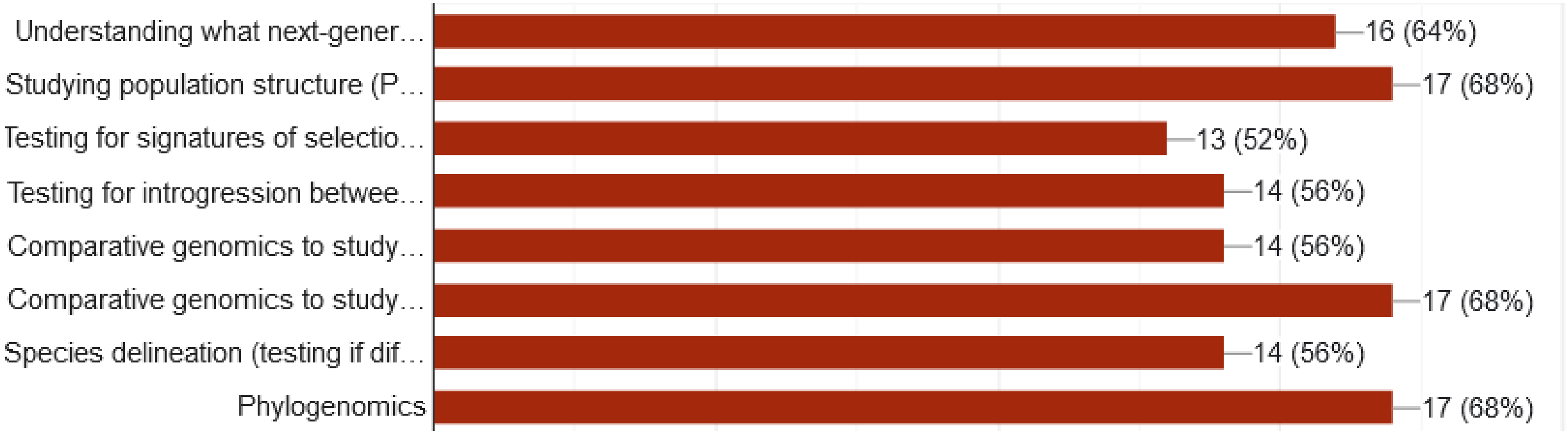
Welcome to the Biodiversity Genomics course!

Joana Meier, Karin Näsvall, Nicol Rueda, Patricio Salazar
Tree of Life Programme, Wellcome Sanger Institute

What you wanted us to cover in the course



Basically everything we offered! We will thus give an introduction to each of these fields

Tentative schedule

- Saturday, 20 July:
  - introduction to biodiversity genomics, Linux, NGS data
  - Filtering and quality checks of Illumina data
- Sunday, 21 July:
  - Mapping reads to a reference genome
  - Variant calling and filtering vcf files
- Monday, 22 July:
  - Population structure with PCA and phylogenomics
  - Genome scans to identify regions under selection
- Tuesday, 23 July:
  - Detecting introgression
  - Comparative genomics
- Wednesday, 24 July:
  - Phylogenomics with reference genomes
  - Buffer, topics of interest & discussing own projects

Daily sessions:

- 9:00-10:30: First session (1.5 hours)
- 10:30-11:00: Coffee break (30 min)
- 11:00-12:30: Second session (1.5 hours)
- 12:30-1:30: Lunch break (1 hour)
- 1:30-3:30: Third session (2 hours)
- 3:30-4:00: Coffee break (30 min)
- 4:00-6:00: Fourth session (2 hours)

Researchers who contributed to organising this course and can help with questions about IKIAM / Tena

Dr Patricio Salazar

Lead organiser

Franz Chandi

Chief of lunch breaks

Kimberly Gavilanes

Chief of coffee breaks & facilities

Alex Arias

Chief of class room infrastructure

María José Sánchez

Jack of all trades

Prof Caroline Bacquet

Key facilitator

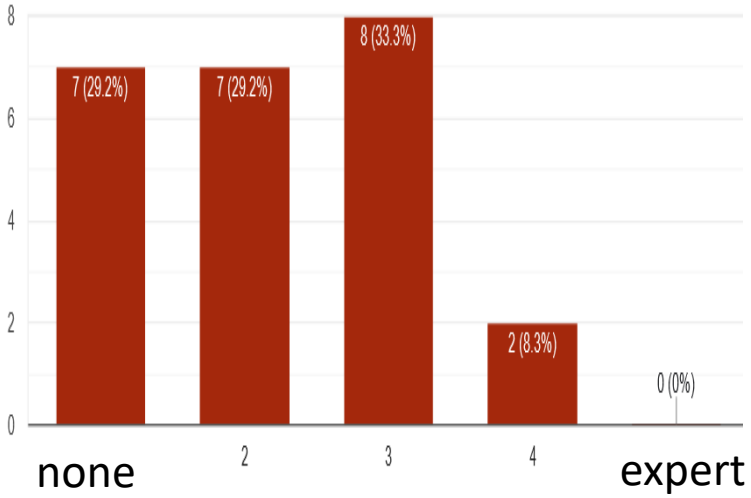
A few tips and housekeeping rules

- Be kind to each other
- Sexism, racism, or any other kind of unfair behaviour is not tolerated. Please let me, Patricio, Nicol or Karin know if you experience or see anything like this.
- To ask questions, please raise your hand. Questions are very much encouraged. If you do not understand something, just ask!
- If you struggle with something, put the green card on your desk and someone will come to help you.
- All course materials (including slides) can be found here: **github.com/rapidspeciation/biodiversity_genomics_course**

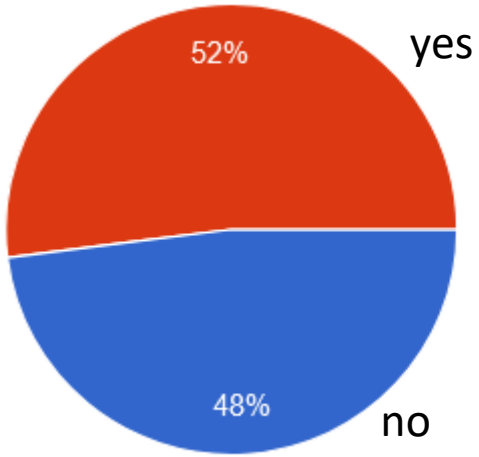
We are a very diverse group, which is great!

- Always feel free to ask questions at any time. It is likely that at least one other person is sooo happy that you asked.
- Our diversity is a big advantage. Let's learn from each other!

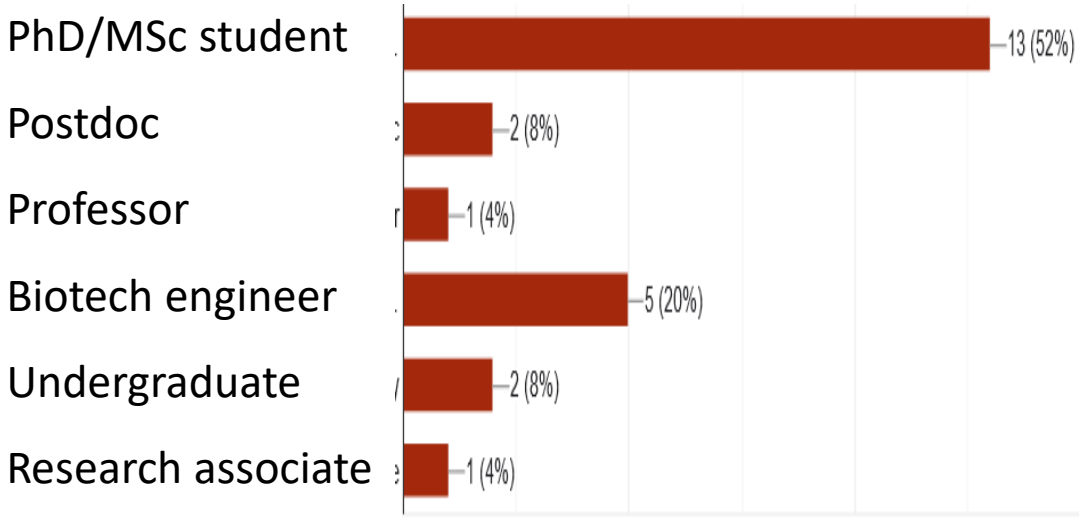
Linux experience



Prior experience with NGS data



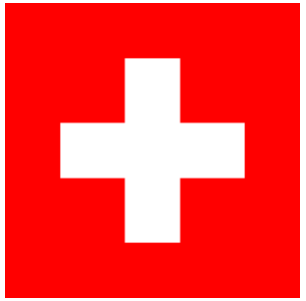
Current academic level/job



Please, introduce yourself with your name and if you wish also your country, study organism or research question

Joana Meier

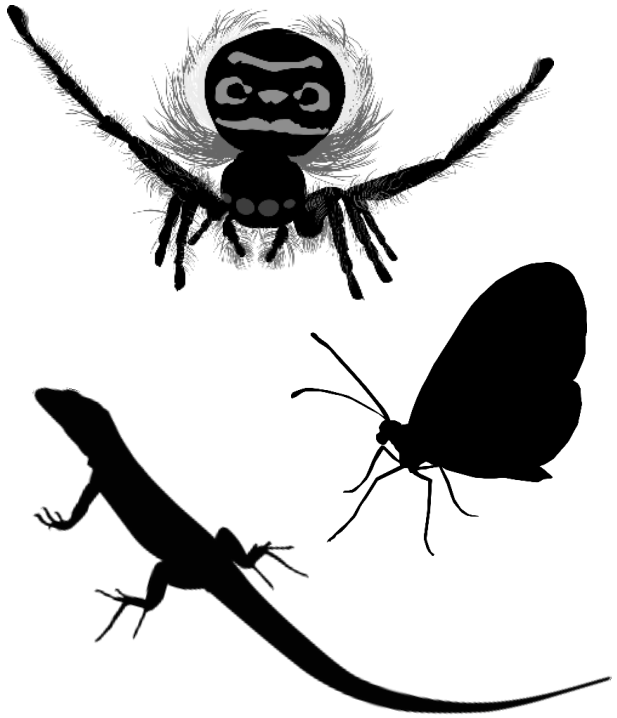
Childhood

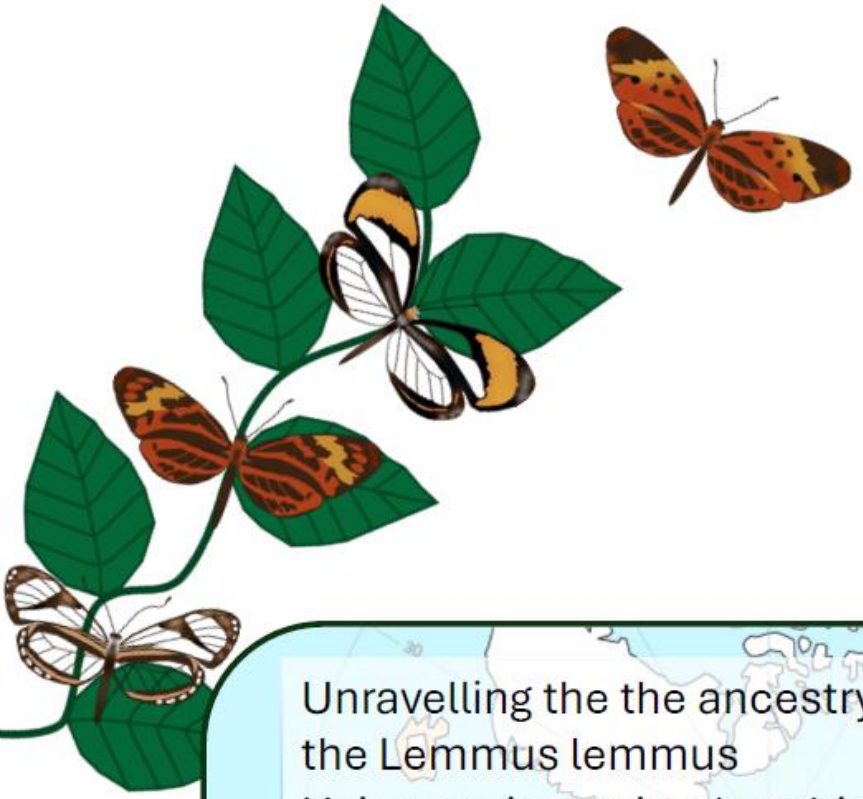


PhD on cichlid fish Speciation in Bern, Switzerland

Postdoc at the University of Cambridge, UK

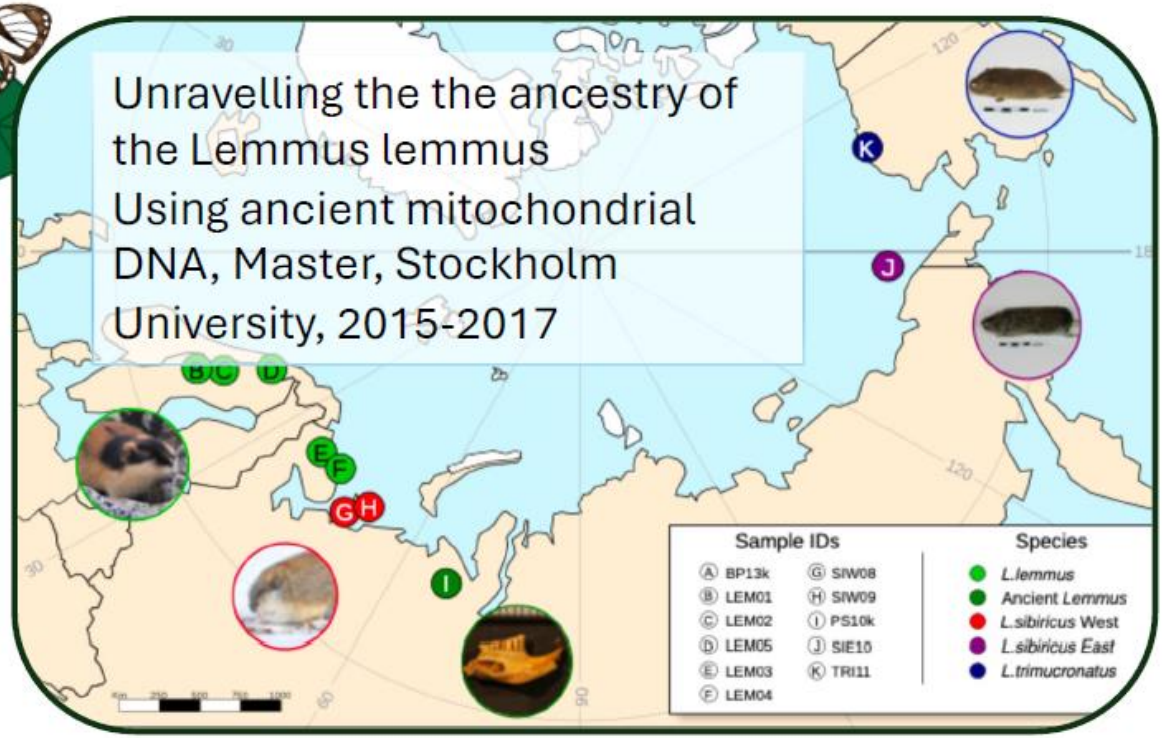
Group leader at the Wellcome Sanger Institute



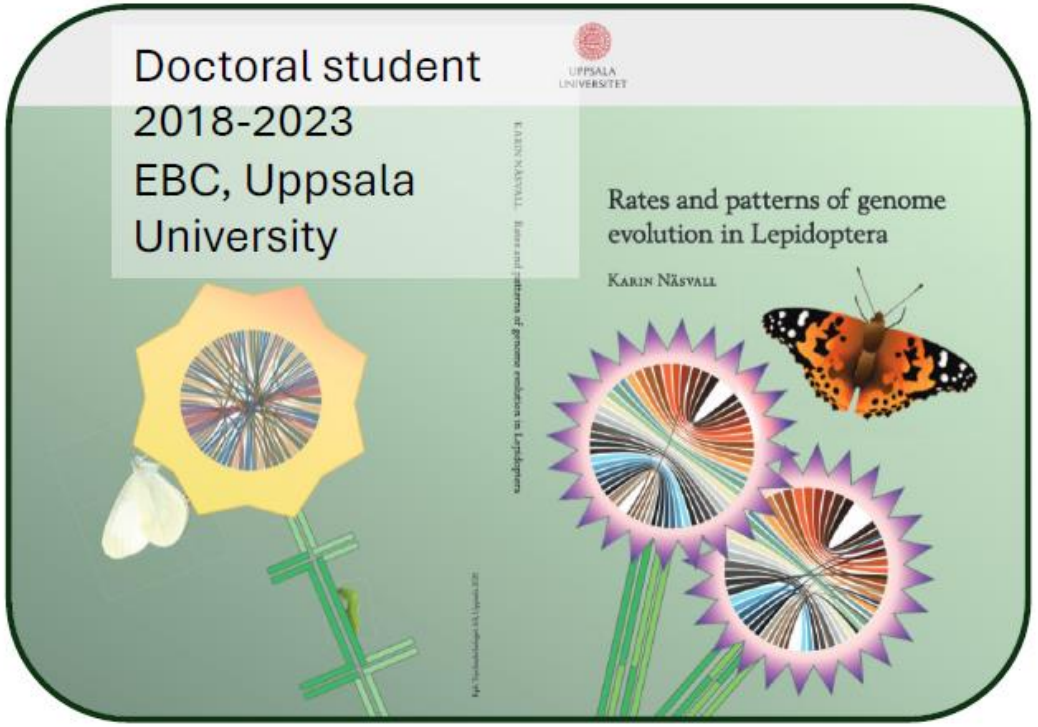


Karin Näsvall, PhD

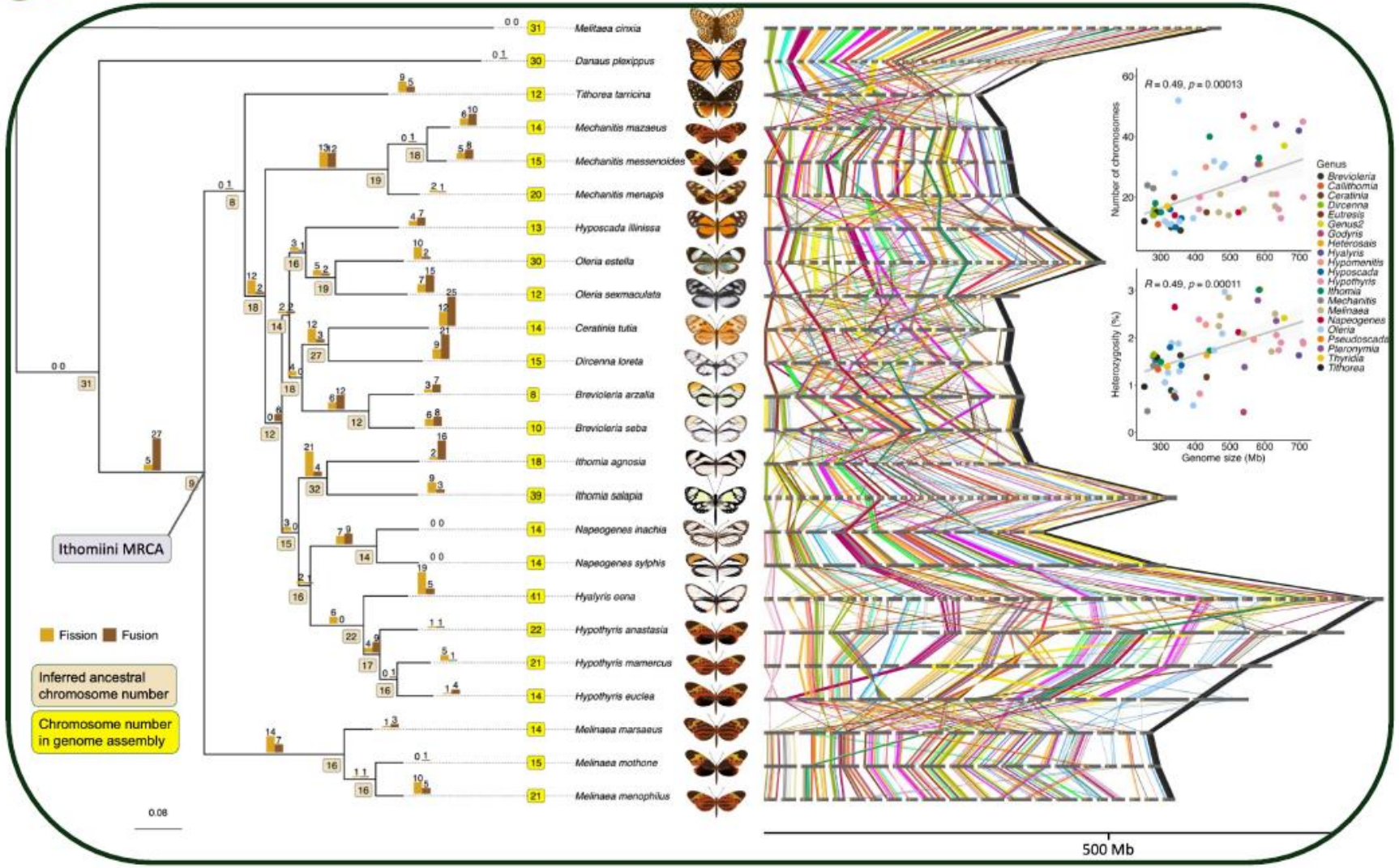
Postdoctoral fellow
Rapid speciation group, Tree of Life
Wellcome Sanger Institute, UK



Unravelling the the ancestry of the Lemmus lemmus Using ancient mitochondrial DNA, Master, Stockholm University, 2015-2017



Doctoral student
2018-2023
EBC, Uppsala University



Association between speciation rate
And chromosomal rearrangements in Ithomiini butterflies

Nicol Rueda

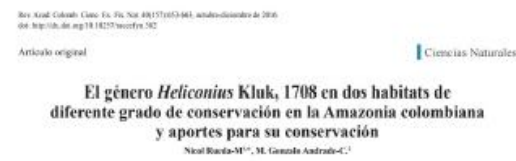
Colombia

- **BSc in Biology - Colombia**
- **Master in Species Conservation in trade (CITES)**
  Universidad Internacional de Andalucia - Spain
- **Master of Science in Biology**
  Universidad Nacional de Colombia
- **PhD candidate in Biology**
  Universidad del Rosario - Colombia

Master of science in Biology

Population monitoring for 7 months in two forests of different conservation status.

Capture-Mark-recapture technique



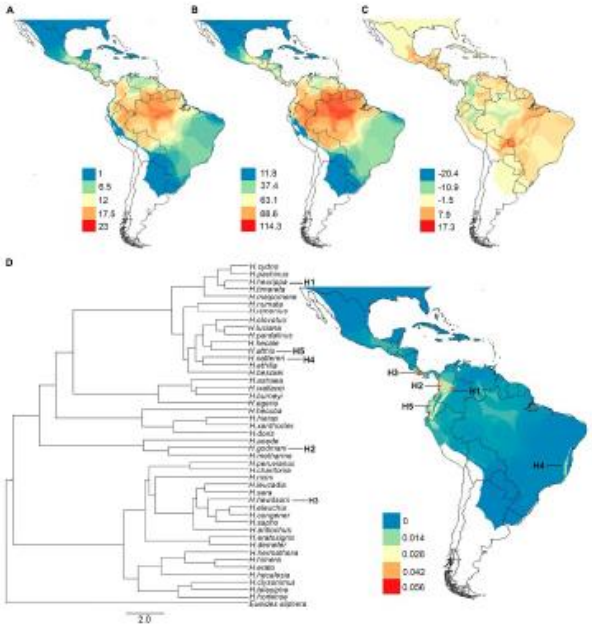
Rev. Acad. Colomb. Cienc. Ex. Fís. Nat. 40(157):653-663, octubre-diciembre de 2016
doi: http://dx.doi.org/10.18257/raccefyn.382

Artículo original | Ciencias Naturales

El género *Heliconius* Kluk, 1708 en dos habitats de diferente grado de conservación en la Amazonia colombiana y aportes para su conservación

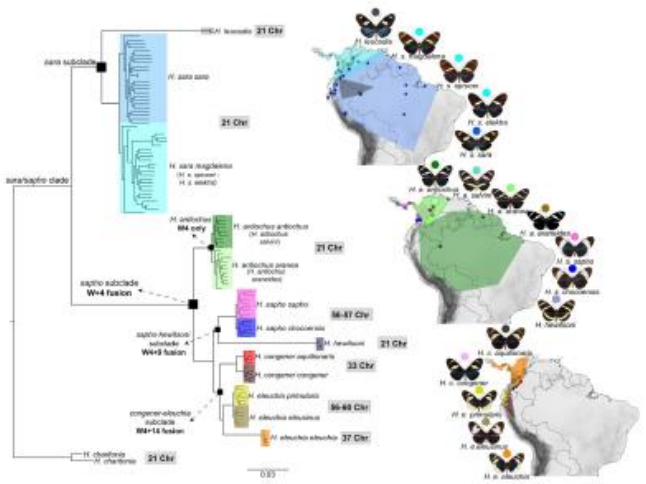
Nicol Rueda-M[1*], M. Gonzalo Andrade-C.[1]

Drivers of diversification in *Heliconius*, with special focus on the *sara/sapho* clade



frontiers in Ecology and Evolution

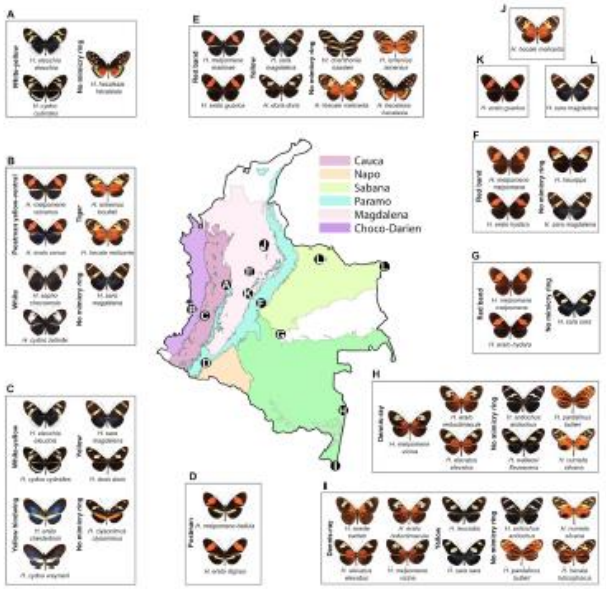
2021 Environmental Drivers of Diversification and Hybridization in Neotropical Butterflies



PLOS GENETICS

2024 RESEARCH ARTICLE

Genomic evidence reveals three W-autosome fusions in *Heliconius* butterflies



WORK IN PROGRESS

Introduction to Biodiversity Genomics

Using genomics to understand and preserve biodiversity from genetic diversity, populations, to species and ecosystems

Resolving the taxonomy

- Placing potentially new species
- Species delineation

Adaptation and speciation

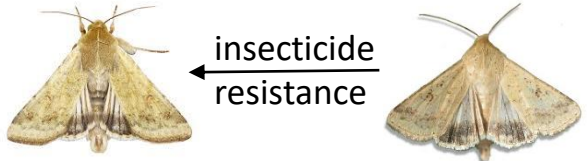
- Identifying genomic regions involved in speciation
- Identifying genes underlying traits

Are the species declining?

- Detecting past inbreeding
- Assessing genetic diversity

Studying gene flow

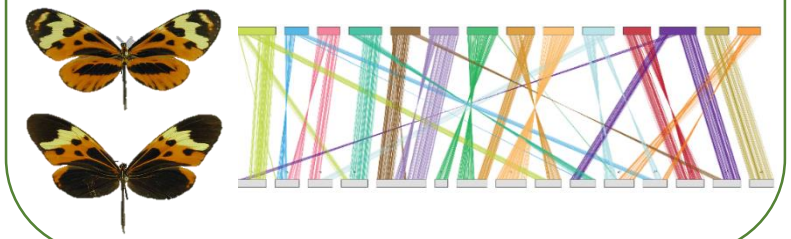
- Are populations/species hybridising or have in the past?
- Finding regions of adaptive introgression



Helicoverpa zea *H. armigera*

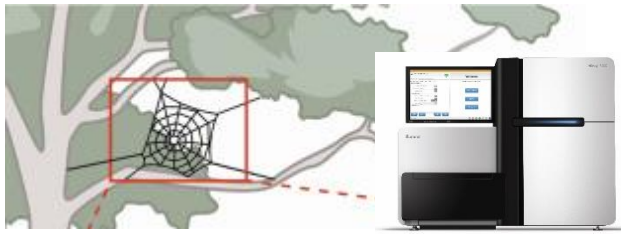
Studying genome evolution

- Gene expansions, e.g. olfactory
- Chromosomal rearrangements
- Genome size evolution (TEs, etc)

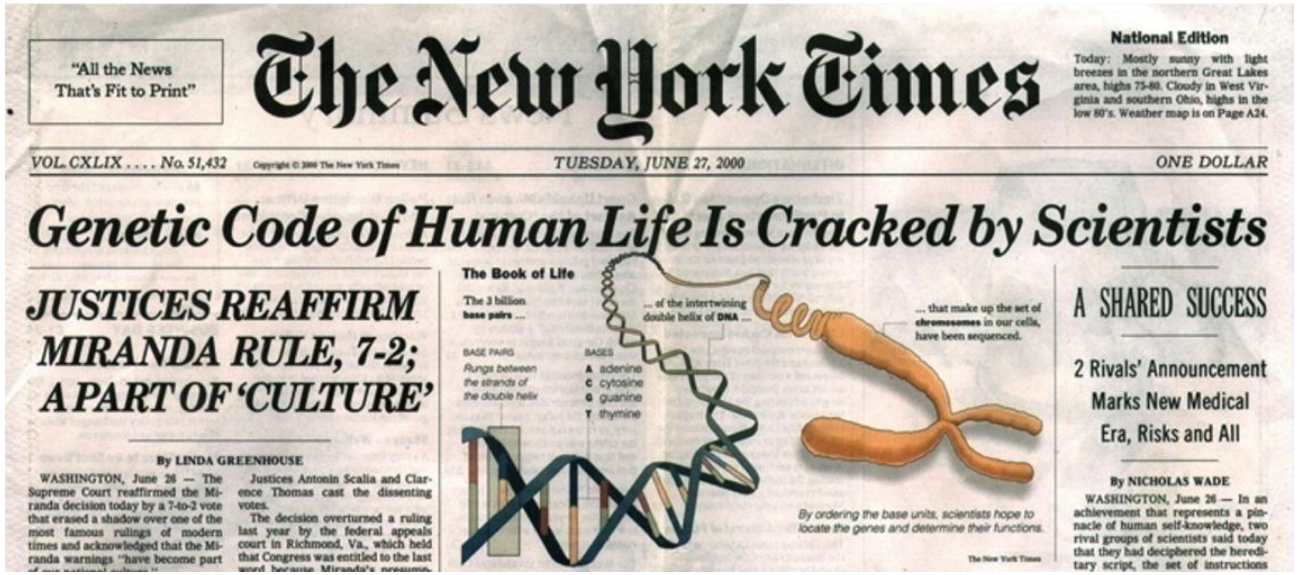


Which species occur here?

- Identifying biodiversity hotspots
- Monitoring effectiveness of conservation strategies



Human Genome Project



"All the News That's Fit to Print"

The New York Times

National Edition

Today: Mostly sunny with light breezes in the northern Great Lakes area, highs 75-80. Cloudy in West Virginia and southern Ohio, highs in the low 80's. Weather map is on Page A24.

VOL. CXLIX . . . . No. 51,432 Copyright © 2000 The New York Times TUESDAY, JUNE 27, 2000 ONE DOLLAR

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption

The Book of Life

The 3 billion base pairs ...

... of the intertwining double helix of DNA ...

... that make up the set of chromosomes in our cells, have been sequenced.

BASE PAIRS Rungs between the strands of the double helix

BASES A adenine C cytosine G guanine T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

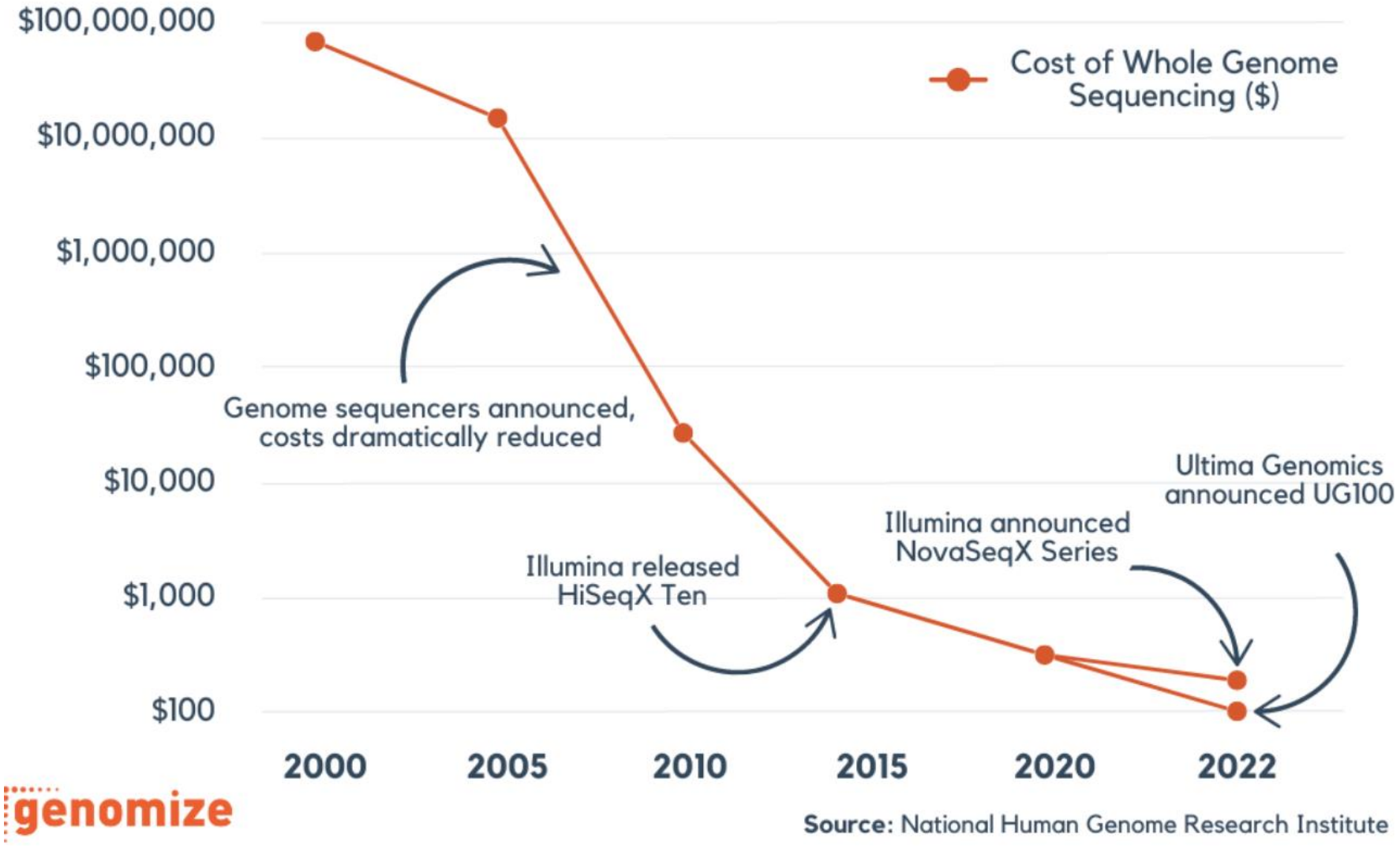
WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions

- Human genome project – started in 1990, completed in 2003
- Sequenced across ~20 institutions worldwide
- Cost an approximate $5 billion US dollars

The first human reference genome transformed modern medicine and understanding of human evolution and physiology

- Comparing populations e.g. to study how humans spread across the globe
- Finding introgression with neanderthals and denisovans
- Identifying genes under selection, like the laktase gene
- Finding genes causing diseases like breast cancer
- Understanding how cancer develops
- Personalised medicine
- Single-cell sequencing to understand which genes are active in which cells

Sequencing costs are decreasing rapidly



Since Oct 2023 PacBio Revio (66 Gbp per lane in 15 kb reads)

EARTH BIOGENOME PROJECT

Darwin TREE of LIFE

Project Psyche

African BioGenome Project

ATLASea
Atlas des génomes marins

VGP VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

BAT 1K

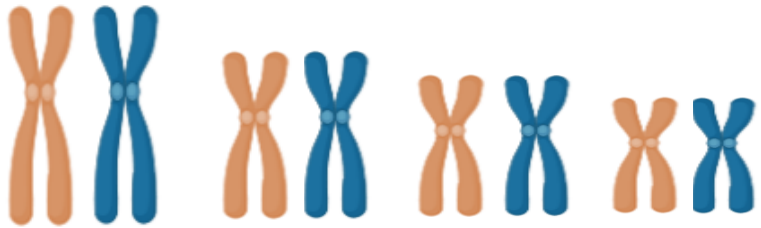
ERGA

CREATING A NEW FOUNDATION FOR BIOLOGY

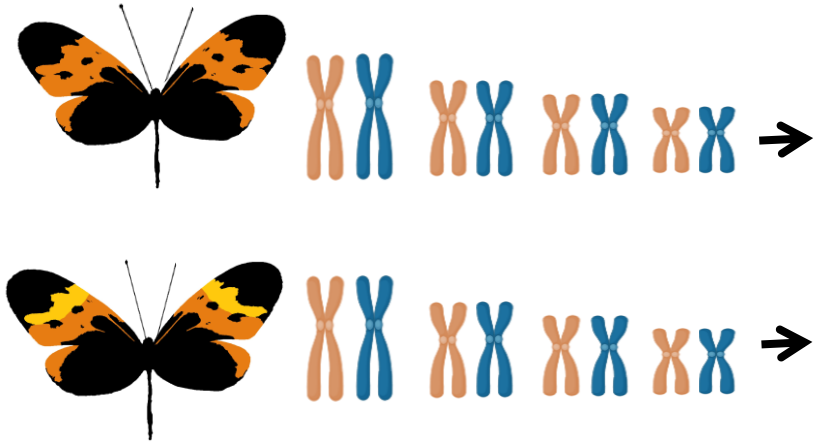
Sequencing Life for the Future of Life

Why do we need a reference genome for whole-genome sequencing projects?

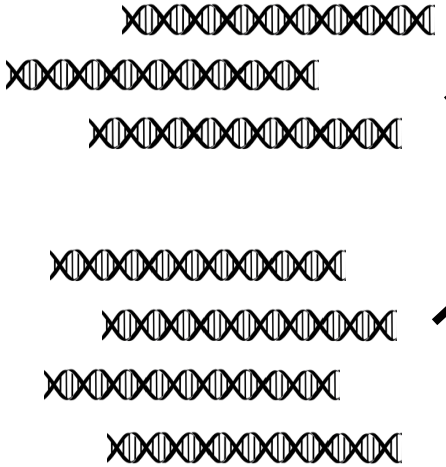
Genome = set of all chromosomes



Problem:
We do not know which of these sequences to compare



Break into many million short DNA fragments



Short-read sequencing machine (e.g. Illumina)

Millions of reads, 150 bp long

GATGCT
ATAGTG
GTGTGG

GTGTAG
GATGCT
CTGAGT
TGCTGA

How do we make a reference genome?

Genome = set of all chromosomes

Break into many million
long (10-50 kbp) DNA fragments

Long-read sequencing machine (e.g. PacBio)

more expensive than short-read sequencing

Millions of reads, 10-20 kbp long

GATGCTGAGTA
ATAGTGTGGAT
GTGTGGATGTG
TGCTGAGTTCG
TGGATGCTGAT
CTGAGTTCTCG

puzzling them together (aligning them to each other)

Reference genome

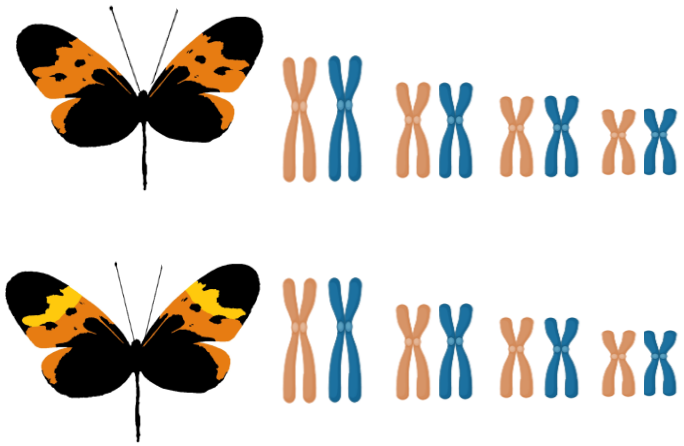
ATAGTGTGGATGCTGAGTTCGT

ATAGTGTGGATG
GTGTGGATGCT
TGGATGCTGAGT
GATGCTGAGTTC
TGCTGAGTTCG
CTGAGTTCGT

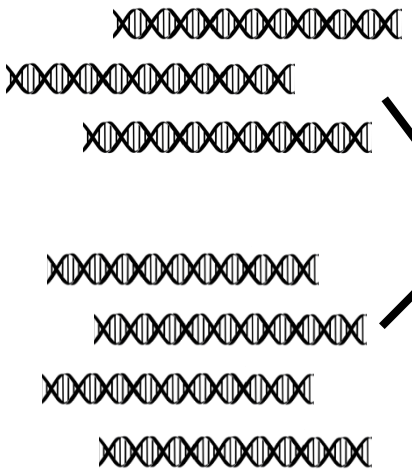
How do we <u>use</u> this reference genome?

Solution:
The reference genome allows us to place the reads so that we can compare them across individuals, populations or species

Genome = set of all chromosomes



Break into many million short DNA fragments



Short-read sequencing machine (e.g. Illumina)

GATGCT
ATAGTG
GTGTGG

GTGTAG
GATGCT
CTGAGT
TGCTGA

Reference genome

ATAGTGTGGATGCTGAGTTCGT

GATGCT
ATAGTG
GTGTGG

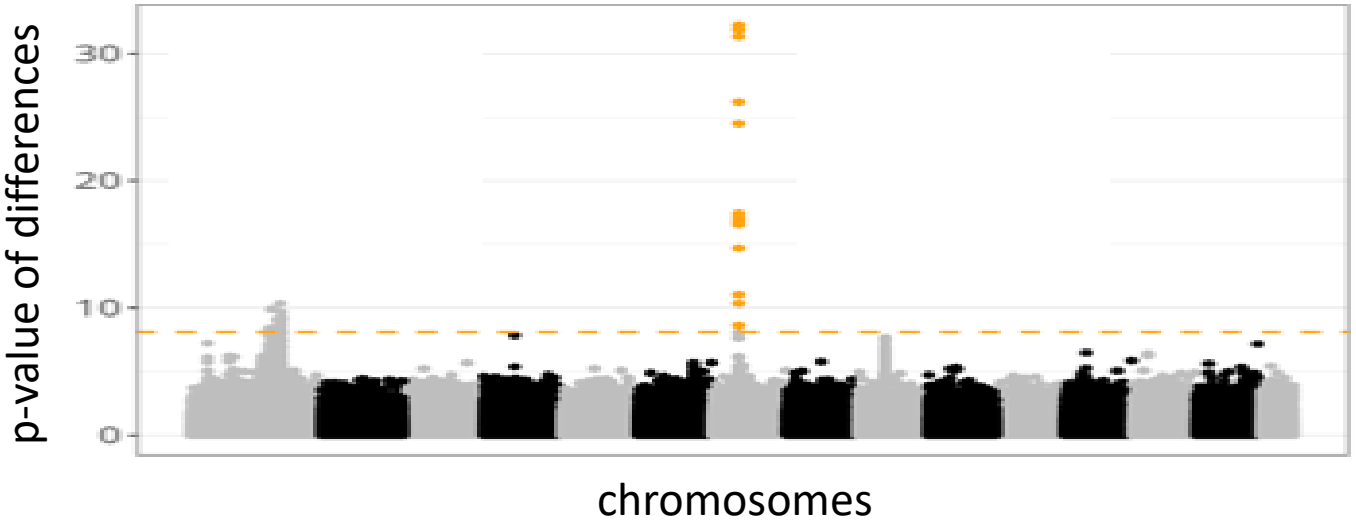
GTGTAG
GATGCT
CTGAGT
TGCTGA

Do these two butterflies belong to different species?

Eva van der Heijden

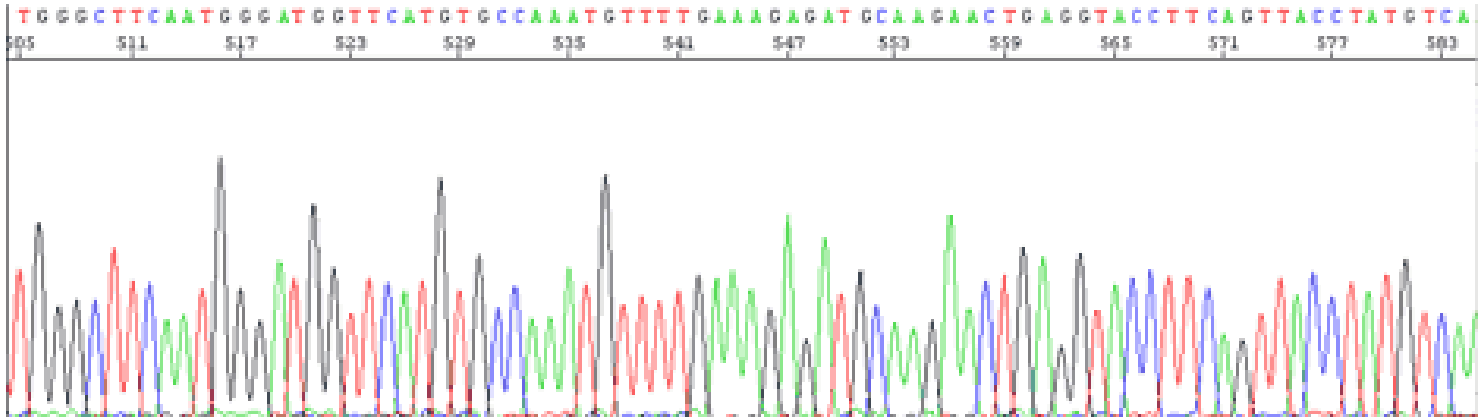
They are only significantly different in one region in the genome, right next to *cortex/ivory,* which are also known to affect colour patterns in other butterflies and moths

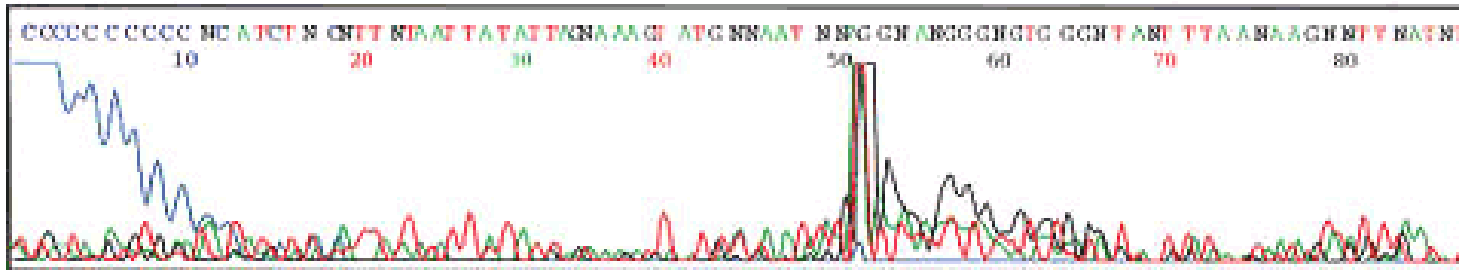
CRISPR butterflies with *cortex/ivory* knocked-out



Introduction to high-throughput or next-generation sequencing

Sanger Sequencing (since 1980s)





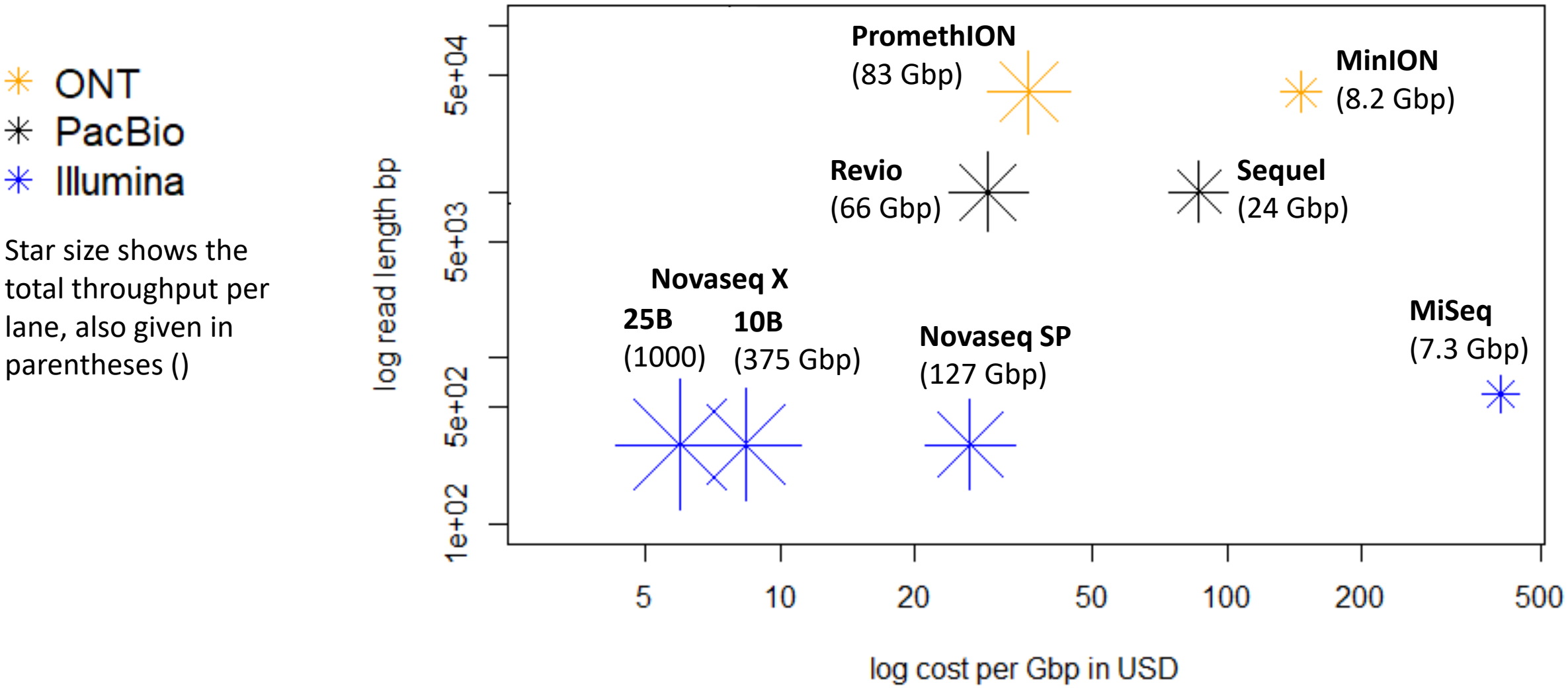
- Possible to manually check each sequence and resequence failed sequences
- Requires primer sequences and has very low throughput (expensive per bp)

Two main types of high-throughput sequencing

- **Short-read sequencing**
  - Reads are typically 150 bp long
  - Cheaper than long-read sequencing
  - E.g. Illumina, soon probably also Ultima Genomics

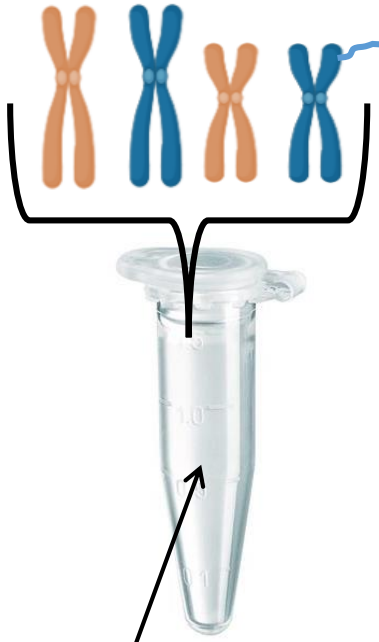
- **Long-read sequencing**
  - Reads are typically >10 kb long (PacBio: 15-20 kbp, Nanopore: 10-100 kbp)
  - More expensive than short-read technologies
  - Required for making a reference genome
  - E.g. PacBio or Nanopore

Read length versus per Gbp sequencing costs for different sequencing machines (note the axes are in logarithmic scale)



Whole-genome sequencing

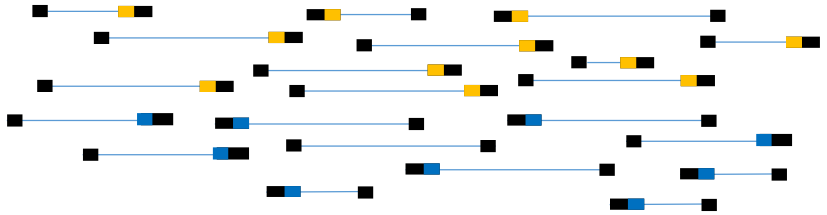
Genome
= complete set of chromosomes



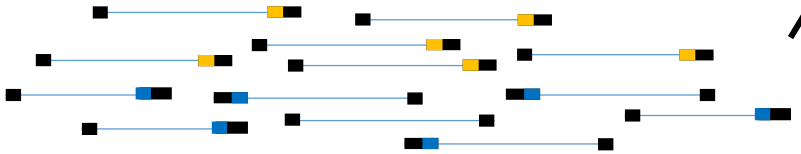
DNA (chromosomes)

Breaking chromosomes into shorter pieces for sequencing

Adding **sequencing adapters** incl. individual index



Size selection

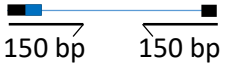


Long-read sequencing (PacBio or Nanopore/ONT)
paired-end sequencing

10-50 kbp

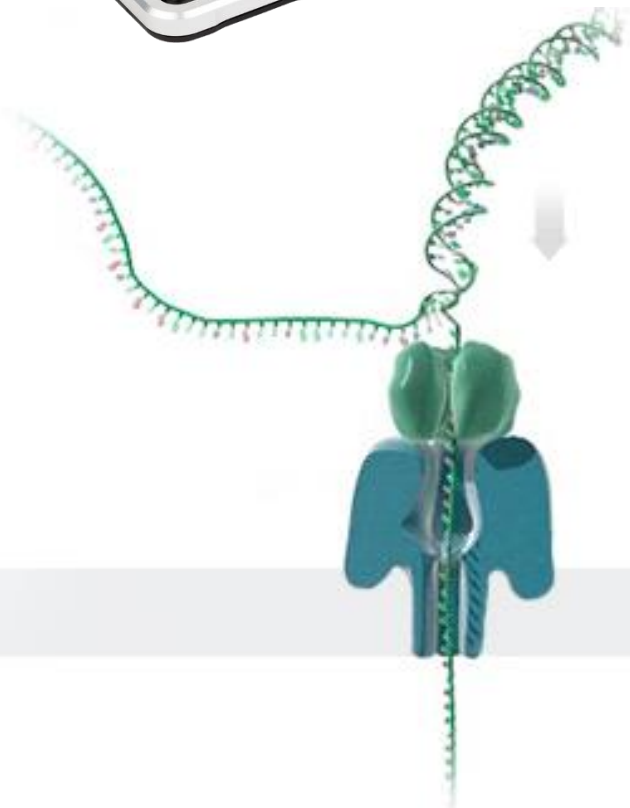
Short-read sequencing (e.g. Illumina)
paired-end sequencing

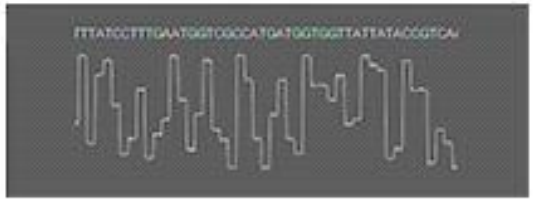
350-500 bp



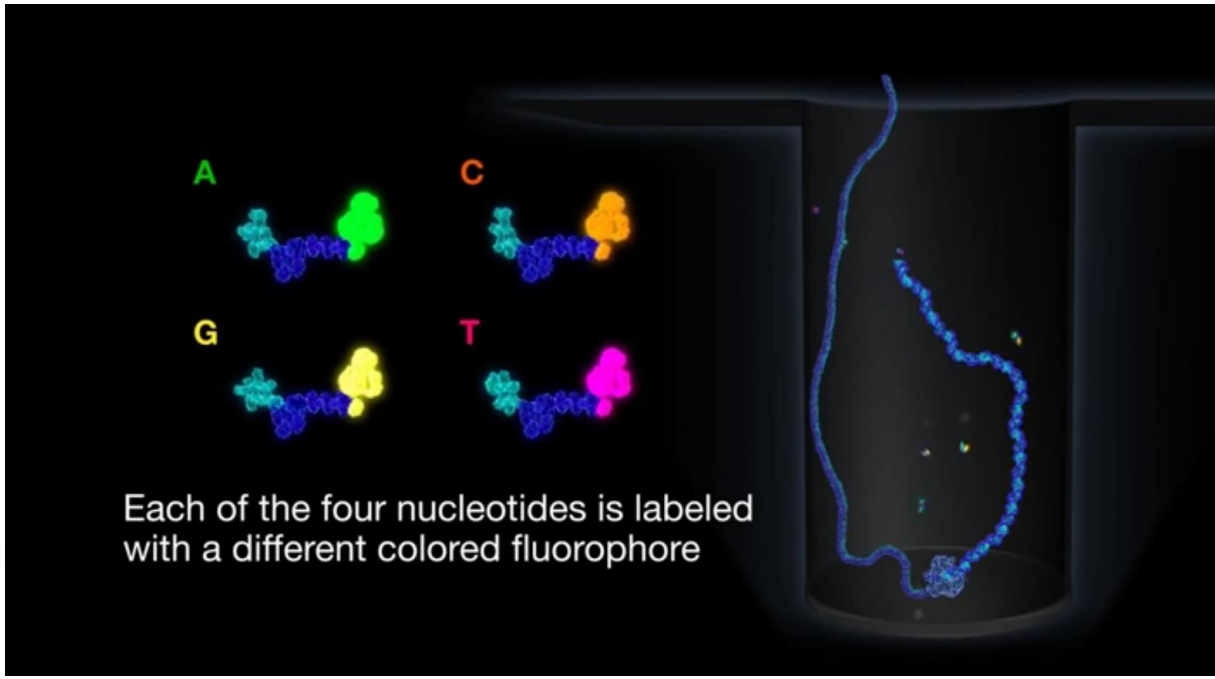
Long read sequencing technologies

Nanopore



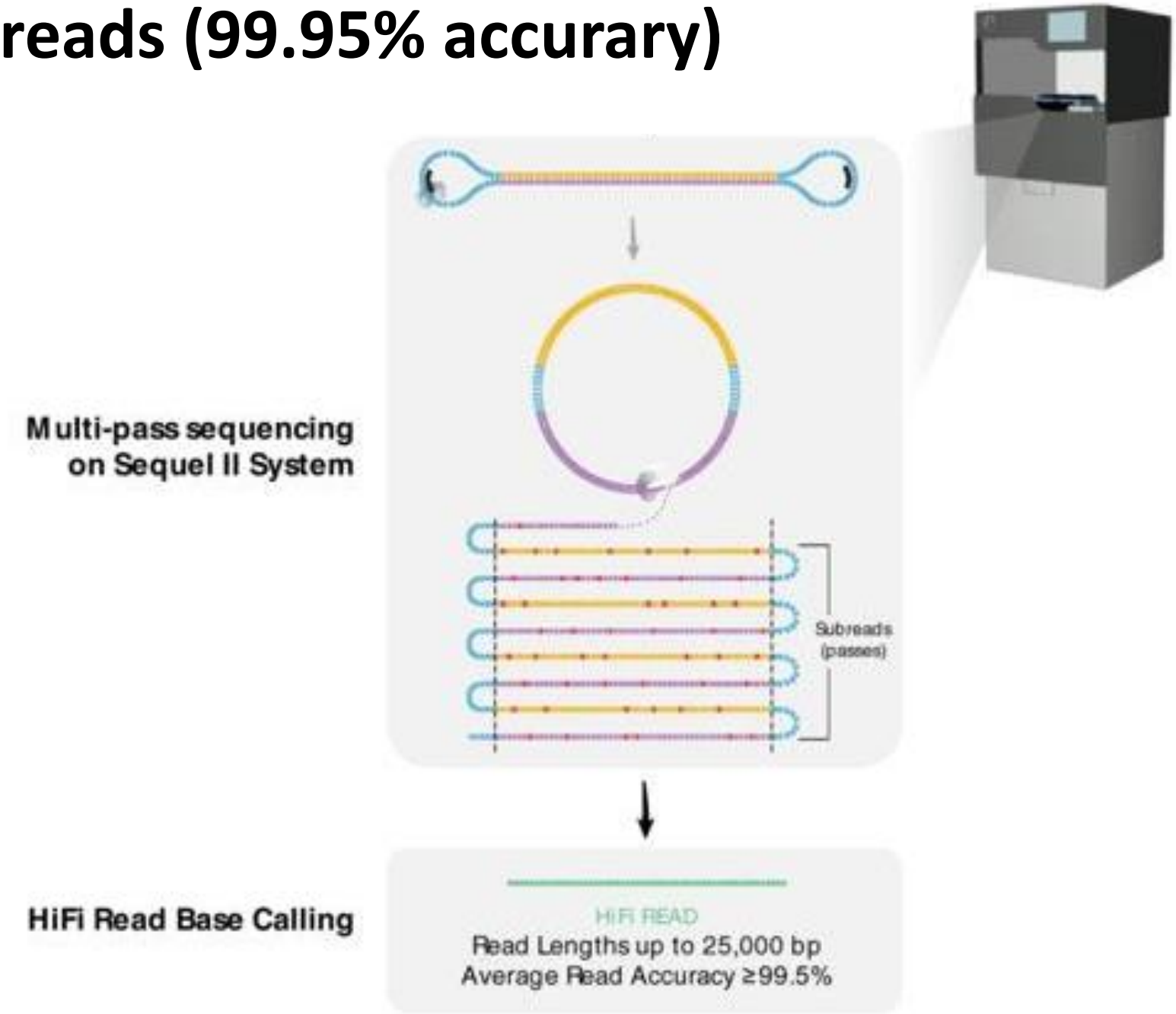


PacBio



PacBio HiFi reads (99.95% accurary)

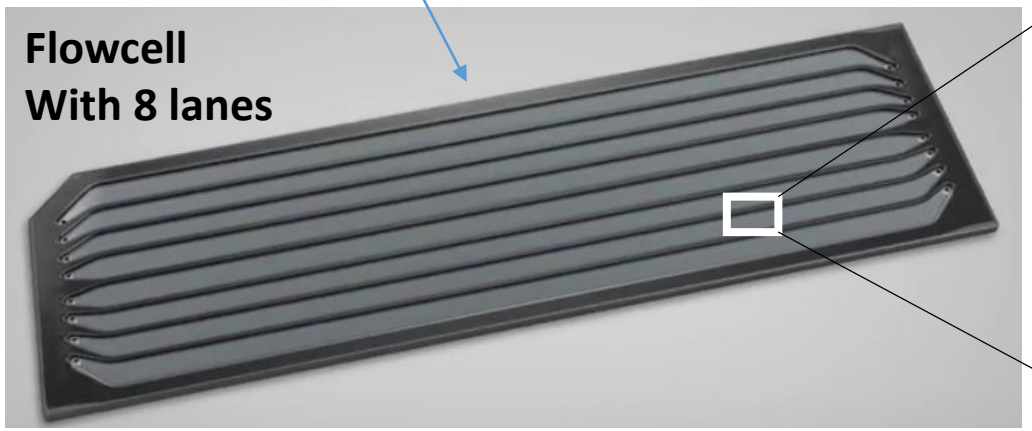
Each DNA-fragment is sequenced many times to get a high-quality consensus (=summary) read

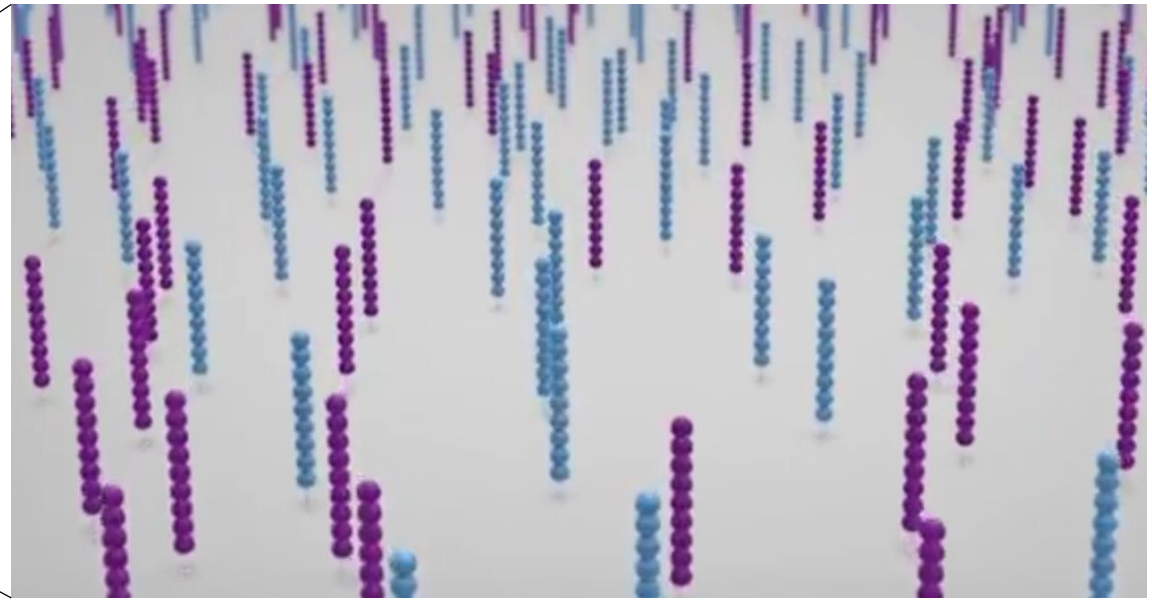


Illumina flowcell

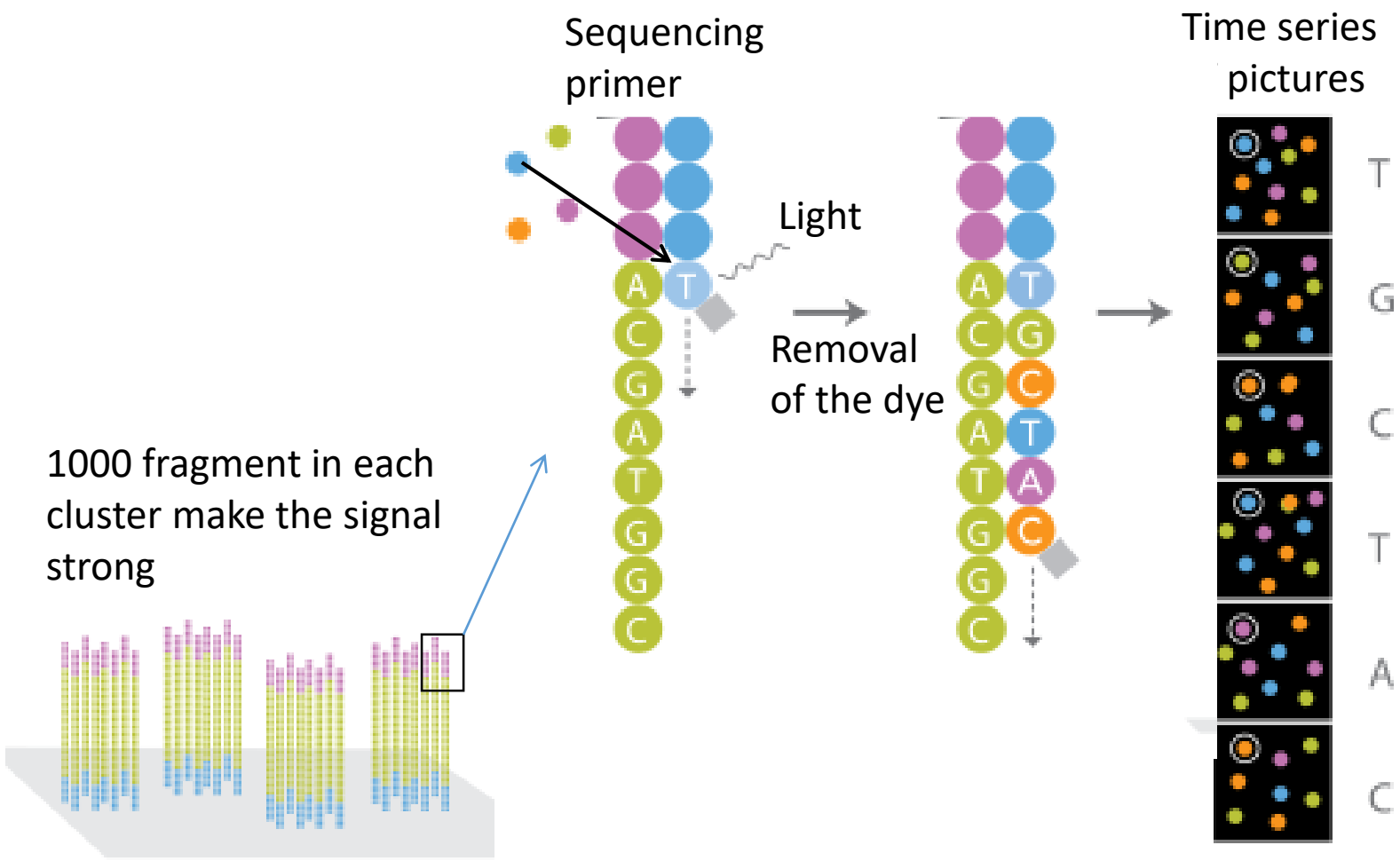
DNA fragments with Illumina adapters

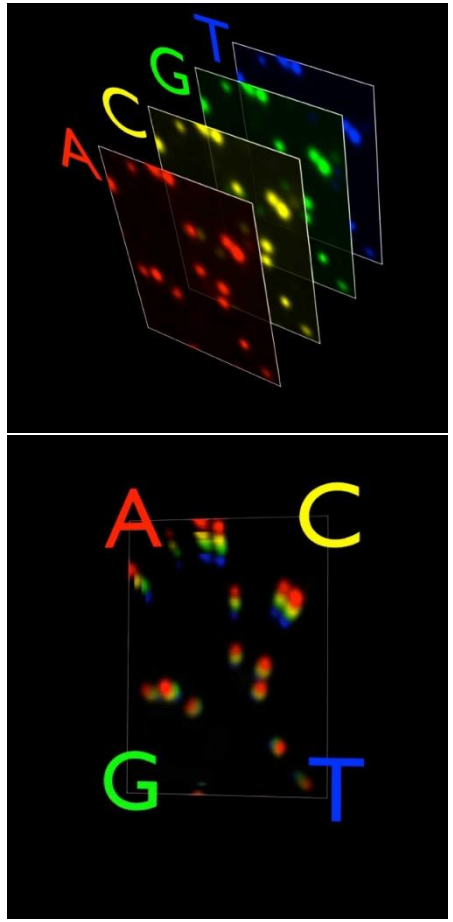
Each lane contains a dense lawn of Illumina primers



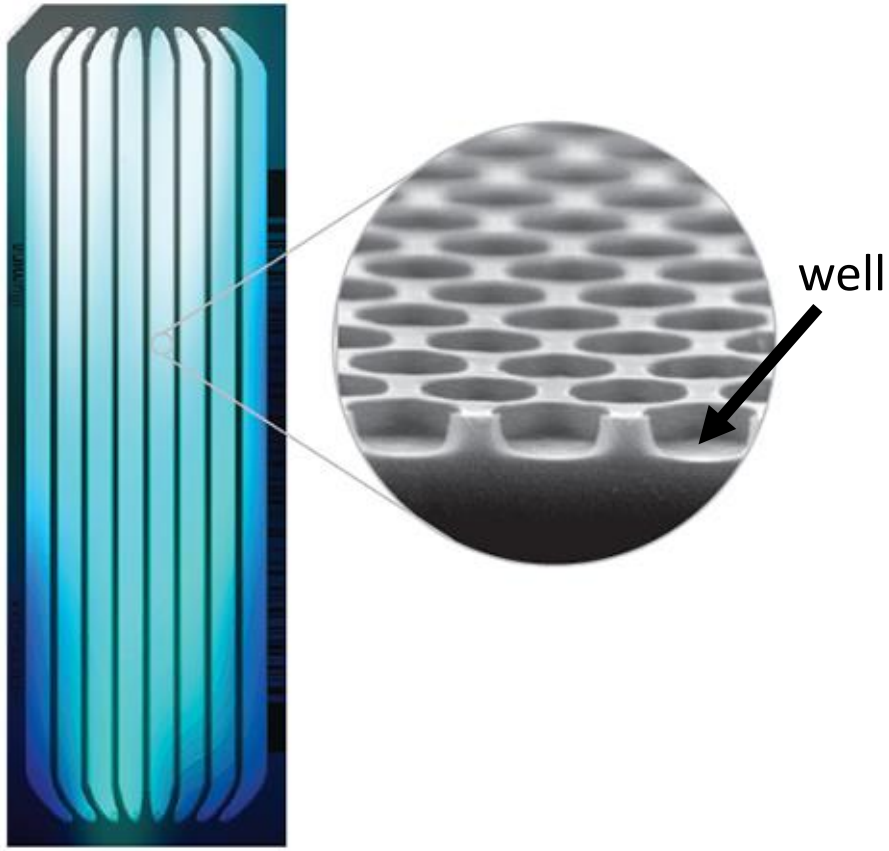


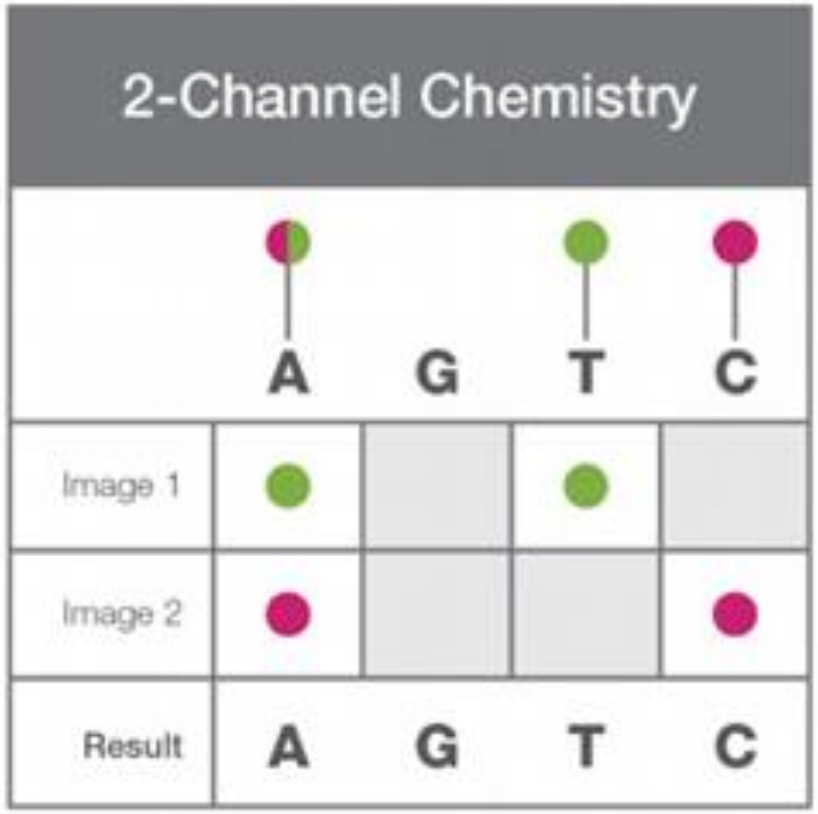
Short-read sequencing with Illumina





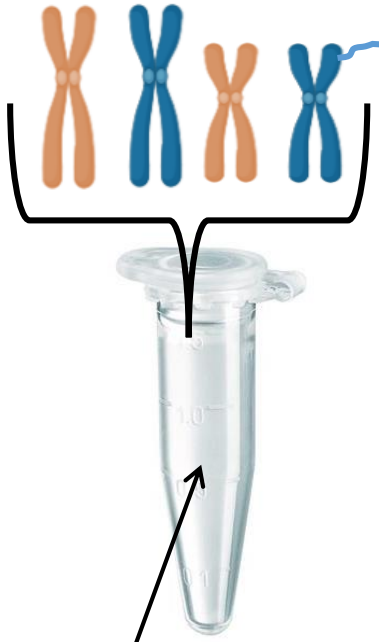
Newer Illumina machines use wells and only 2 colours
(e.g. Novaseq, Nextseq, MiniSeq. This makes it faster and cheaper)





Whole-genome sequencing

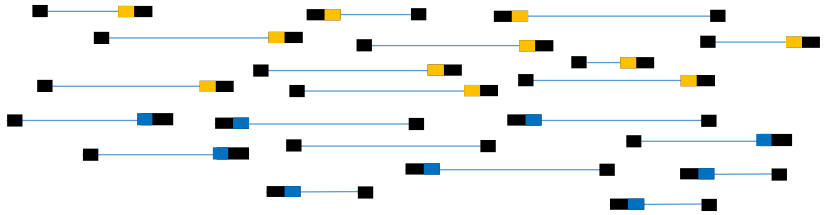
Genome
= complete set of chromosomes



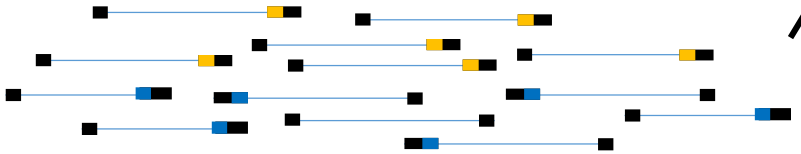
DNA (chromosomes)

Breaking chromosomes into shorter pieces for sequencing

Adding **sequencing adapters** incl. individual index



Size selection

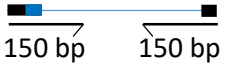


Long-read sequencing (PacBio or Nanopore/ONT)
paired-end sequencing

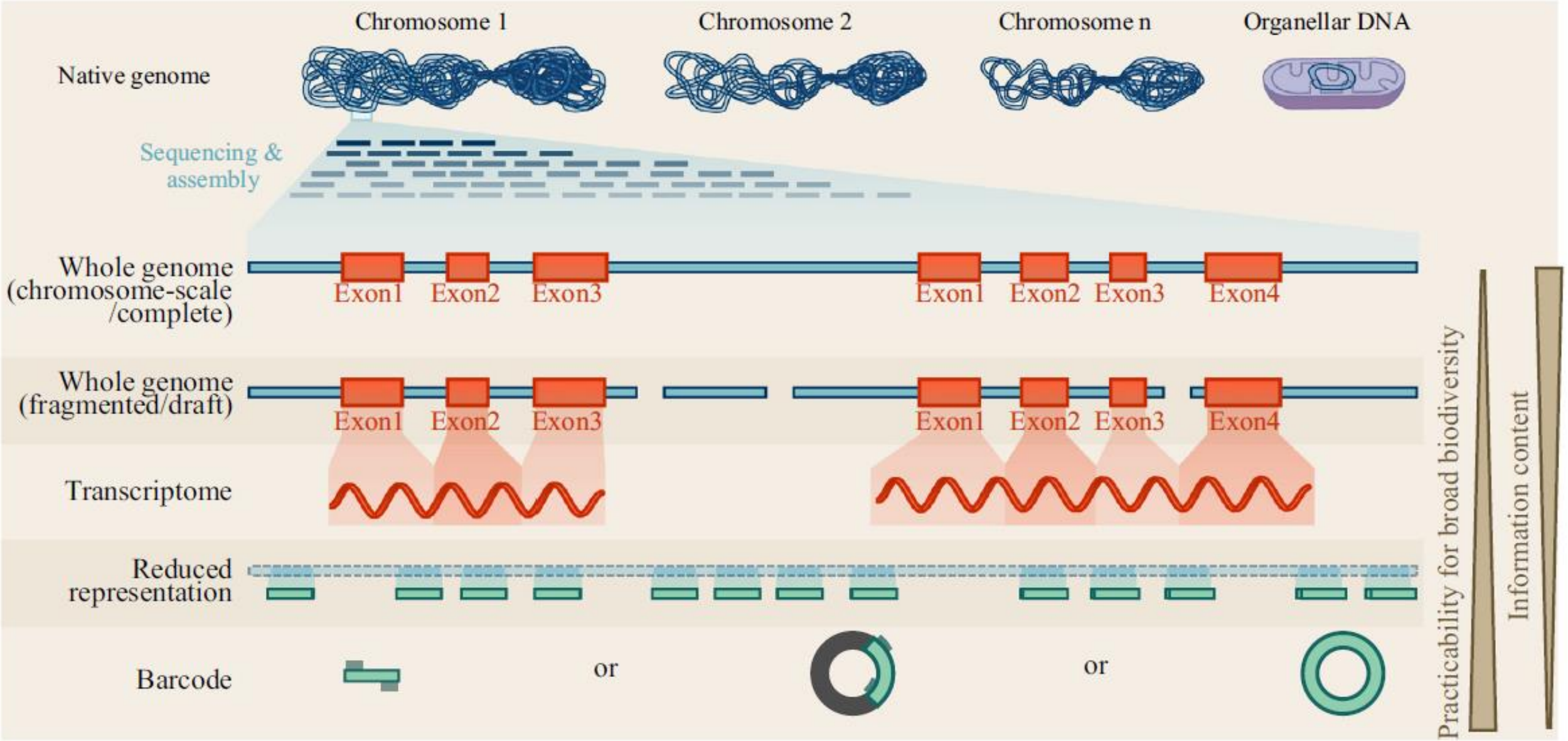
10-50 kbp

Short-read sequencing (e.g. Illumina)
paired-end sequencing

350-500 bp



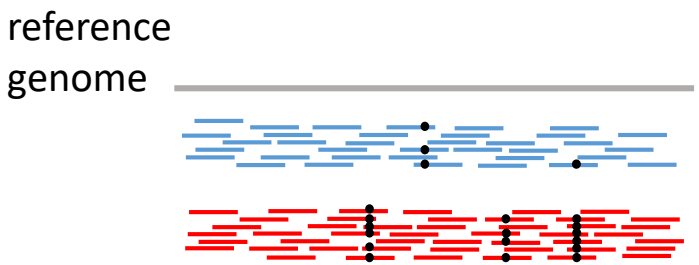
Sequencing approaches for biodiversity genomics



Sequencing approaches for biodiversity genomics

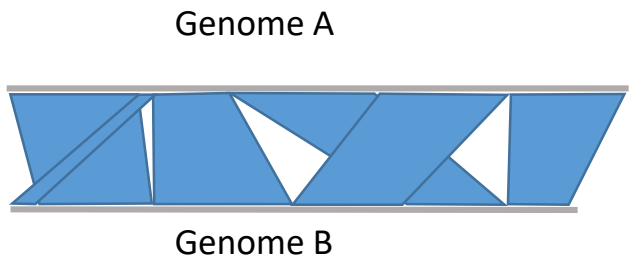
Whole-genome requencing (short-read data)

- Requires a reference genome
- individuals need to be from the same or closely related species
- Complete genome sequenced



Genome assembly comparisons (long-read data)

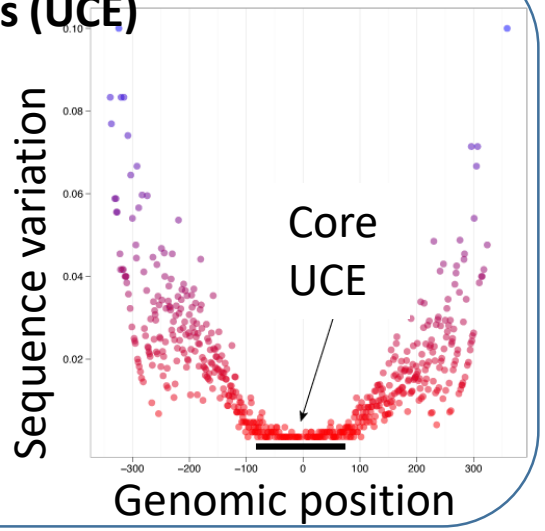
- Comparative genomics – studying structural variation between species, can be distantly related
- Gene expansions, transposable elements etc
- Phylogenomics across deeply divergent species
- Pangenomics – multi-genome assemblies to study within-species variation in structural variants



Reduced-representation techniques (only parts of the genome sequenced)

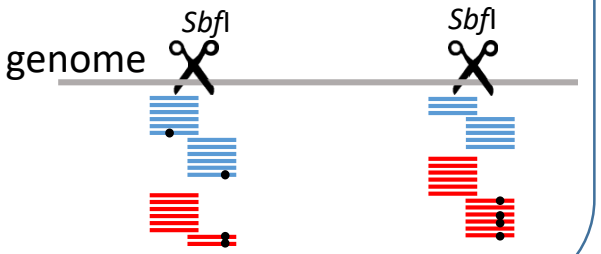
Ultra-conserved elements (UCE)

- Sequence capture with baits based on genomic regions that are conserved across many species
- Works with highly divergent species



Restriction Associated Sequencing (RAD)
(similar methods: GBS, ddRAD)

- does not require primers/baits or reference genome
- individuals need to be from the same or closely related species
- Information from thousands of loci distributed across the genome

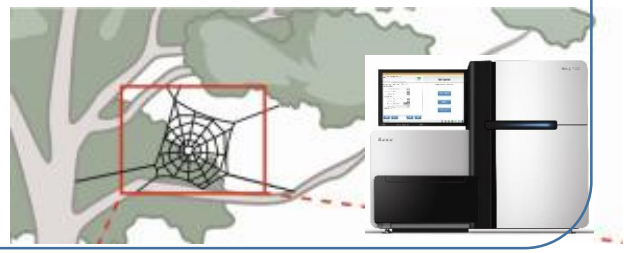


Targeted or amplicon sequencing, e.g. barcoding

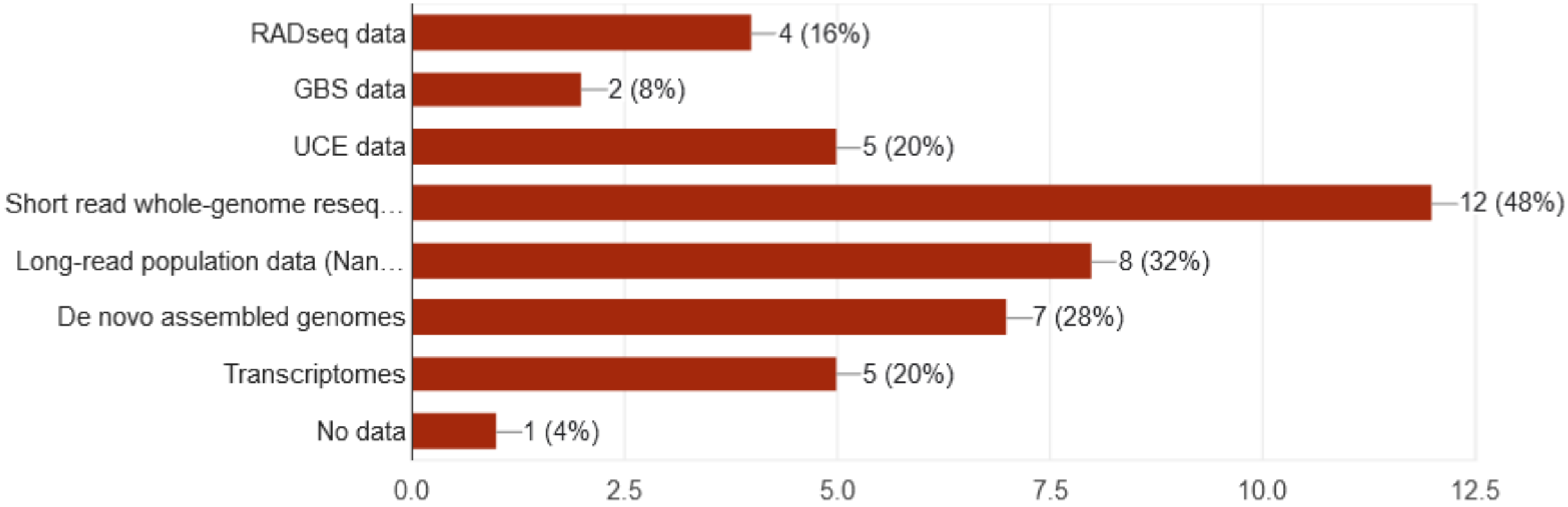
- Sequencing one or few genes
- requires primers
- e.g. CO1 (mitochondrial barcoding region), advantage: large database (BOLD) available to compare to for species identification

Environmental DNA (eDNA)

- Mostly CO1 sequencing from soil, water, air (spider webs)
- Identifying local species
- Studying species richness



Data that course attendees are working on



Trade-offs: Splitting reads (i.e. costs) among:

Total data gets divided by:

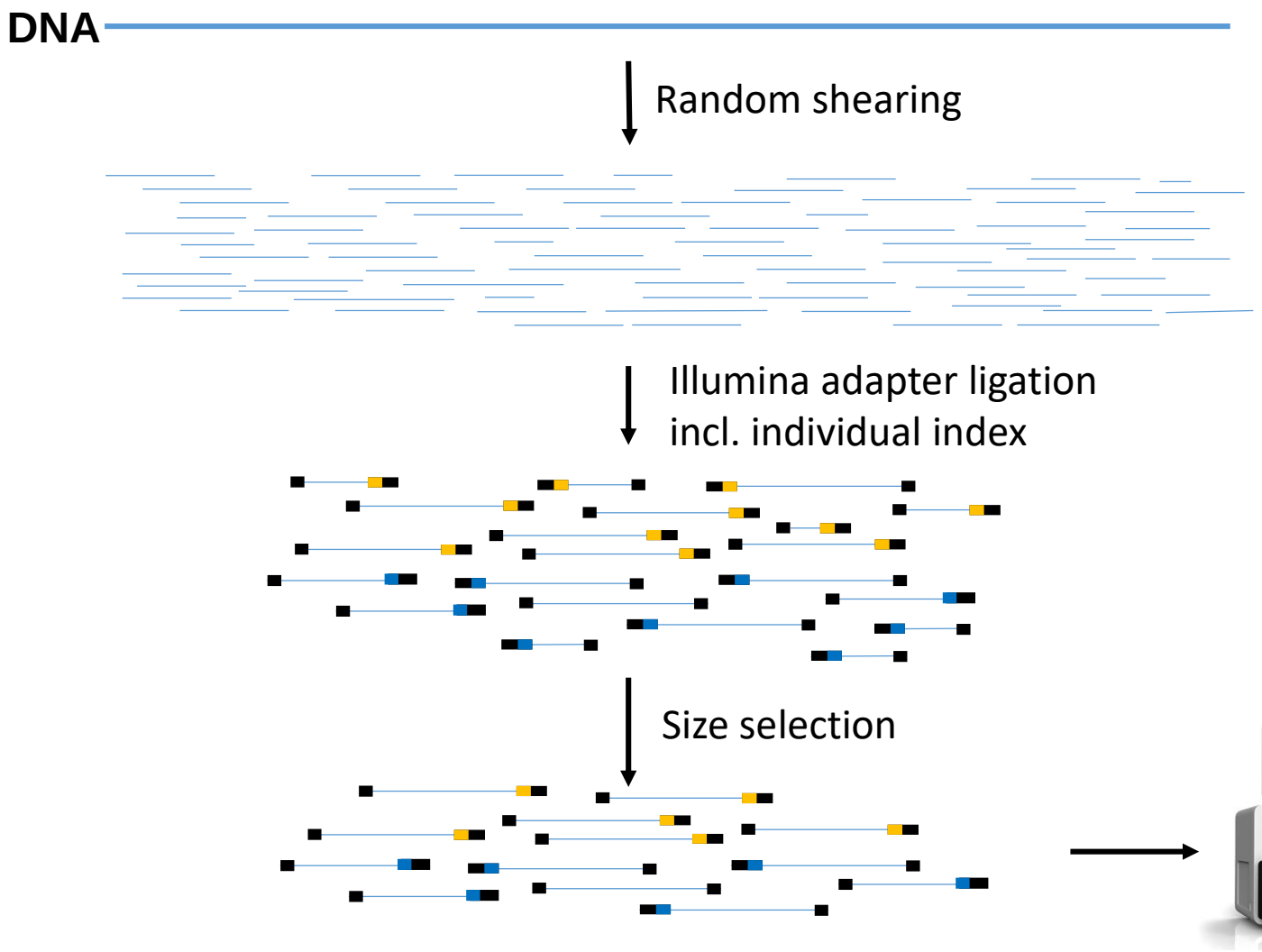
- Number of sites to sequence
  - Depends on genome size and sequencing stragegy, e.g. RAD versus whole-genome
- Sequencing depth (e.g. sequencing at 10x depth of coverage)
- Number of specimens to sequence

- Example: 1 Novaseq X 10 B lane
  ~2.5 billion paired-end reads of 150 bp each -> 375 Gbp data
  - 100 whole-genomes of a species with 0.375 Gbp genome size at 10x coverage
  - 19 whole-genomes of a species with 1 Gbp genome size at 20x coverage
  - 375 individuals sequenced with a RAD sequencing approach resulting in 50 Mbp at a sequencing depth of 20x

Now let's get started with handling genomic data!

First, a brief overview of what we will do now

Whole-genome short-read sequencing



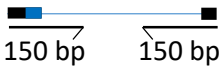
paired-end sequencing

Up to 20 billion read pairs

1. Quality check and trimming raw reads

paired-end Illumina sequencing

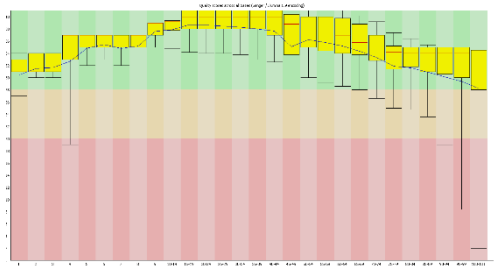
Up to 20 billion read pairs



Forward reads

Reverse reads

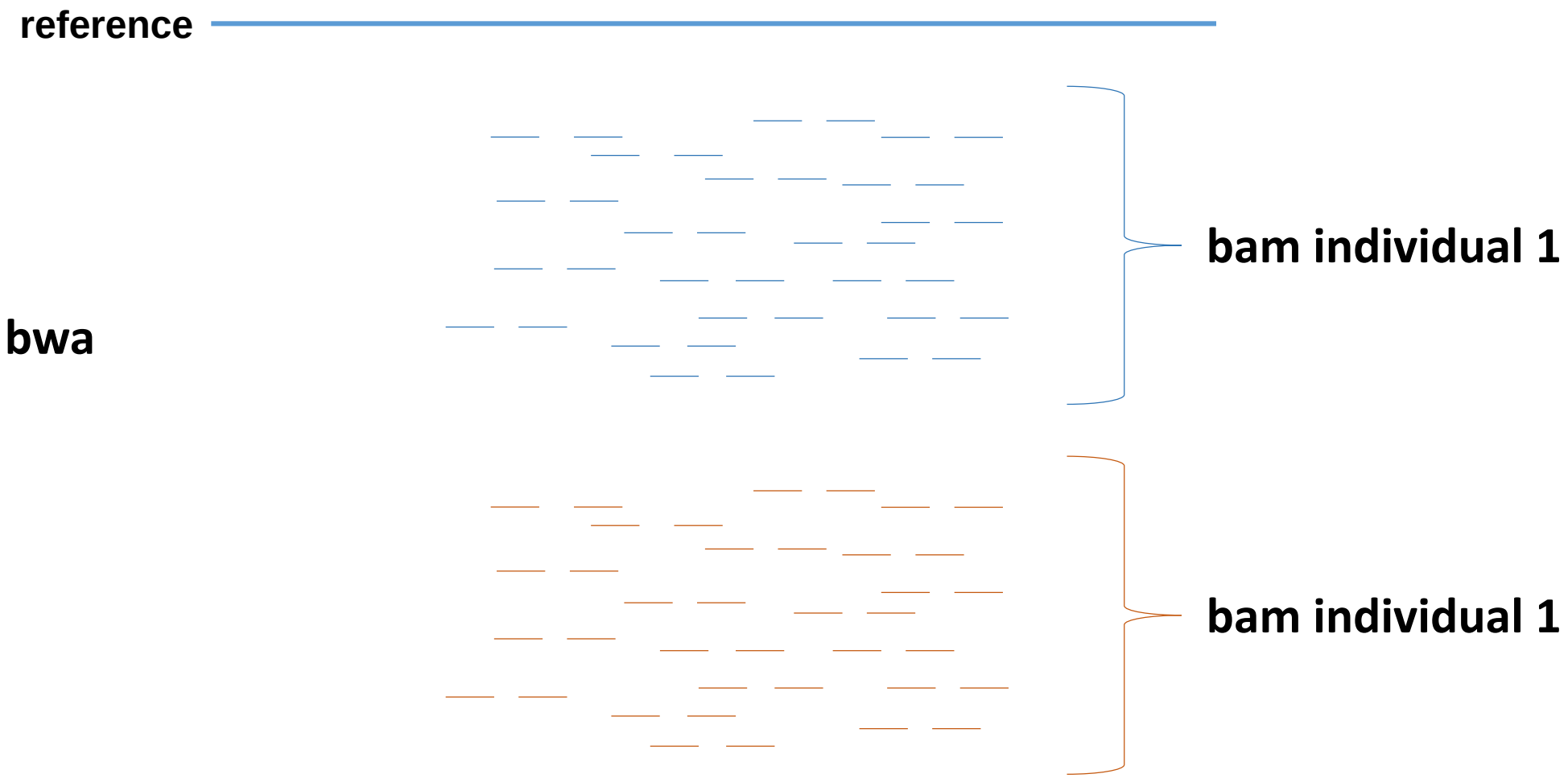
Fastqc quality plots



fastp

Trimmed and filtered reads of good quality

2. Alignment to the reference genome with bwa



3. Variant and genotype calling with bcftools