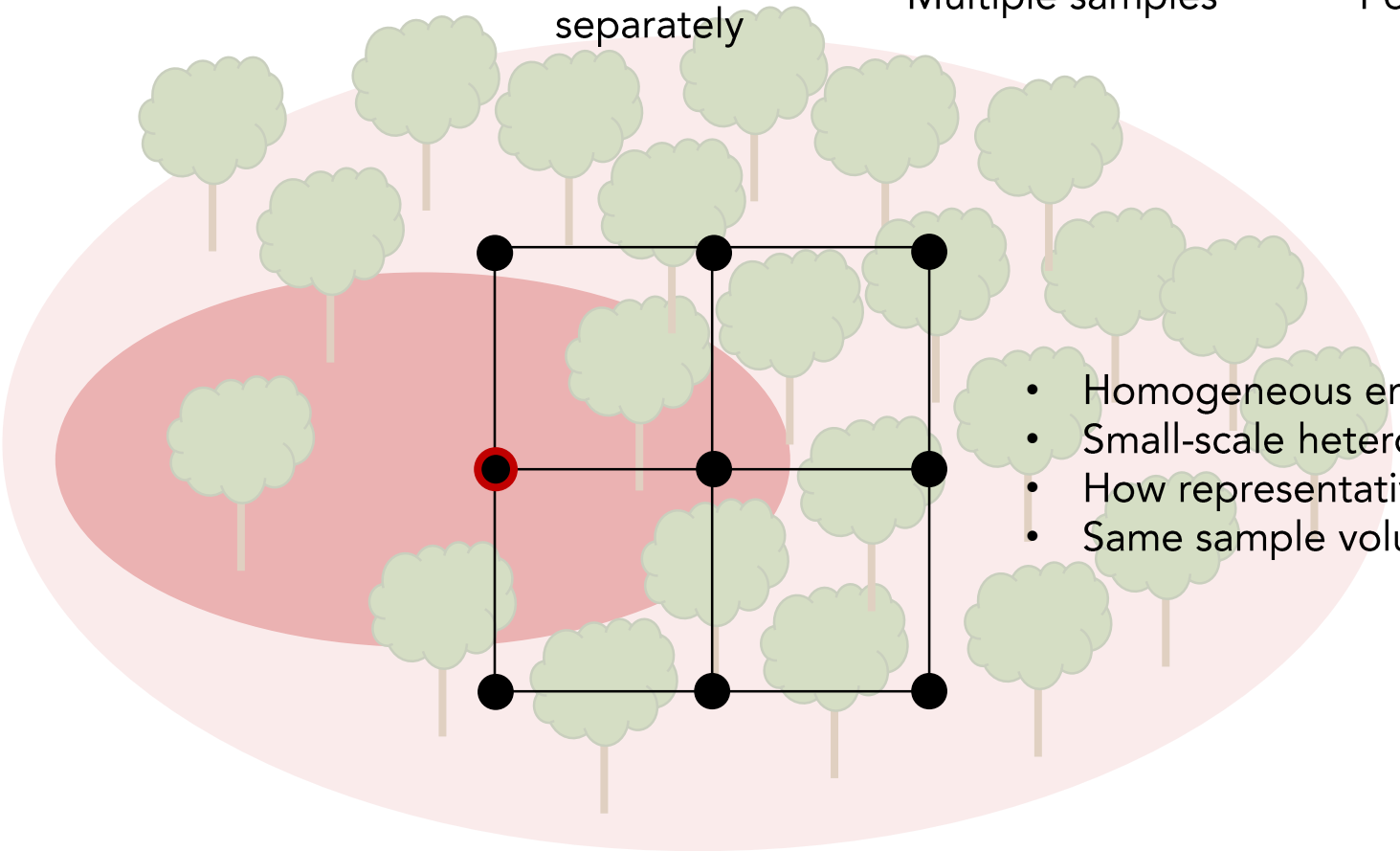


Sampling and sequencing strategies

Microbial communities are very heterogeneous and vary temporally/spatially at a small scale

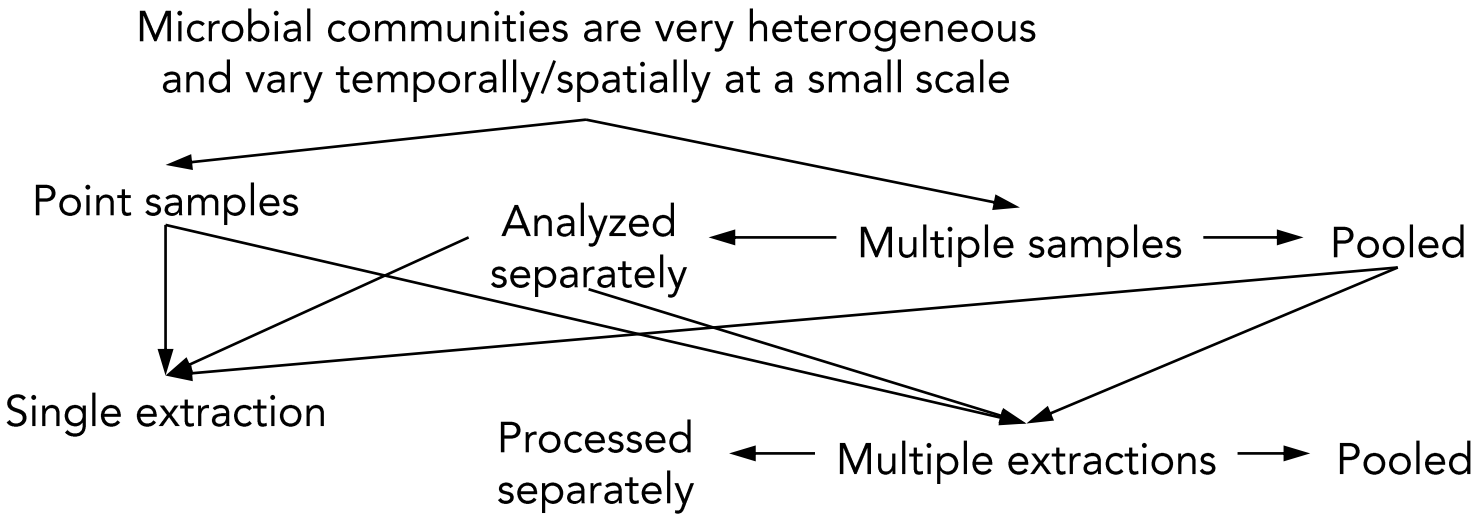
Point samples

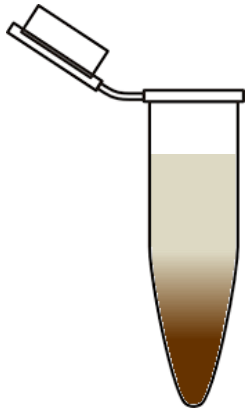
Analyzed separately ← Multiple samples → Pooled



- Homogeneous environment?
- Small-scale heterogeneity?
- How representative is the average?
- Same sample volume!

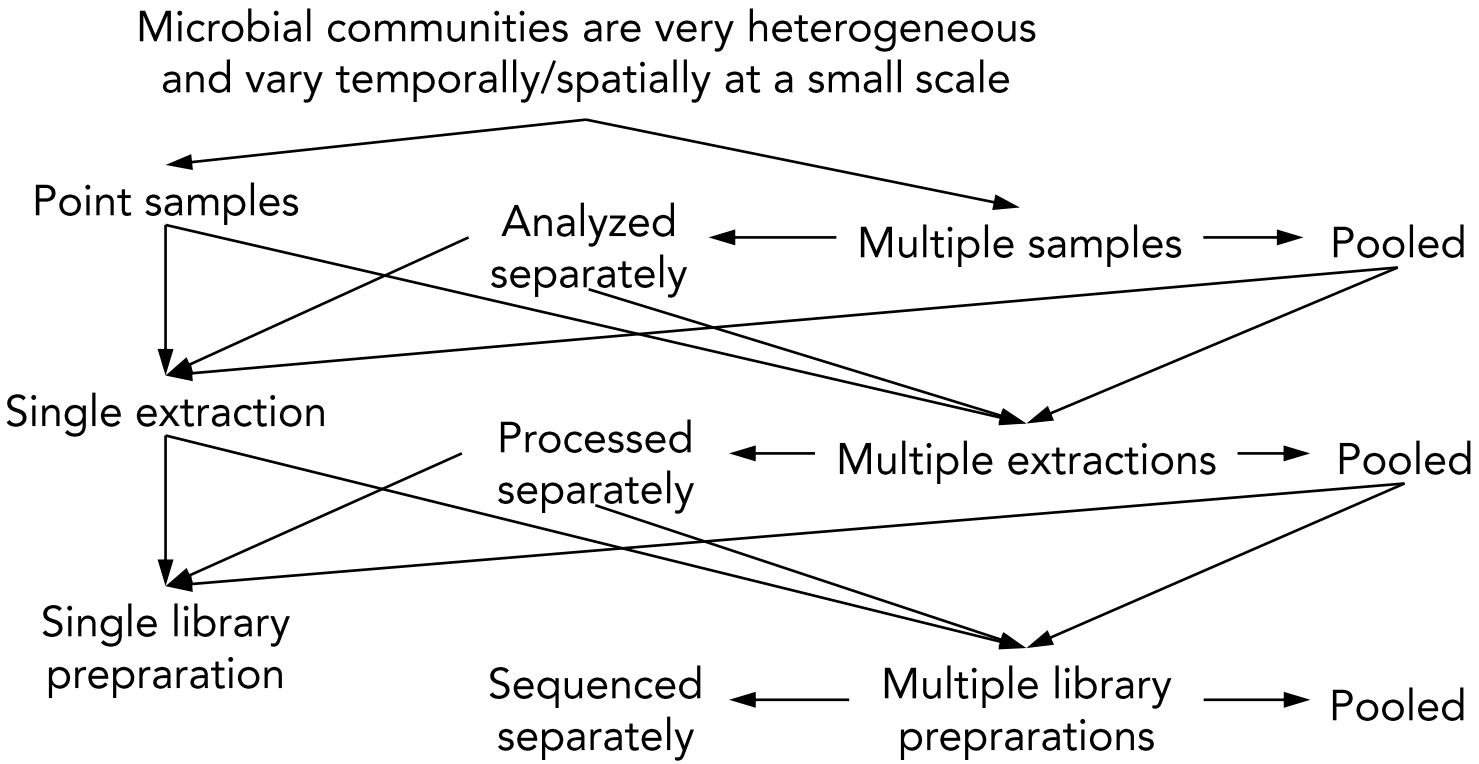
Sampling and sequencing strategies

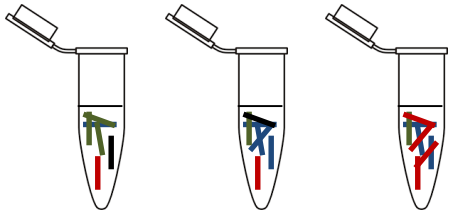




- Efficiency of sample homogenization?
- Random variation?
- Extraction yield?

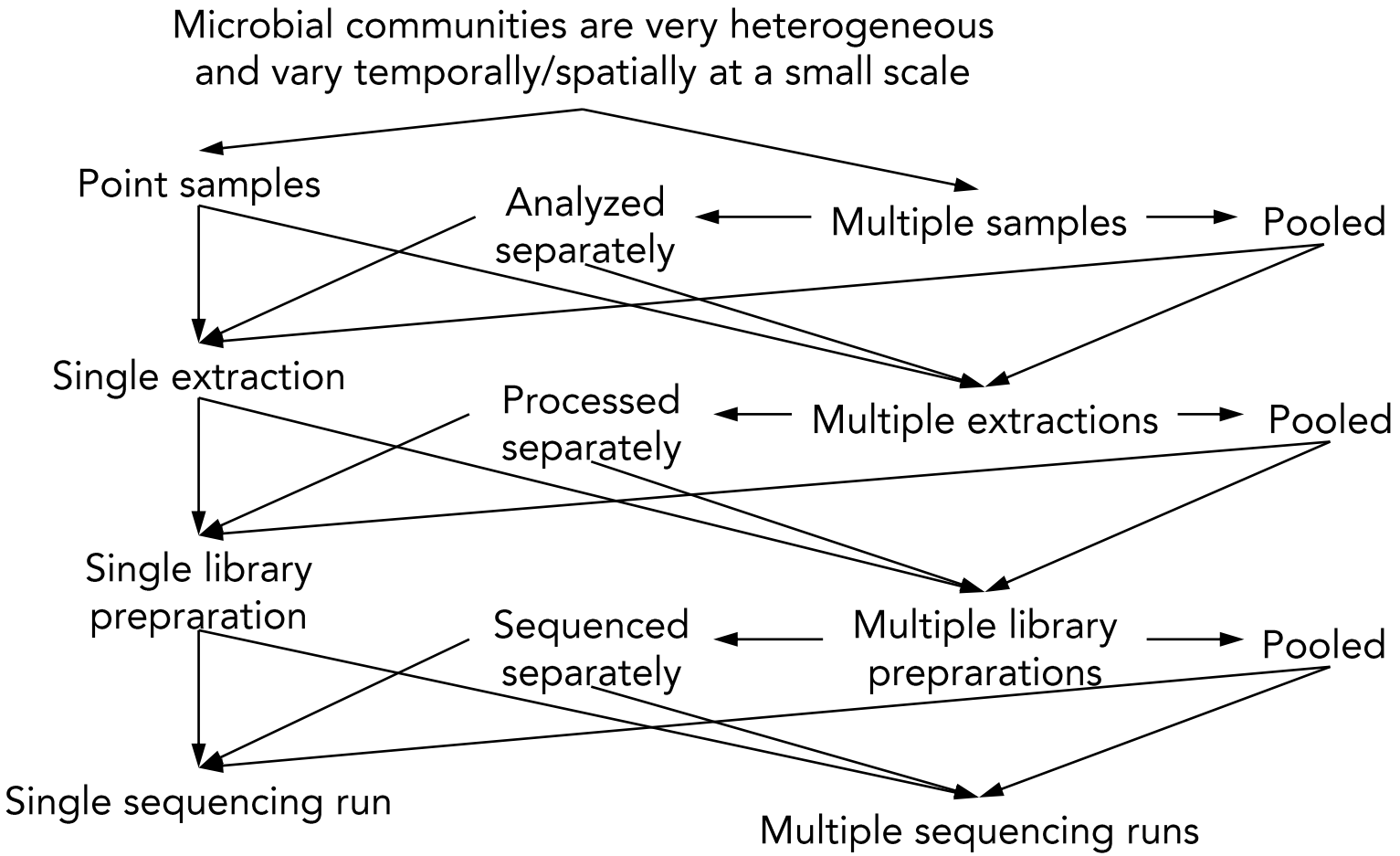
Sampling and sequencing strategies



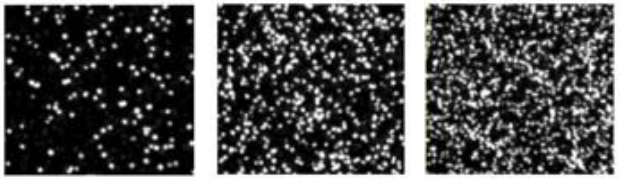


- Pipetting variation?
- Random variation?
- Differences between replicate PCRs?

Sampling and sequencing strategies



- Sufficient sequencing depth?
- Sequencing error profiles?
- Random variation?



https://www.well.ox.ac.uk/ogc/sequencing-quality-monitoring-run/

Confounding factors

- What else varies with your factor of interest?
- Wrong conclusions about effect
- ➢ Manipulative studies: controls
- ➢ Correlative studies: detailed understanding of system

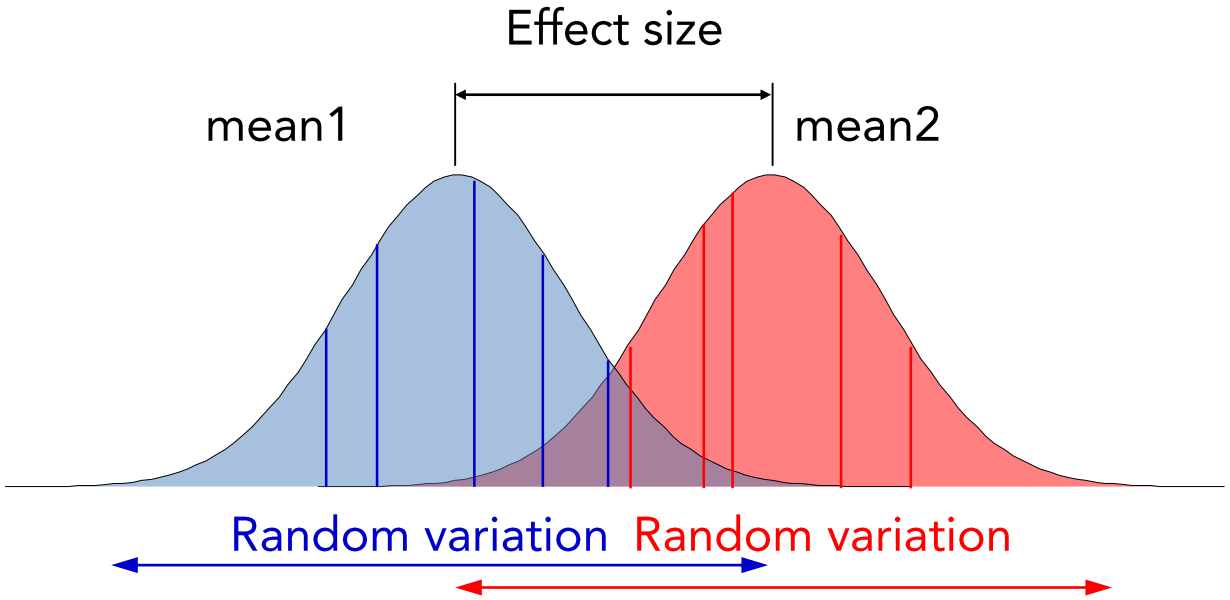
http://msandersonbio.weebly.com/exp-design.html

Power analysis

- Statistical power = probability to detect difference if there is one
- Depending on:
  - Effect size
  - Random variation
  - Sample size

Educated guesswork

This we can modify



Plans vs. Reality

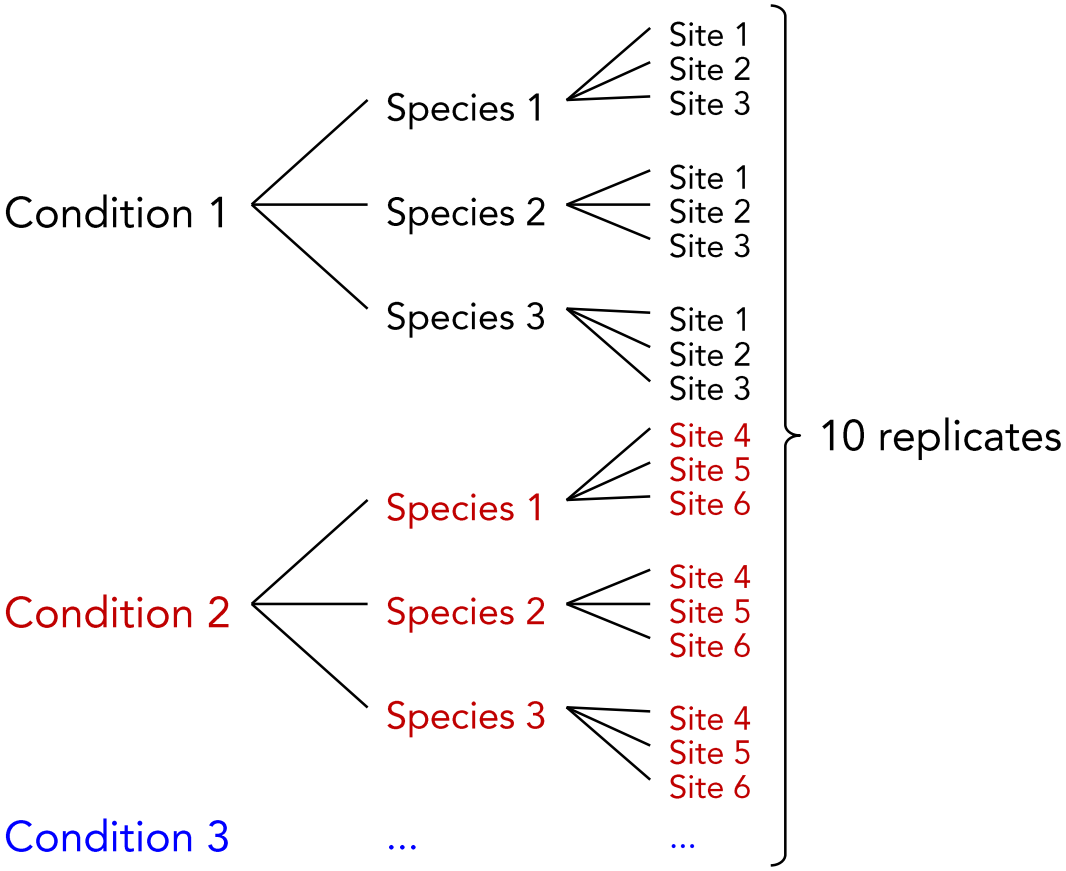
- > 10 replicates for each treatment
- Additional technical replicates
- Balanced sampling design
- Normally distributed data
- No missing data
- No outliers
- No (observer) bias
- No confounding variables

- ~ 3 replicates because of logistic constraints
- Technical replicates not comparable
- Unbalanced sampling design
- Irregular data distribution
- Missing data due to failed measurements
- Many outliers
- Strong biases
- Highly confounded environmental data
- …

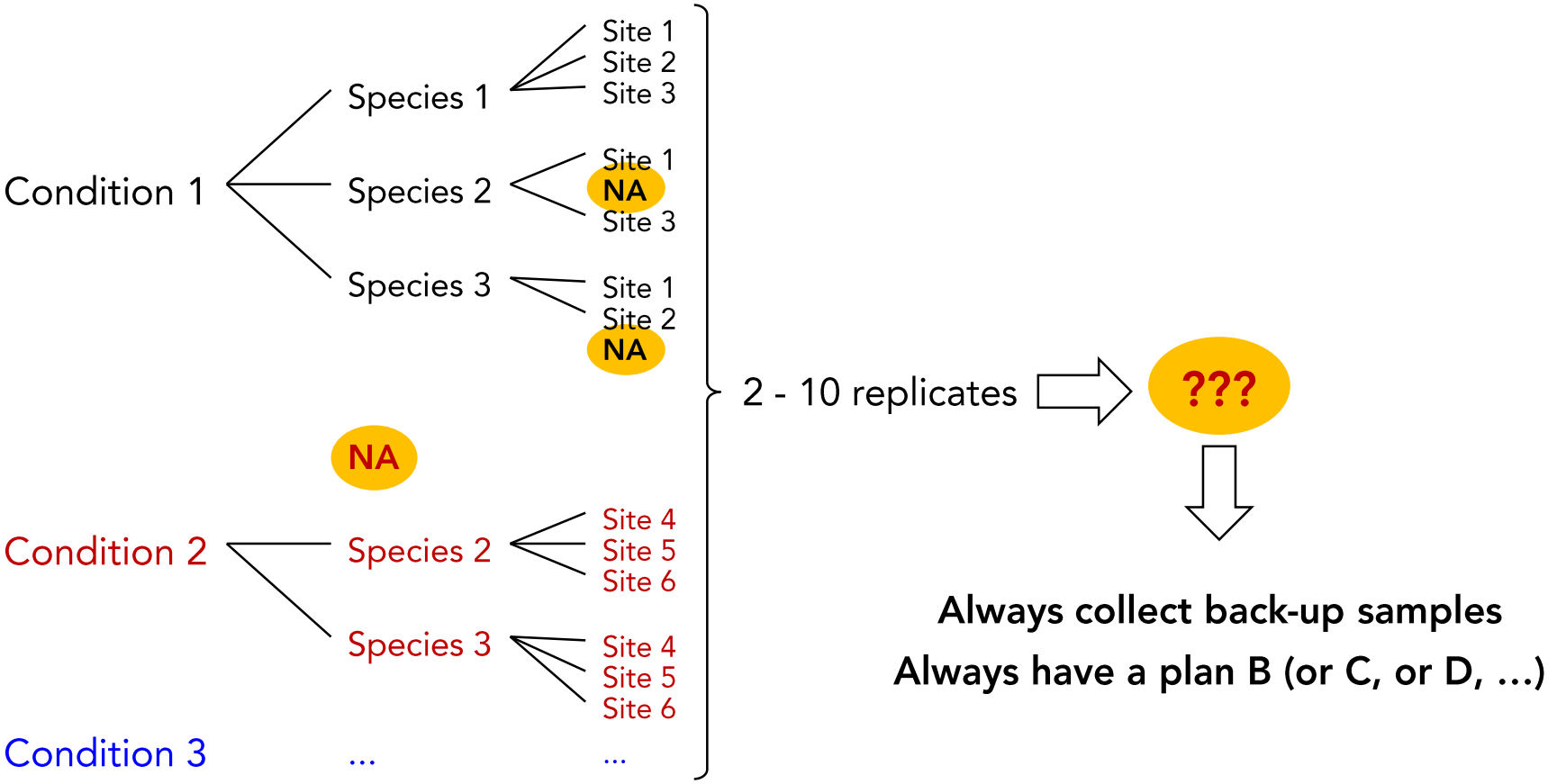
Compromise

Irregular data distributions:	Simple sampling design Consider non-parametric tests and/or permutation tests
Logistic constraints:	Instead of reducing the number of replicates, reduce the number of treatments
Missing data:	Collect more samples than necessary, even if they are not going to be analyzed

Plans vs. Reality



Plans vs. Reality



Always collect back-up samples

Always have a plan B (or C, or D, …)

Check list

- What are your **hypotheses**?
- Which **conditions** do you want to compare?
- How many **factors** do you have, with how many **levels**?
- What kind of **design** is best suited for your hypotheses?
- Which **kind of data** are you working with?
- How much **time** and **money** do you have?
- How many **replicates** are feasible?
- What is your level of **independent** replication?
- How do you want to **analyze** your data?
- Which **tests** are suitable for your data and experiment, and which **assumptions** have to be met?
- What are your **plans B**, **C**, **D**, ...?

Good scientific practice

- Research must be reproducible!
  - Take detailed notes and write code (that others can understand)
  - Document any modification to raw data (scripts)
  - Back-up your work immediately/regularly (electronic and hard copy)
  - Archive your data: raw data and code!

" We forgot to back up our files, so we're asking everyone to remember everything they have typed during the past 10 days. "

https://www.linkedin.com/company/clifford-technologies-ltd

Data archiving

Dr. Ivaylo Kostadinov

https://www.gfbio.org/