



Next-Generation Sequencing (NGS)

Rob Colautti

Biological Sciences

Contact:

BioSci 4325a

613-533-2353

bit.ly/colautti

robert.colautti@queensu.ca



@ColauttiLab

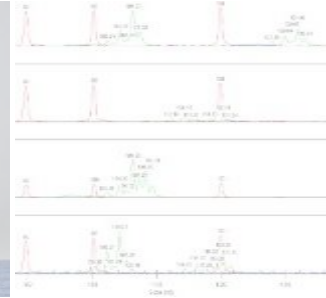
University of Windsor & GLIER – Hs Bsc, MSc



Hugh MacIsaac



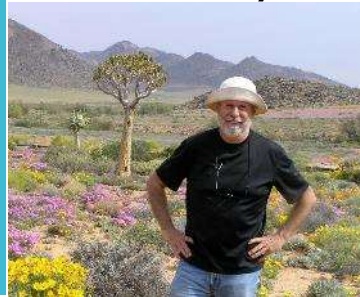
Dan Heath



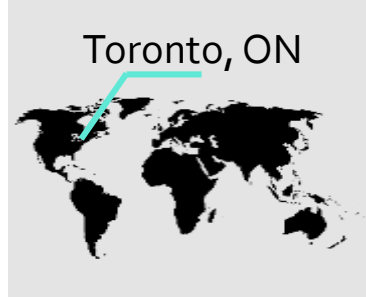
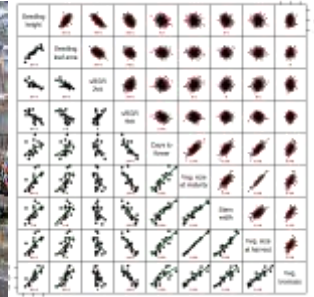
Windsor, ON

Research Background

University Toronto – PhD



Spencer Barrett



Toronto, ON

Research Background

Duke University – Postdoc



Tom Mitchell-Olds

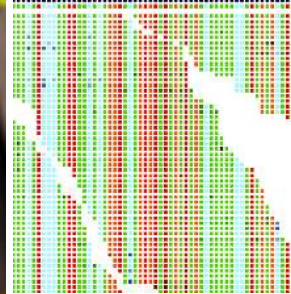
Jill Anderson

University of British Columbia – Postdoc



Loren Rieseberg

University of Tuebingen – Postdoc



Oliver Bossdorf

Durham, NC



Vancouver, BC



Tuebingen, DE



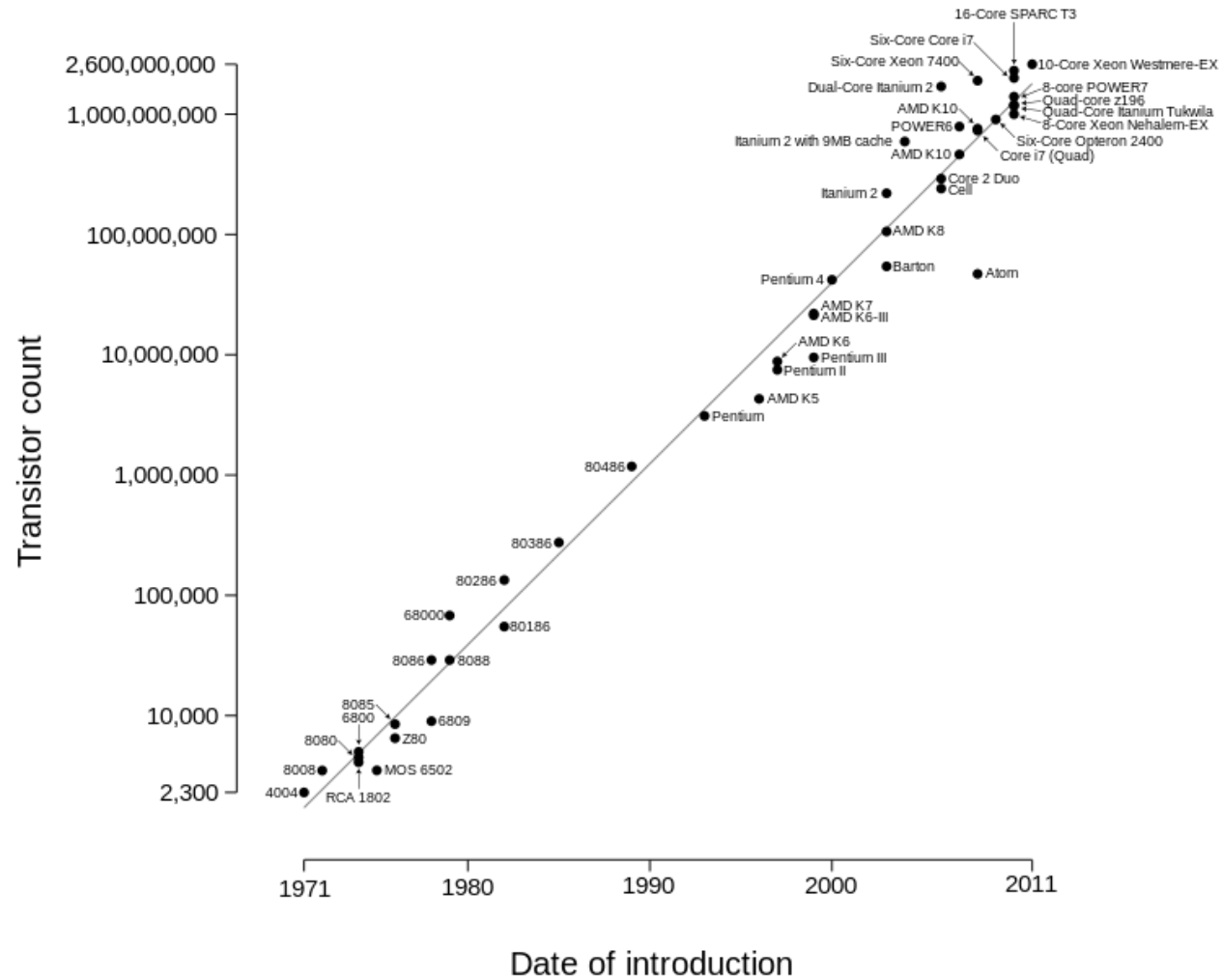
Overview

- What is Next-Generation Sequencing (NGS)?
- How does NGS work?
- Why is NGS revolutionizing the biological sciences?



Introduction to NGS

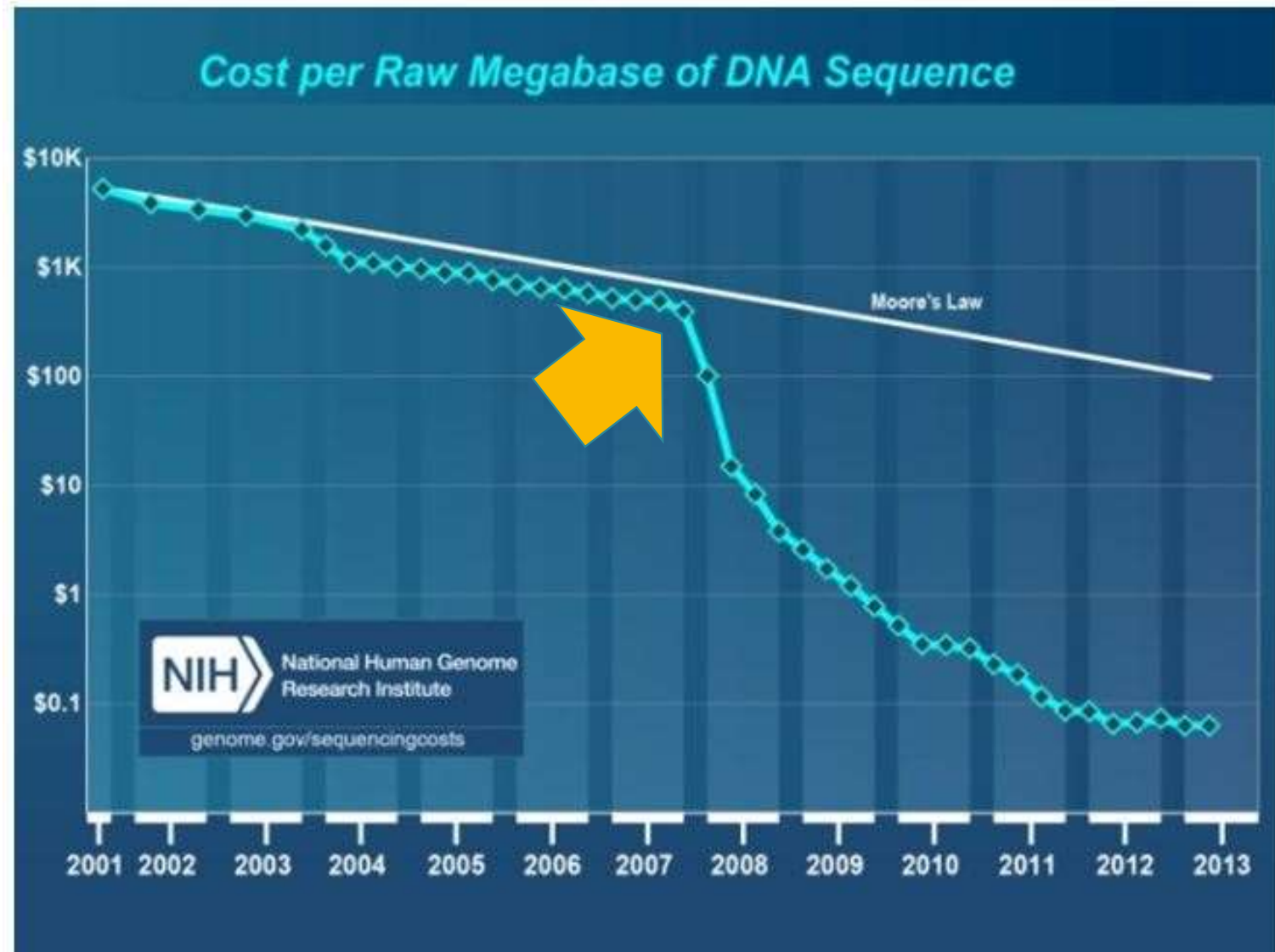
Moore's Law



- "the number of transistors in a dense integrated circuit doubles approximately every two years"

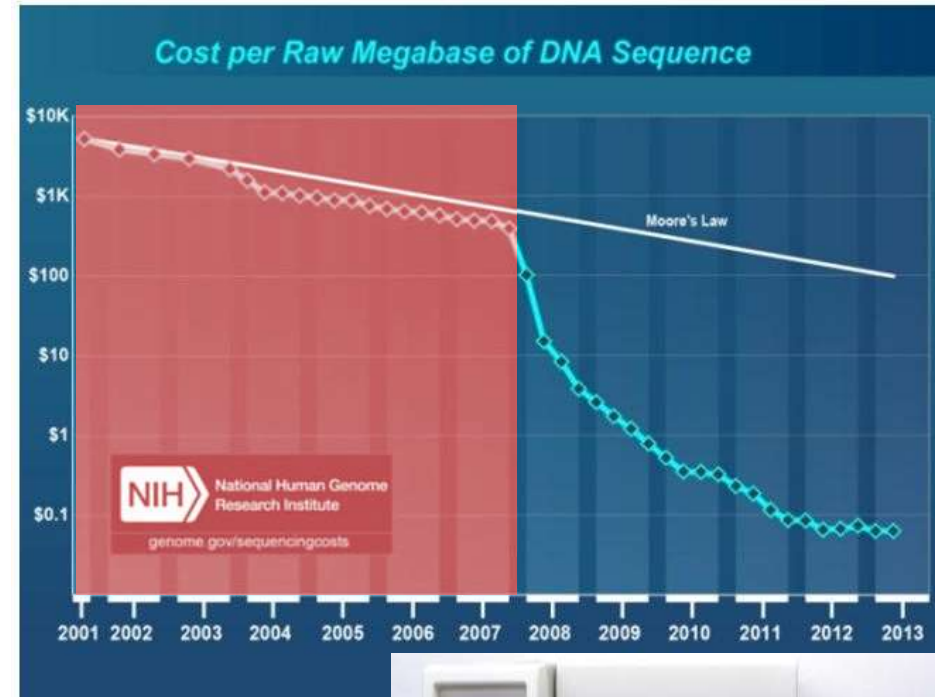
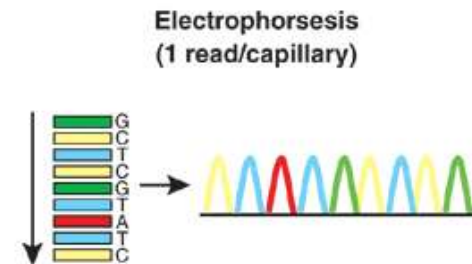
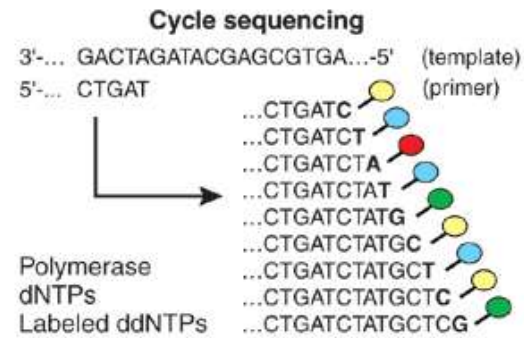
-- Wikipedia (Feb 2, 2016)

Moore's Law vs DNA Sequencing technology



Sanger Sequencing

Dideoxy chain termination sequencing

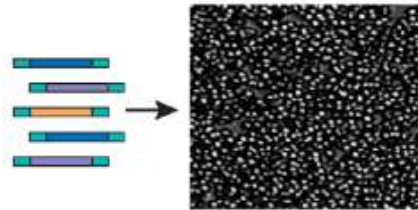


Beckman Coulter CEQ 8000

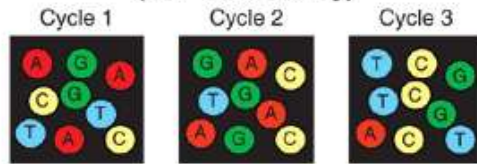
Next-generation sequencing (NGS)

Sequencing by synthesis

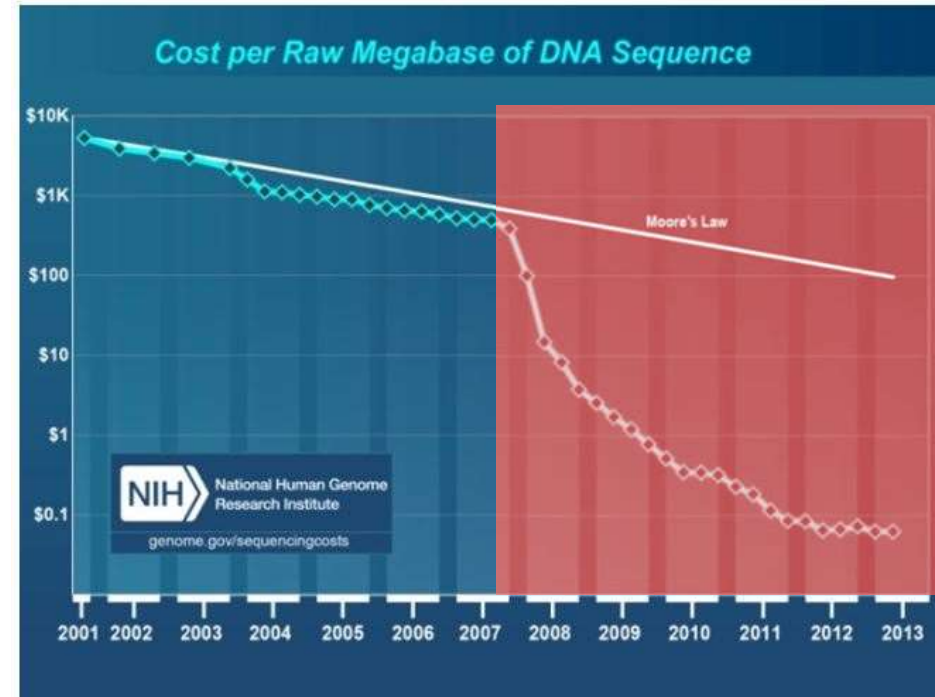
Generation of polony array



Cyclic array sequencing
($>10^6$ reads/array)

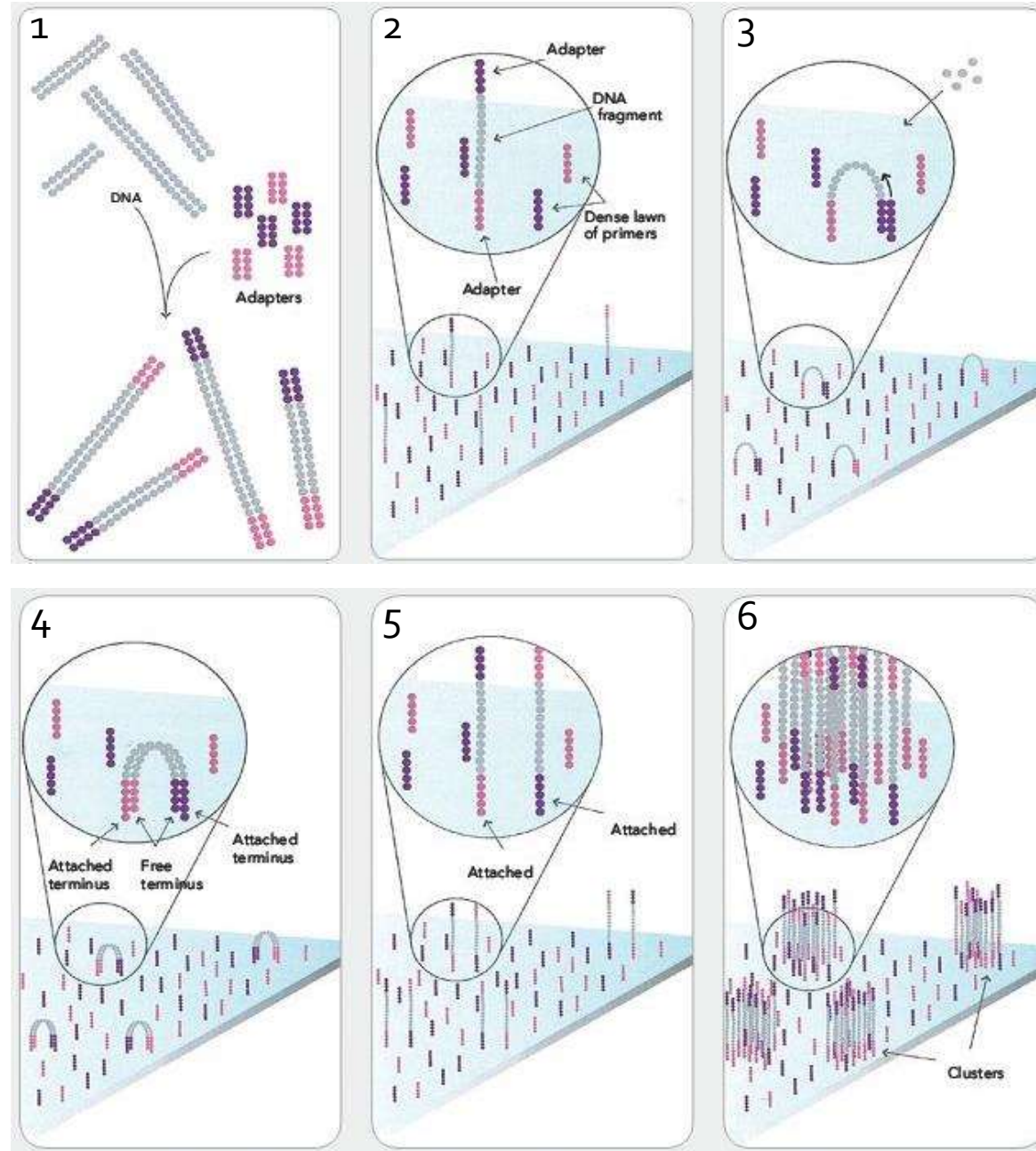


What is base 1? What is base 2? What is base 3?



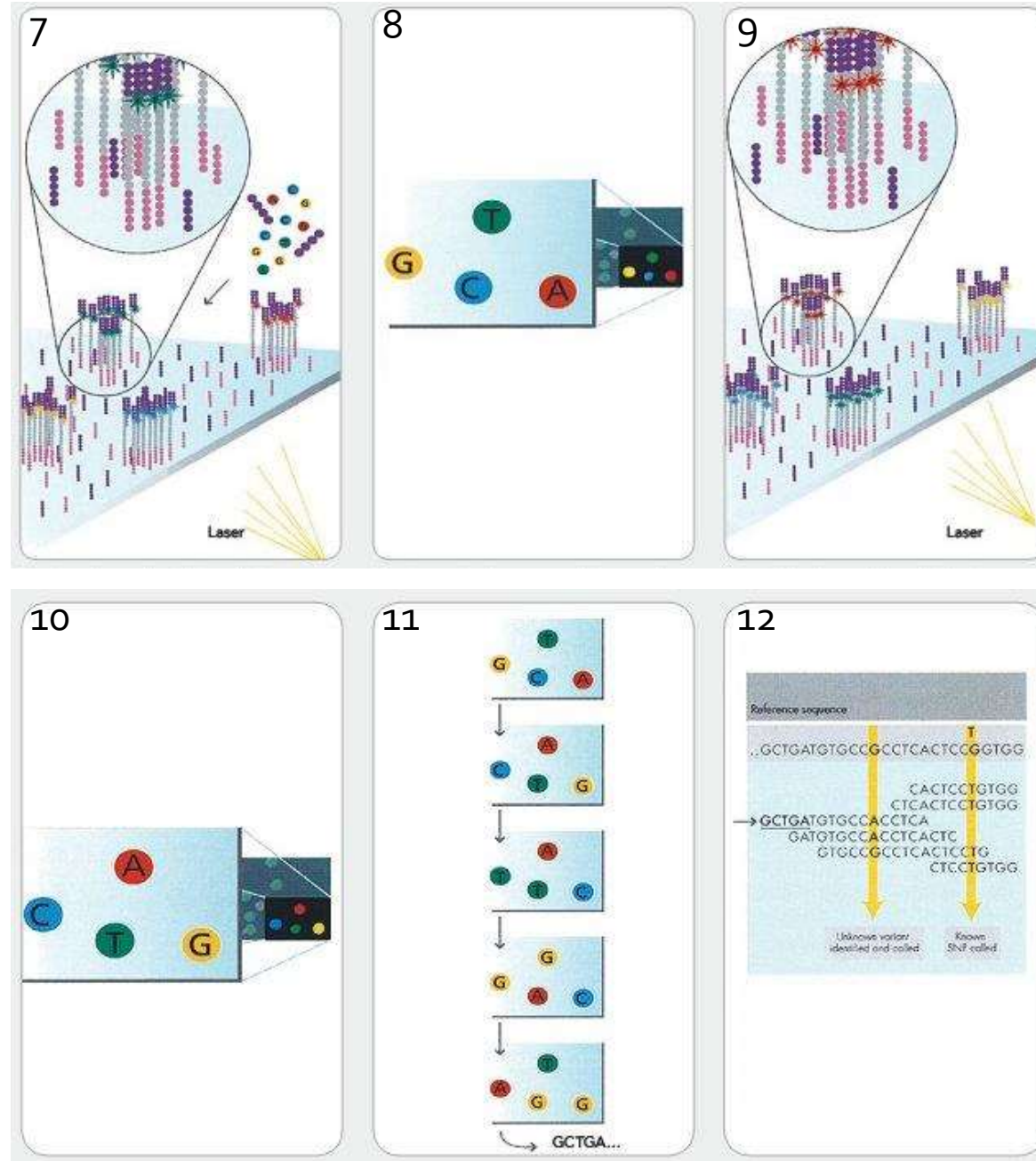
Illumina HiSeq & MiSeq

Illumina (Sequencing by synthesis)



1. Shear DNA into short fragments (~100-500bp) & ligate sequencing primers (adapters)
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double stranded molecules
6. Cluster amplification

Illumina

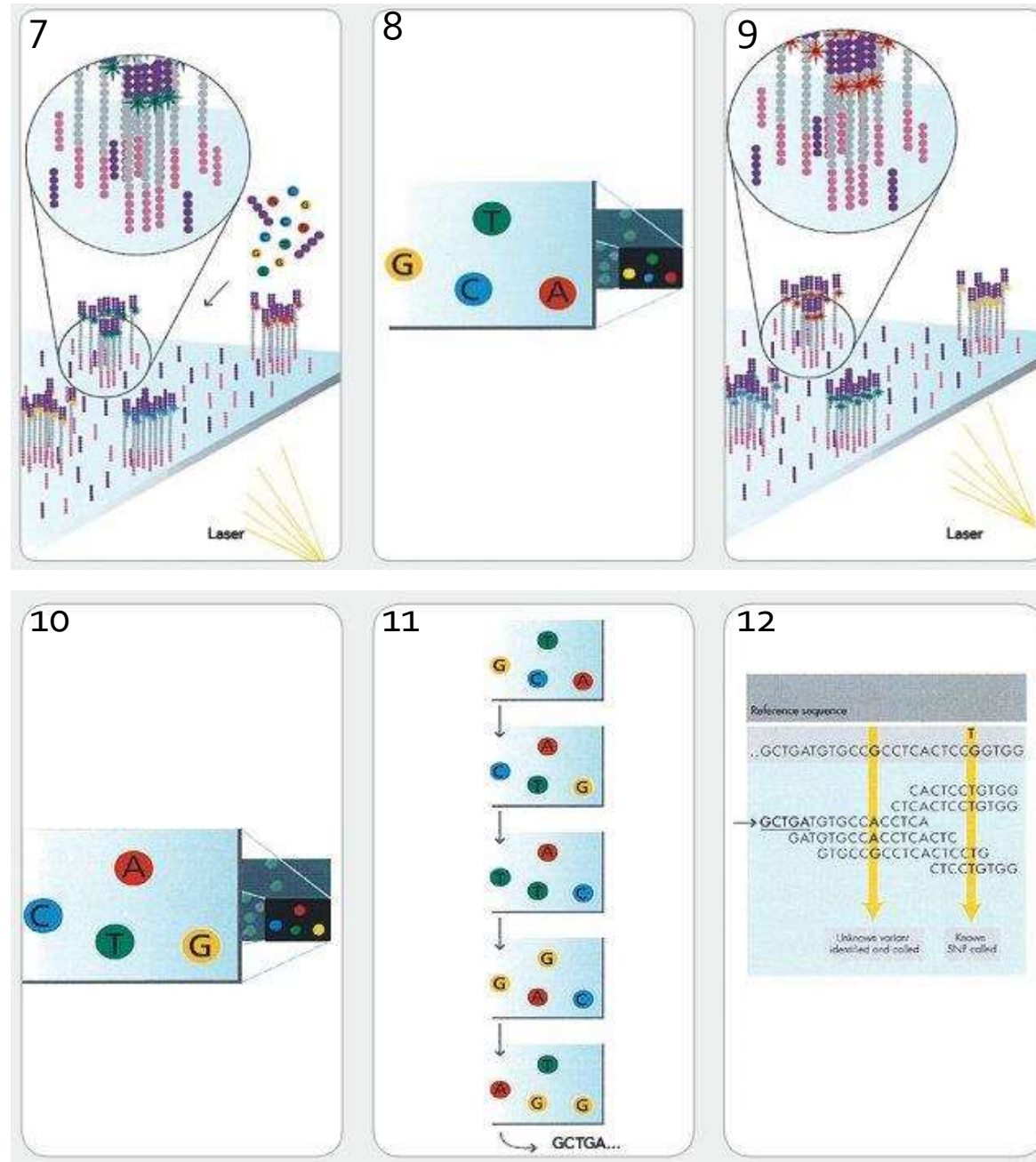


7. Determine first base
8. Image first base
9. Determine second base
10. Image second base
11. Sequence reads over multiple cycles
12. Align data

HiSeq 3000/4000
up to 5 billion
fragments/flow cell
2x 150bp fragment length

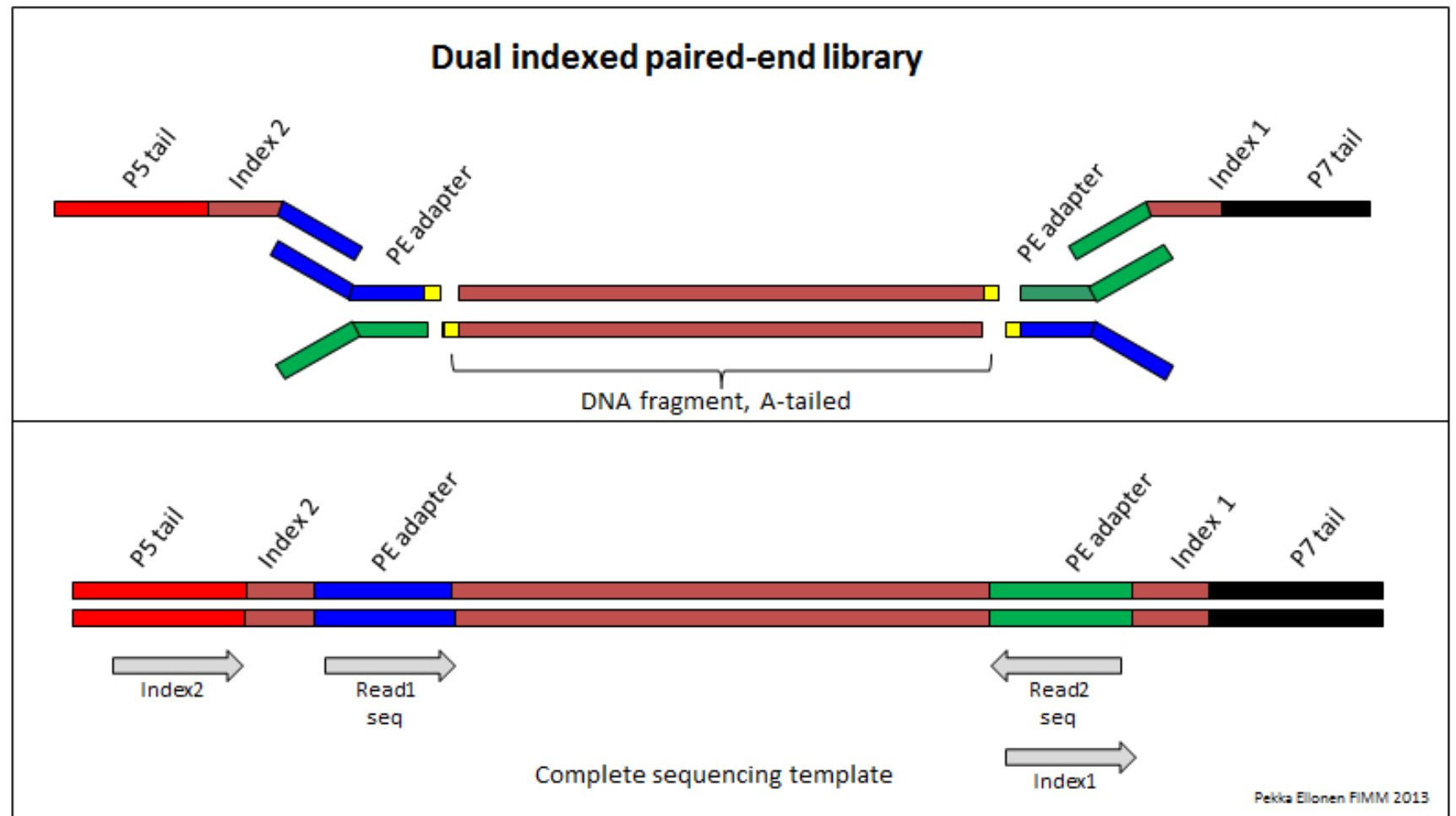
MiSeq
up to 25 million
2x300bp

Illumina



7. Add fluorescent-labelled nucleotides
8. Image first base
9. Flush flow cell and repeat
10. Image second base
11. Sequence reads over multiple cycles
12. Align data

DNA construct



Advantages of Illumina

- The industry standard
- Versatile (see applications later)
- Economic (price per Mb of data)
 - **MiSeq** – up to 2 x 300bp x 15M reads = **9 Gb**
 - ~\$2,500 (\$0.27)
 - **HiSeq** – up to 2 x 125bp x 200M reads = **50 Gb**
 - ~\$3,000 (\$0.06)

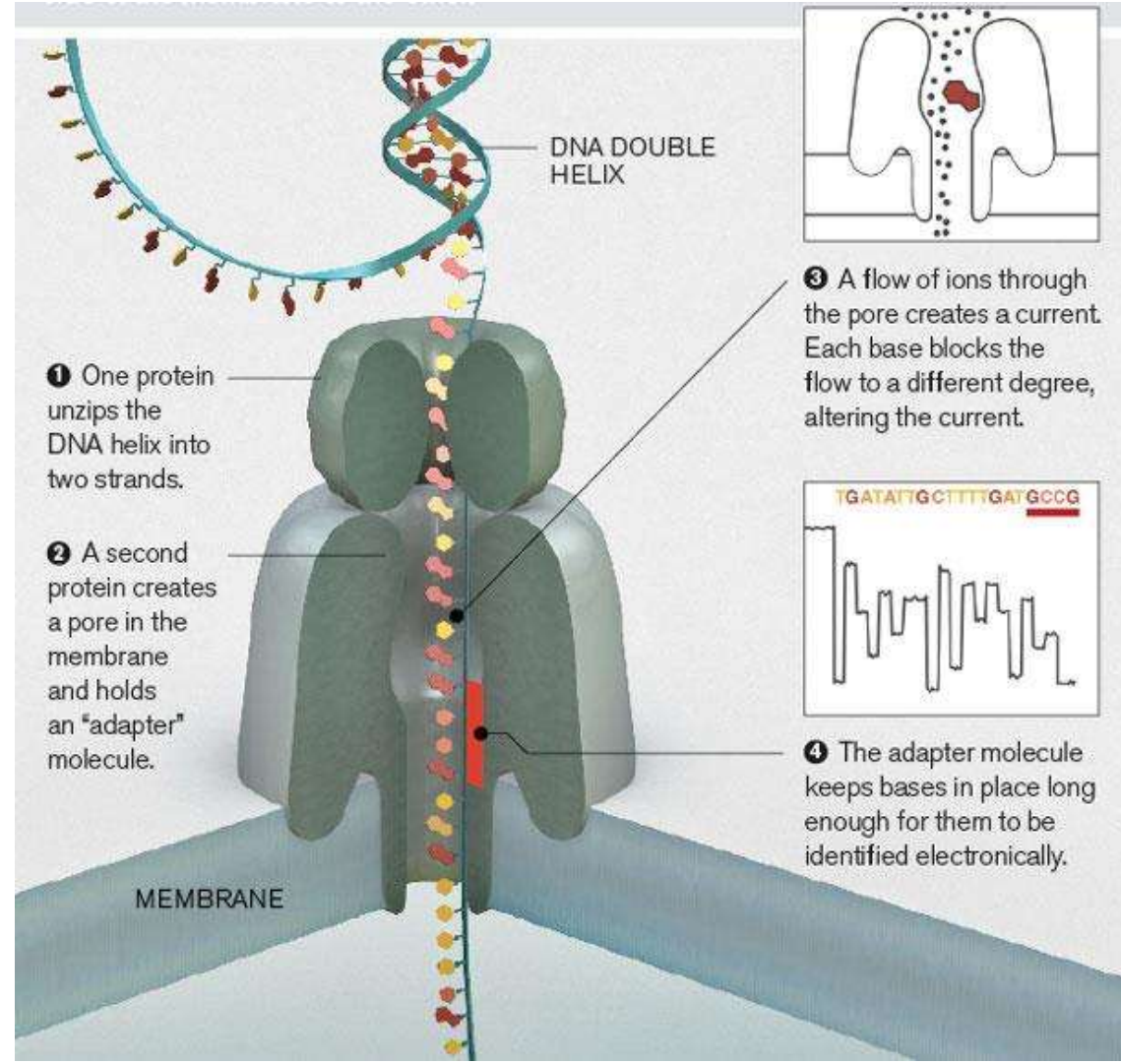
Compare Platforms:

<http://www.molecular ecologist.com/next-gen-fieldguide-2016/>

Other NGS Platforms

- **PacBio** – longest reads (>20kb)
 - genome assembly and full-length cDNA sequencing
 - DNA structural differences (e.g. inversions, transposon jumps)
 - BUT: see [10x genomics](#) for Illumina
- **Ion Torrent** – Fewer reads (1 Gb)
 - Lower up-front investment (\$80k vs \$150k+ for Illumina)
 - pre-existing variant panels (model organisms)
 - Amplicon sequencing (e.g. microbiome)
 - User-friendly, fast turnaround
- **Roche 454** – Obsolete; Lowest throughput (1 Mb)
 - cheaper long reads (up to 1 kb)
 - Microbiome studies

Experimental: MinION (Oxford Nanopore)



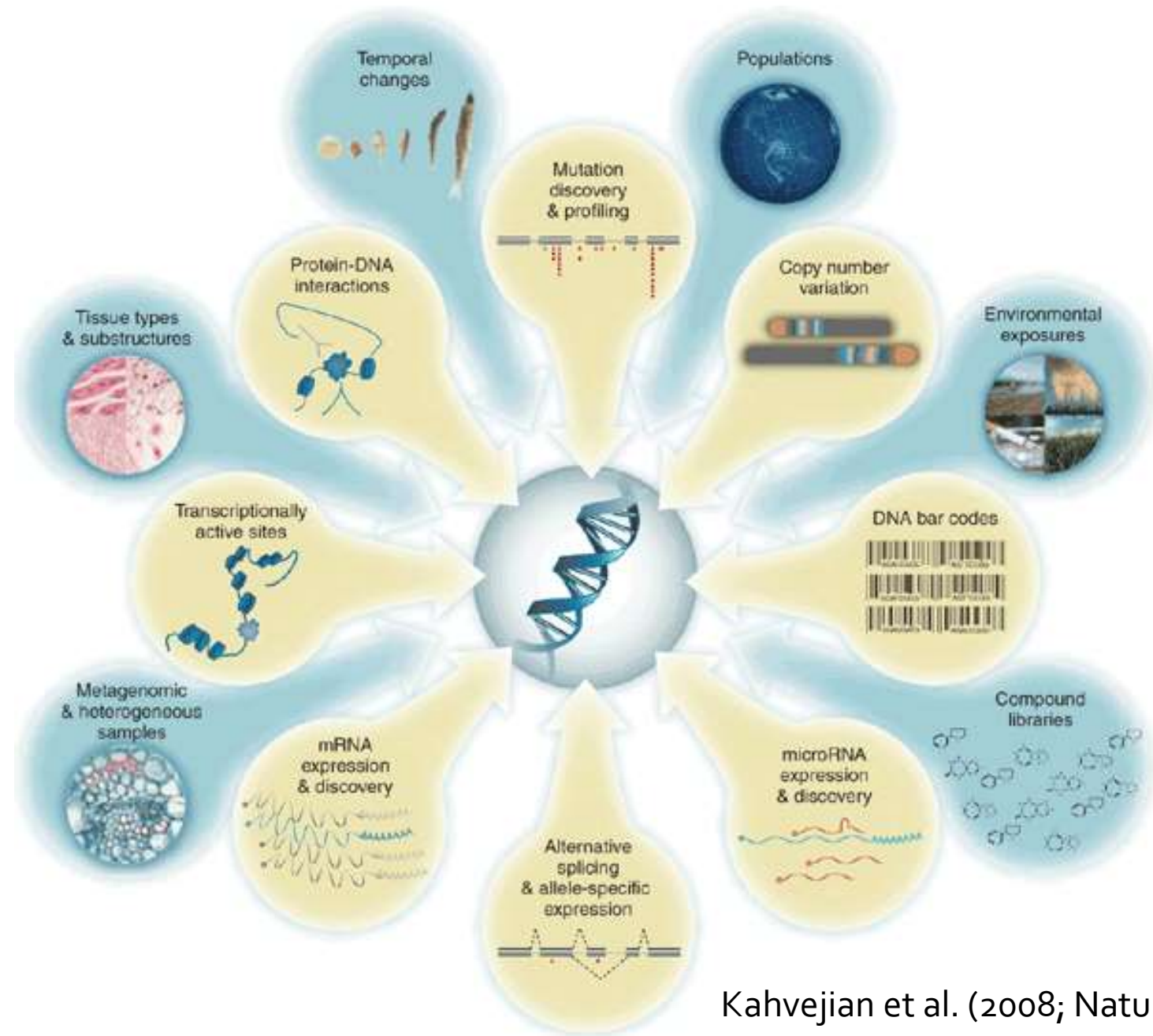
Why is NGS revolutionizing Biology?

- NGS-tools translate across disciplines.
For example:
- Metagenomics:
 - Soil, water, human gut microbiomes
- Variant detection & GWAS
 - QTL mapping for agriculture, adaptive traits, environmental stressors, human health



Applications

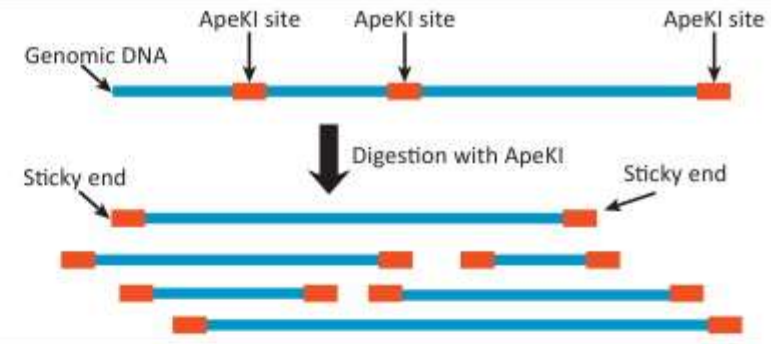
Applications



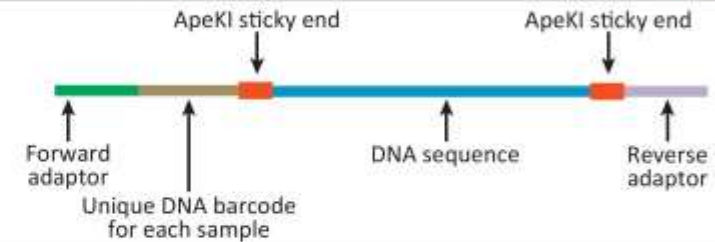
Kahvejian et al. (2008; Nature Biotech)

Genotype-by-sequencing (GBS)

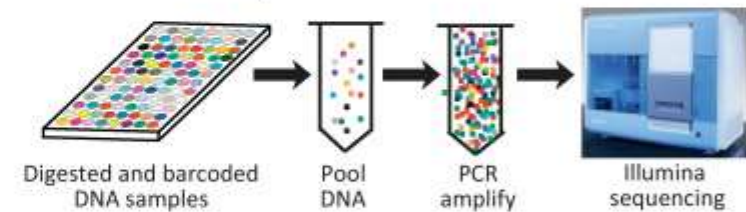
Step 1
Construct reduced representation libraries (RRLs) by digesting each DNA sample with a restriction enzyme (ApeKI)



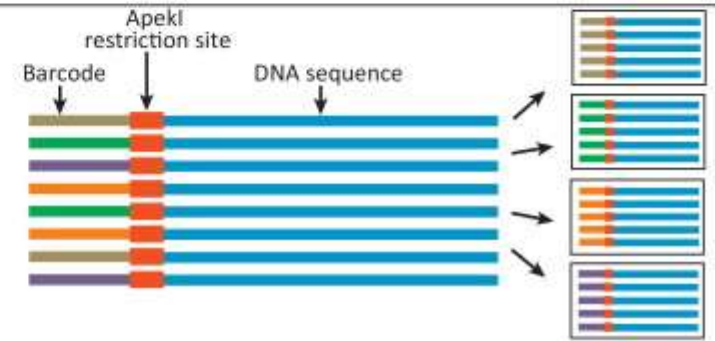
Step 2
Ligate custom 'barcoded' adaptors to sticky ends of restriction site. Each sample has its own unique barcode sequence



Step 3
Pool digested and barcoded DNA into a single tube. Perform PCR amplification, library preparation, and sequencing on Illumina platform

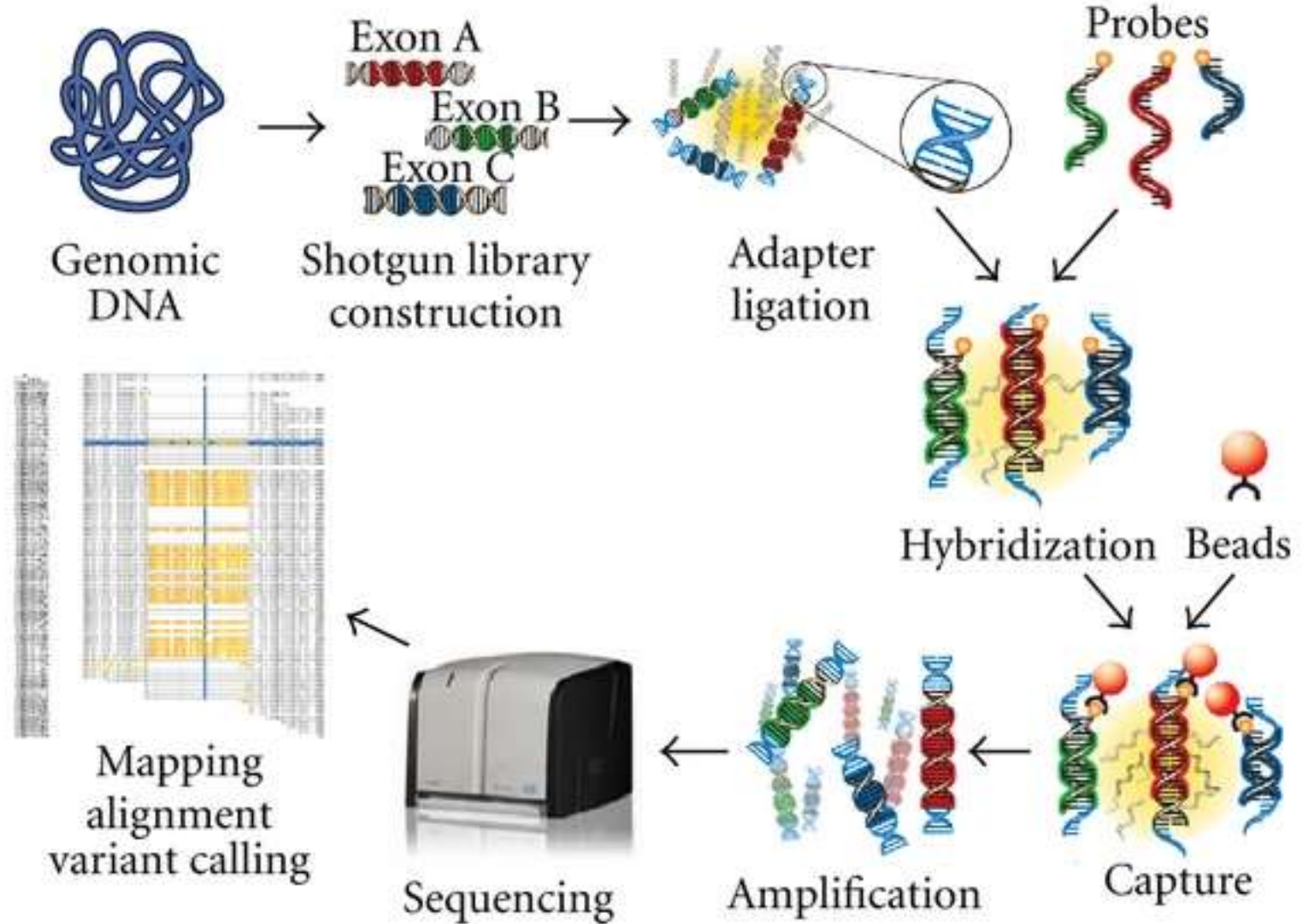


Step 4
Use barcodes to assign sequences to samples. Produce a file of DNA sequence data for each sample

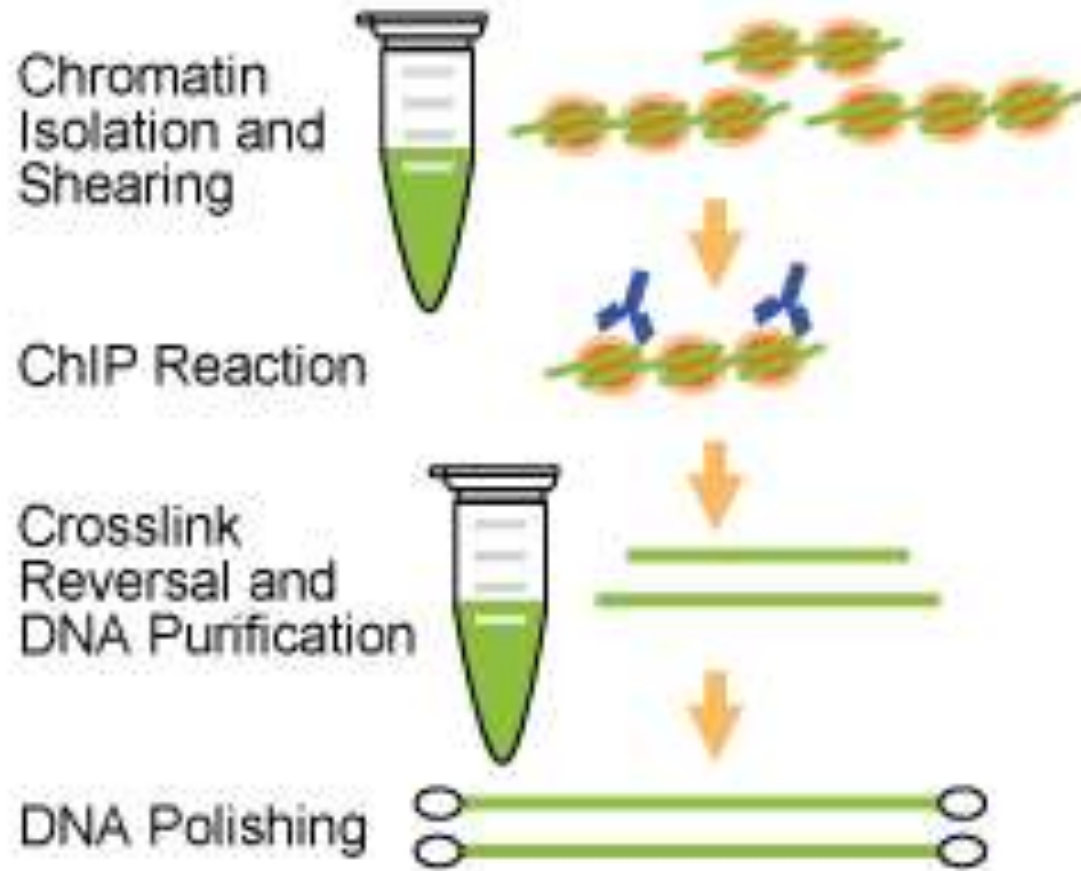


TRENDS in Genetics

Target-capture sequencing

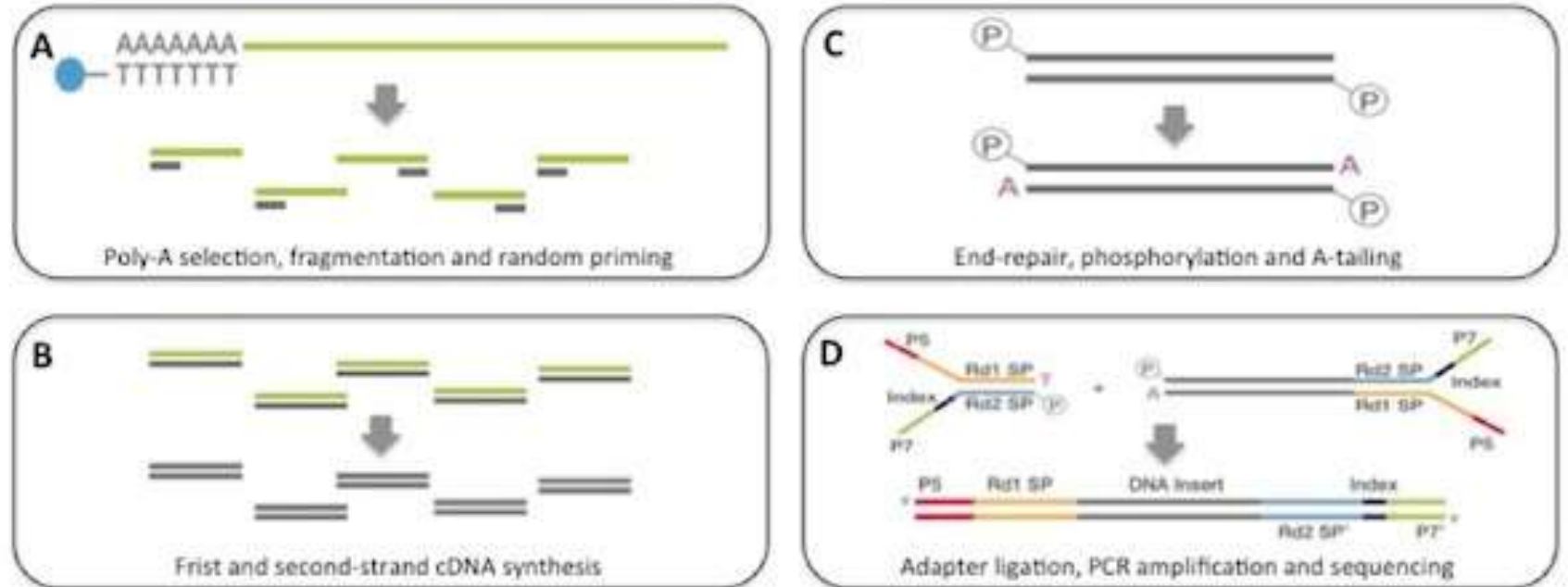


ChIP-seq (protein-targets)



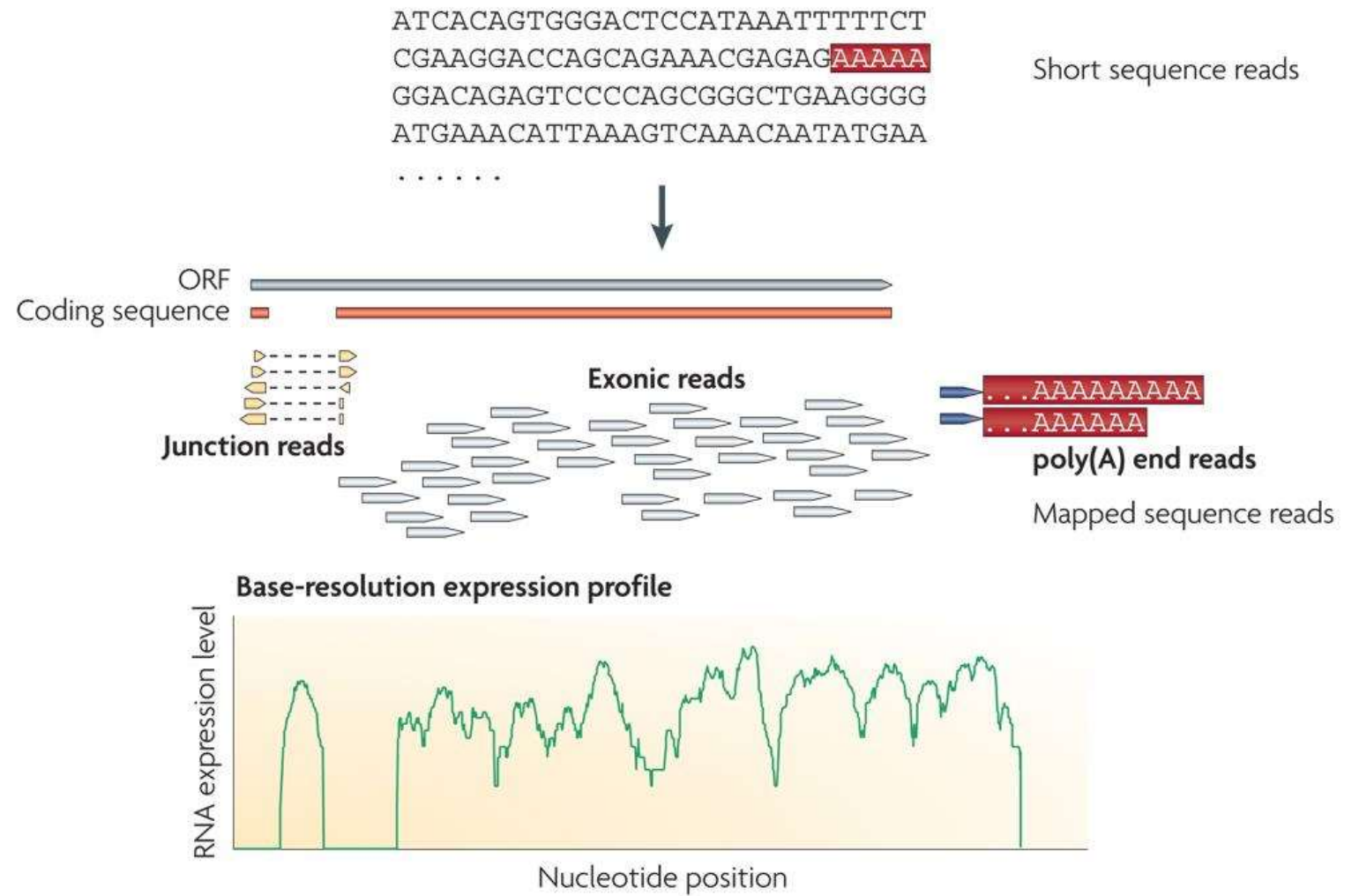
RNA-Seq

Illumina Tru-Seq RNA-seq protocol

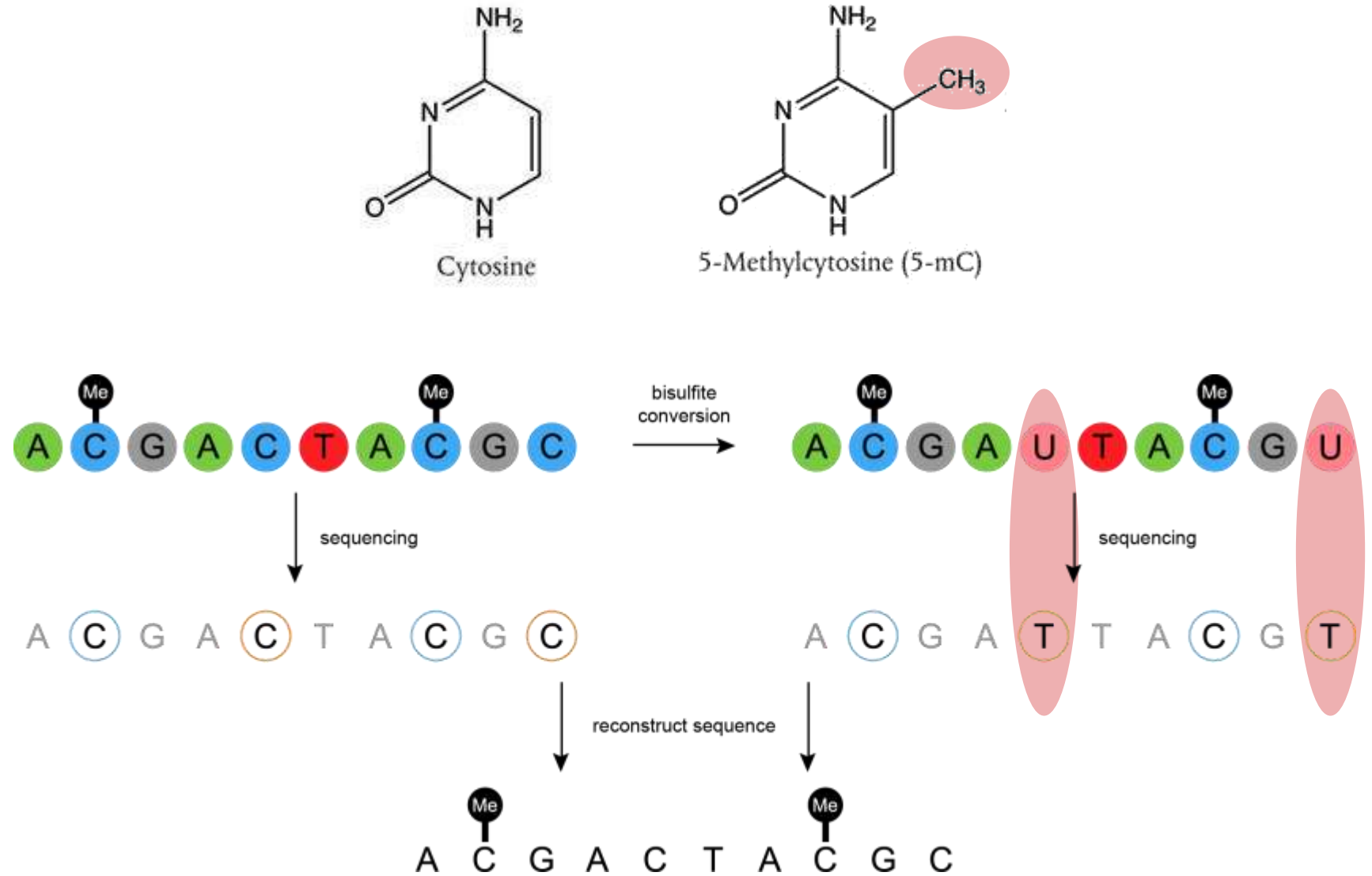


Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

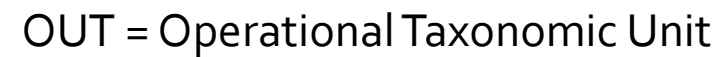
RNA-Seq expression profiles



Bisulfite sequencing

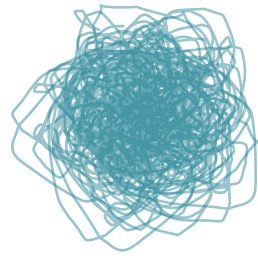


Metagenomics

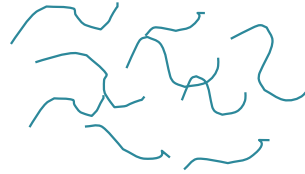


OUT = Operational Taxonomic Unit

Whole-Genome Shotgun (WGS) sequencing



Extract



Fragment



Sequence

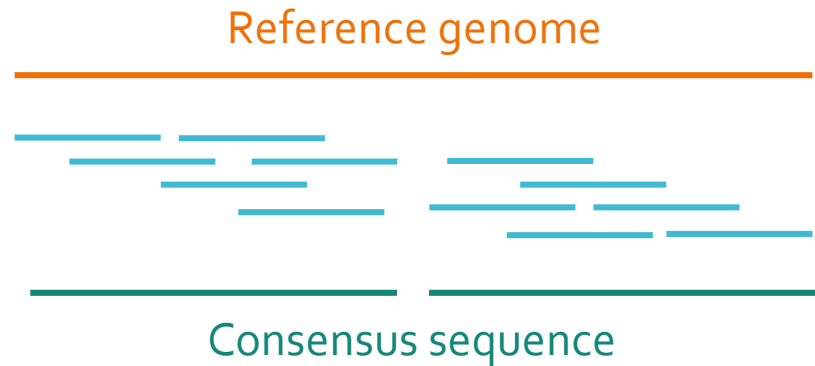


Align

Assembly with or without a reference

Alignment to reference

- sometimes called 'resequencing'
- common in 'model systems'
- e.g. human, mouse, *Drosophila*, *C. elegans*, *Arabidopsis*



de novo assembly

- 'non-model' organisms



De novo assembly

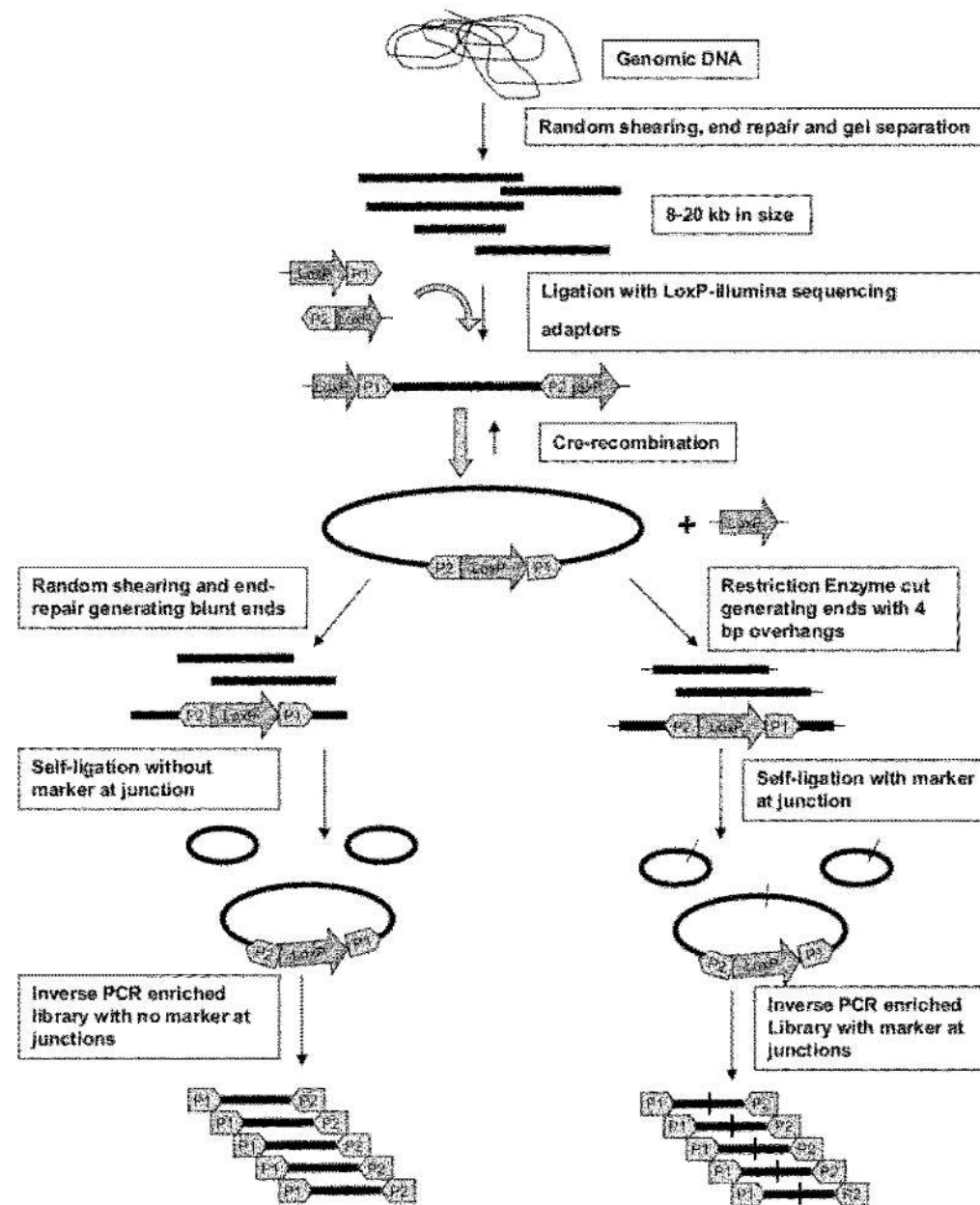
Figure 5. *De Novo* Assembly with Mate Pairs



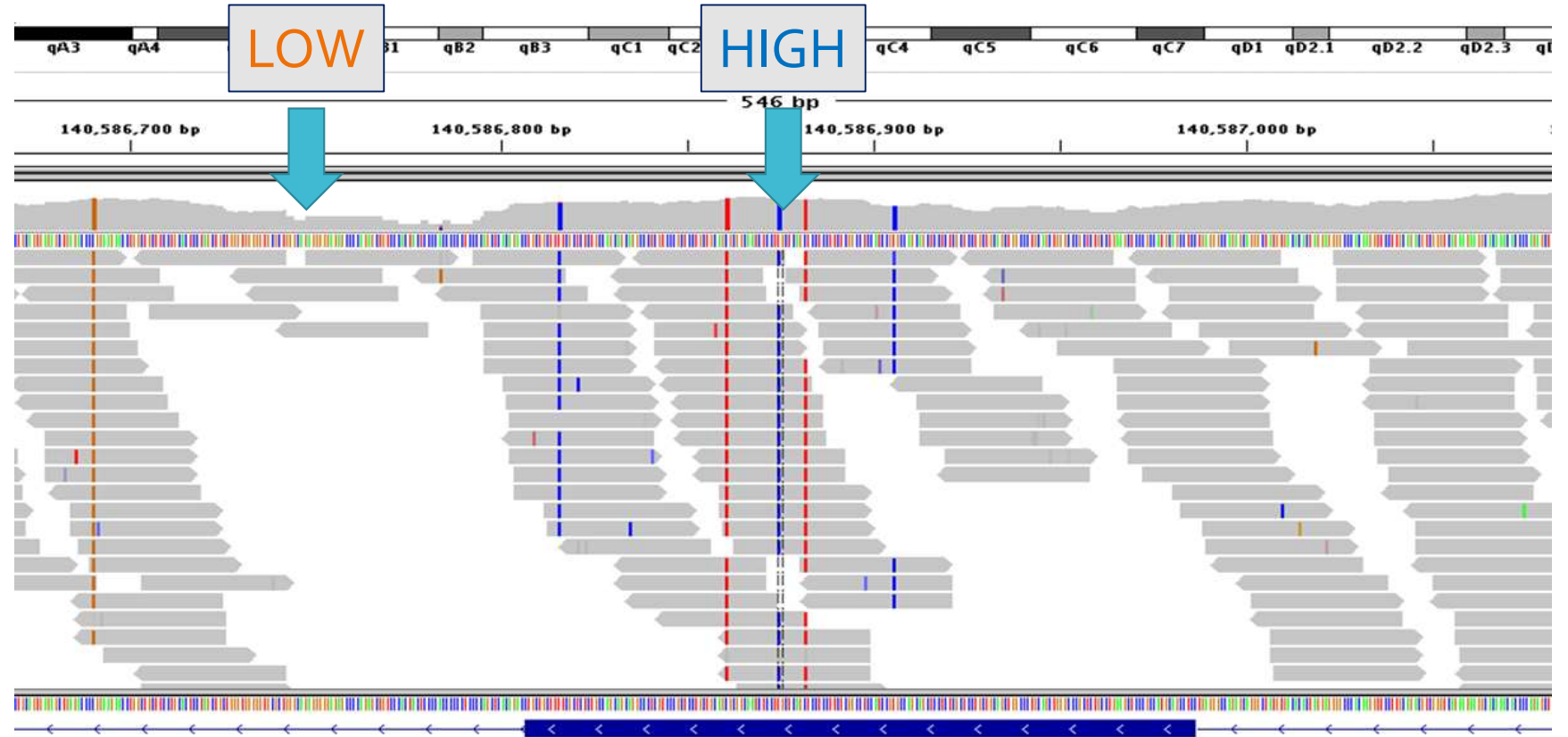
Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

- IMPORTANT: Paired End vs Mate-Pair

Mate pair library



Coverage



- = average # reads at any given location (base pair)
- A 'random' sampling process
- (Higher is better)

de novo assembly programs*

- ABySS
- ALLPATHS-LG
- CORTEX
- CLC Genomics Workbench
- DISCOVAR de novo
- Geneious
- IDBA
- MaSuRCA
- MIRA
- PLATANUS
- RAY
- SOAP de novo

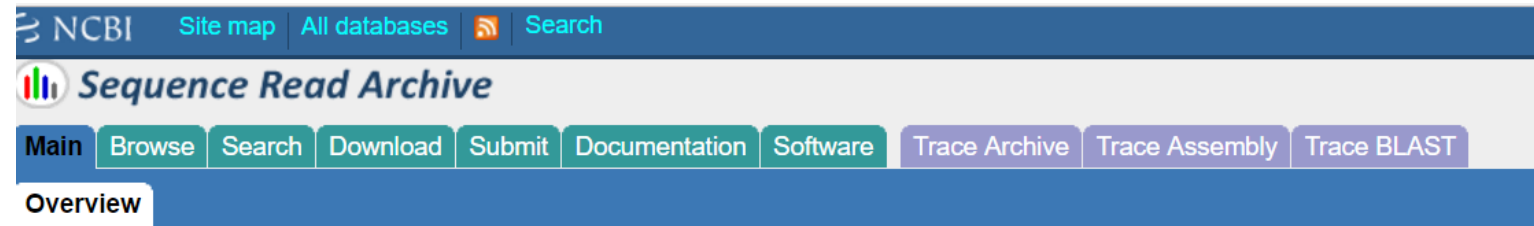
*Large genome assemblers (>0.5Gb)
Gold = commercial/restricted license



Challenges

'Big Data' Computation

<https://trace.ncbi.nlm.nih.gov/Traces/sra/>



The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.

SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

Please check [SRA Overview](#) for more information.

Submitting to SRA

Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

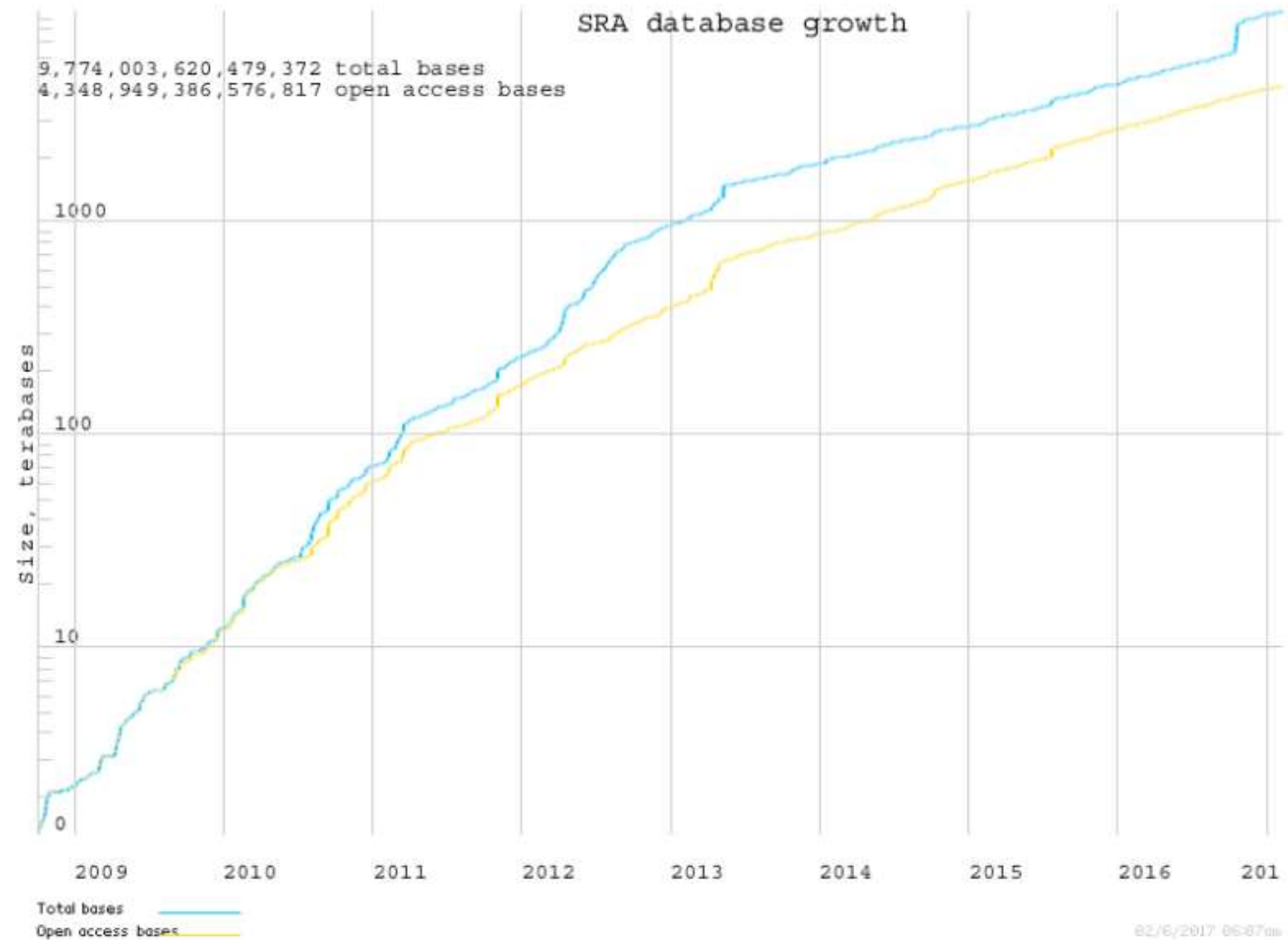
- [Submission Quick Start](#)
- [Frequently Asked Questions](#)
- [Submitter Login](#)

Using SRA Data with SRA Toolkit

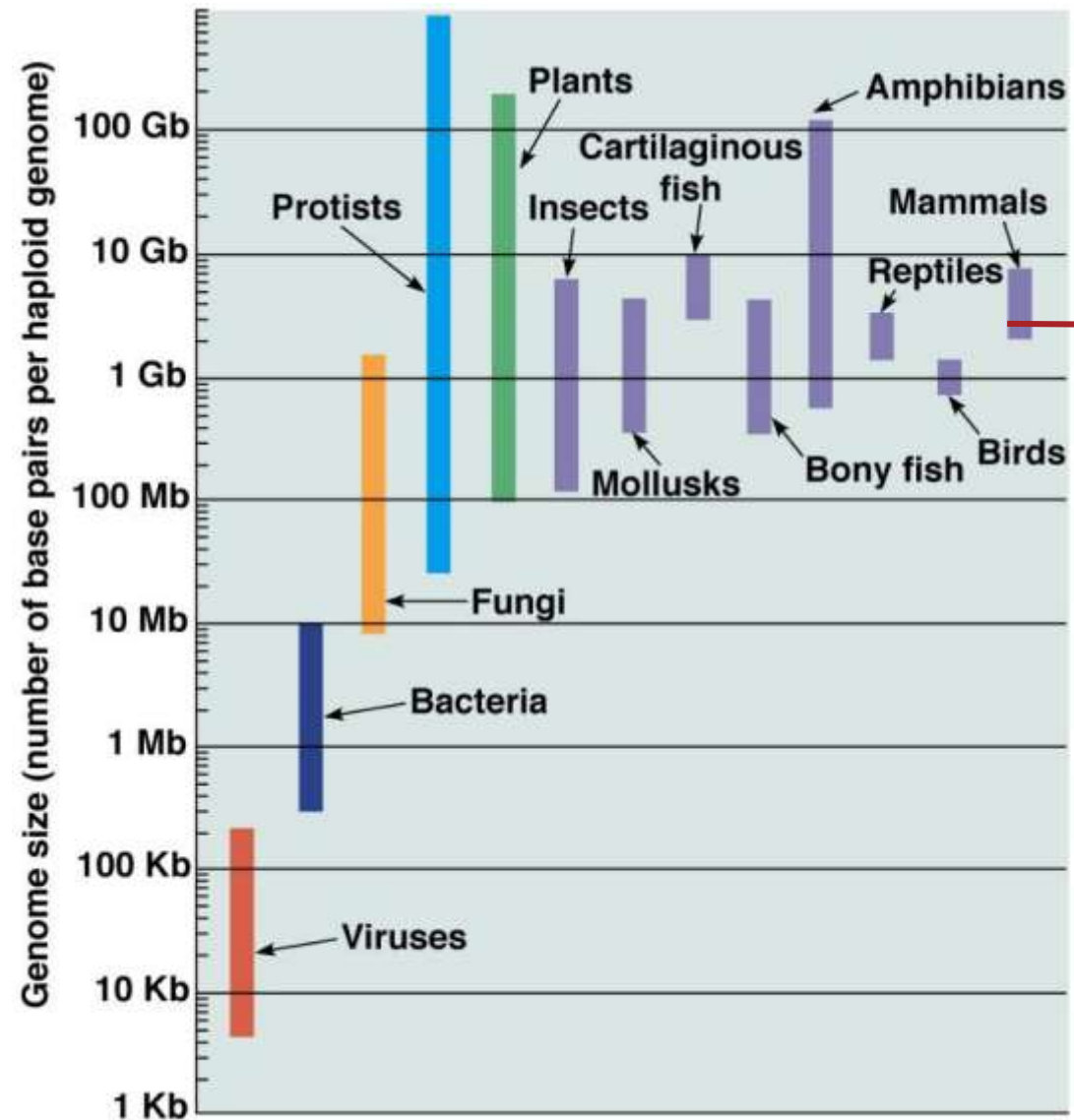
Use SRA data to validate experimental results, increase sample sizes, determine variance and open up new avenues of research.

- [Documentation](#)
- [Usage Guide](#)
- [Download](#)
- Get sources code on [GitHub](#) (for developers using SRA)

'Big Data'
(~10 Petabytes)



Large genomes



Human
(~3.2Gb)

Assembly challenges

- Large repeats – creates gaps
- Errors – cannot be assembled
- Low coverage – creates gaps
- Unequal coverage – confuses error correction



NGS and Ecological Genomics

in the Colautti Lab

Ecological Genomics

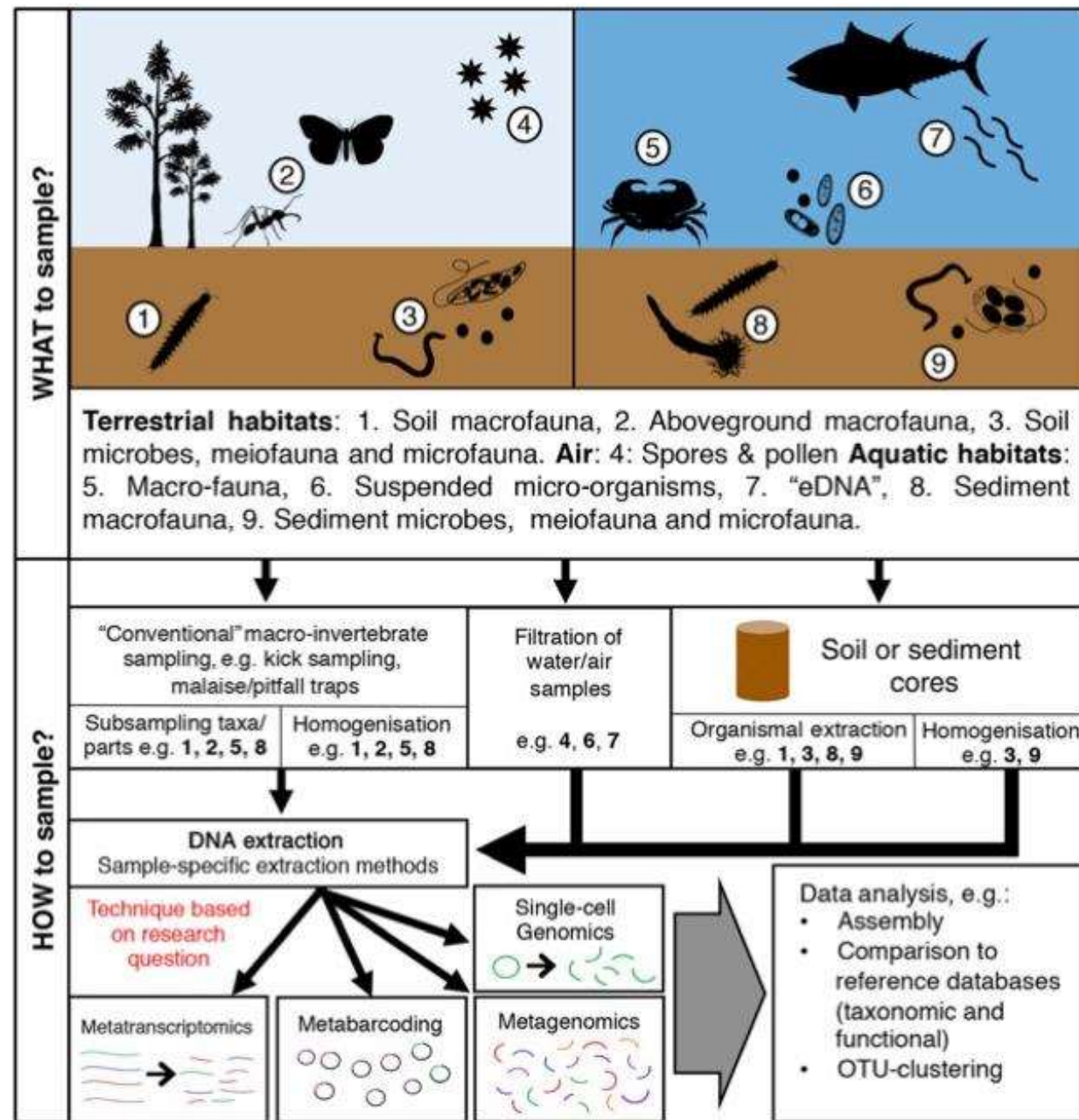


Figure 1. (Creer et al., 2016) Schematic of molecular ecology workflow.

Ecology and Evolution in the Anthropocene

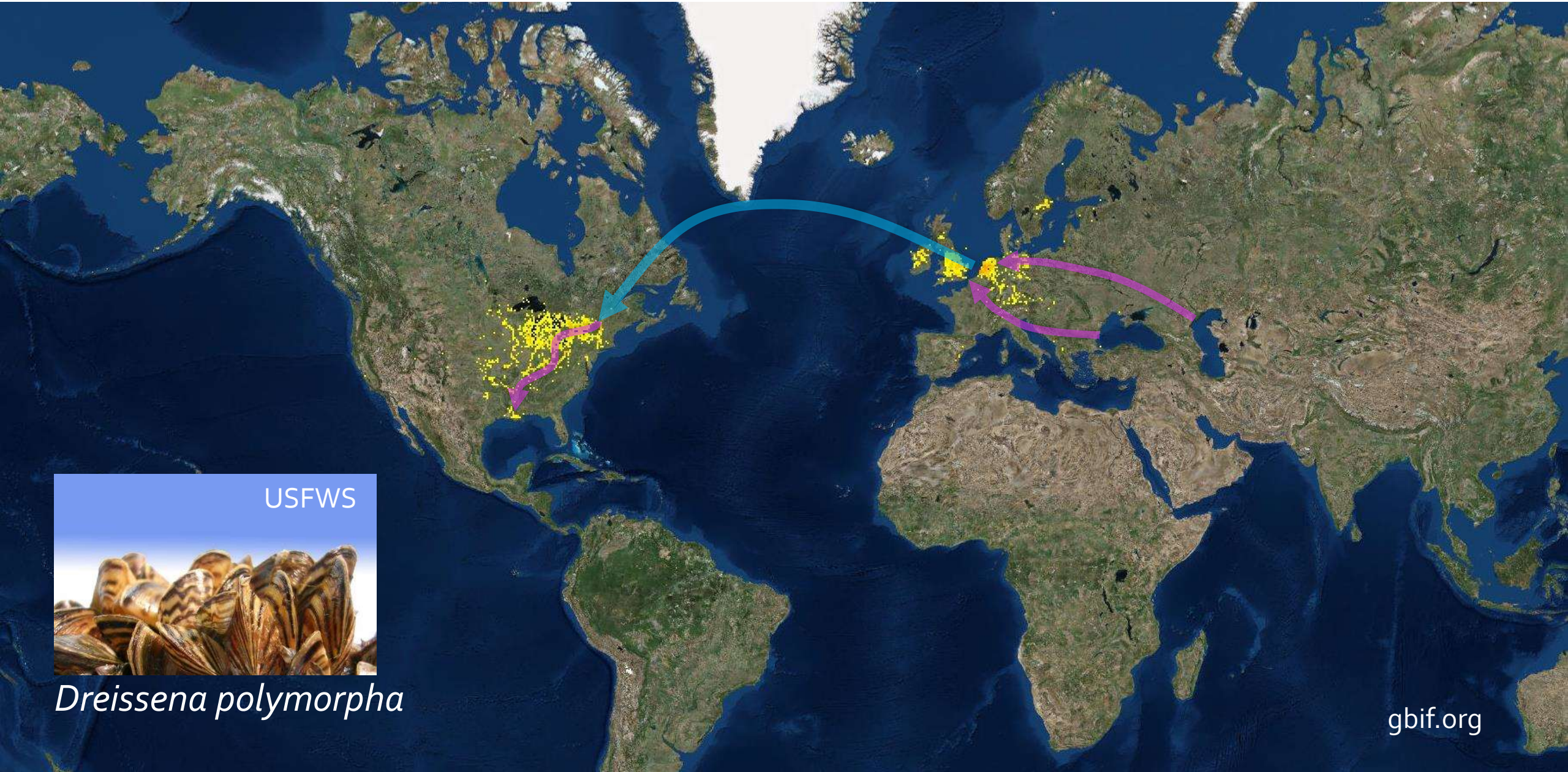


Magellan I Sverige

James Hoznik

Environment --> Natural Selection --> Genome Evolution

Ecological genomics of invasive species



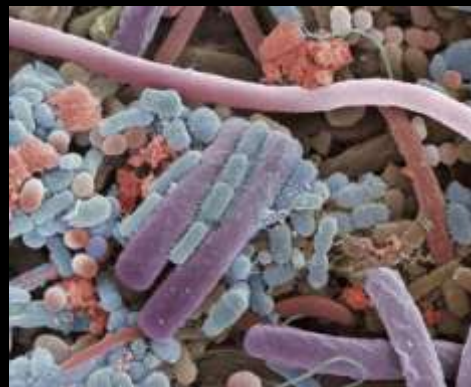
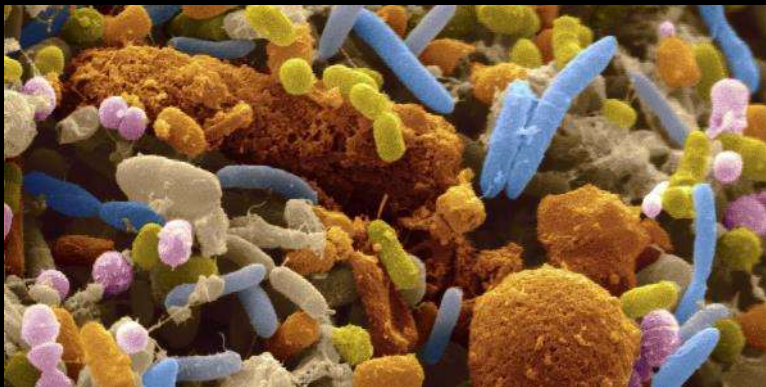
USFWS

Dreissena polymorpha

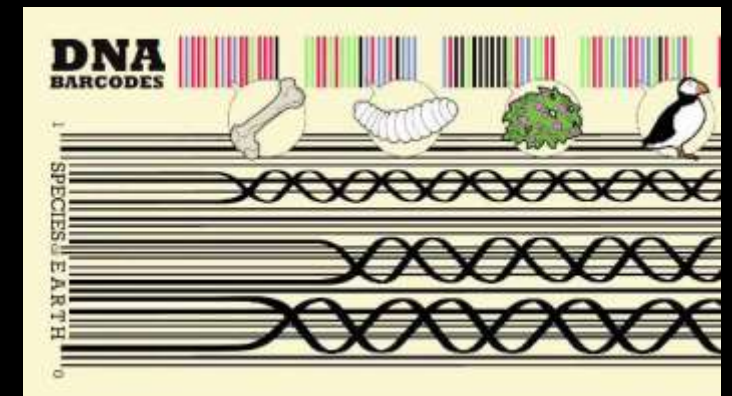
Invasions: grand, unplanned ecological experiments



eDNA and DNA barcodes for environmental monitoring



Barcode of Life Project
www.boldsystems.org



<https://www.youtube.com/watch?v=ZImiXgU6bCk>

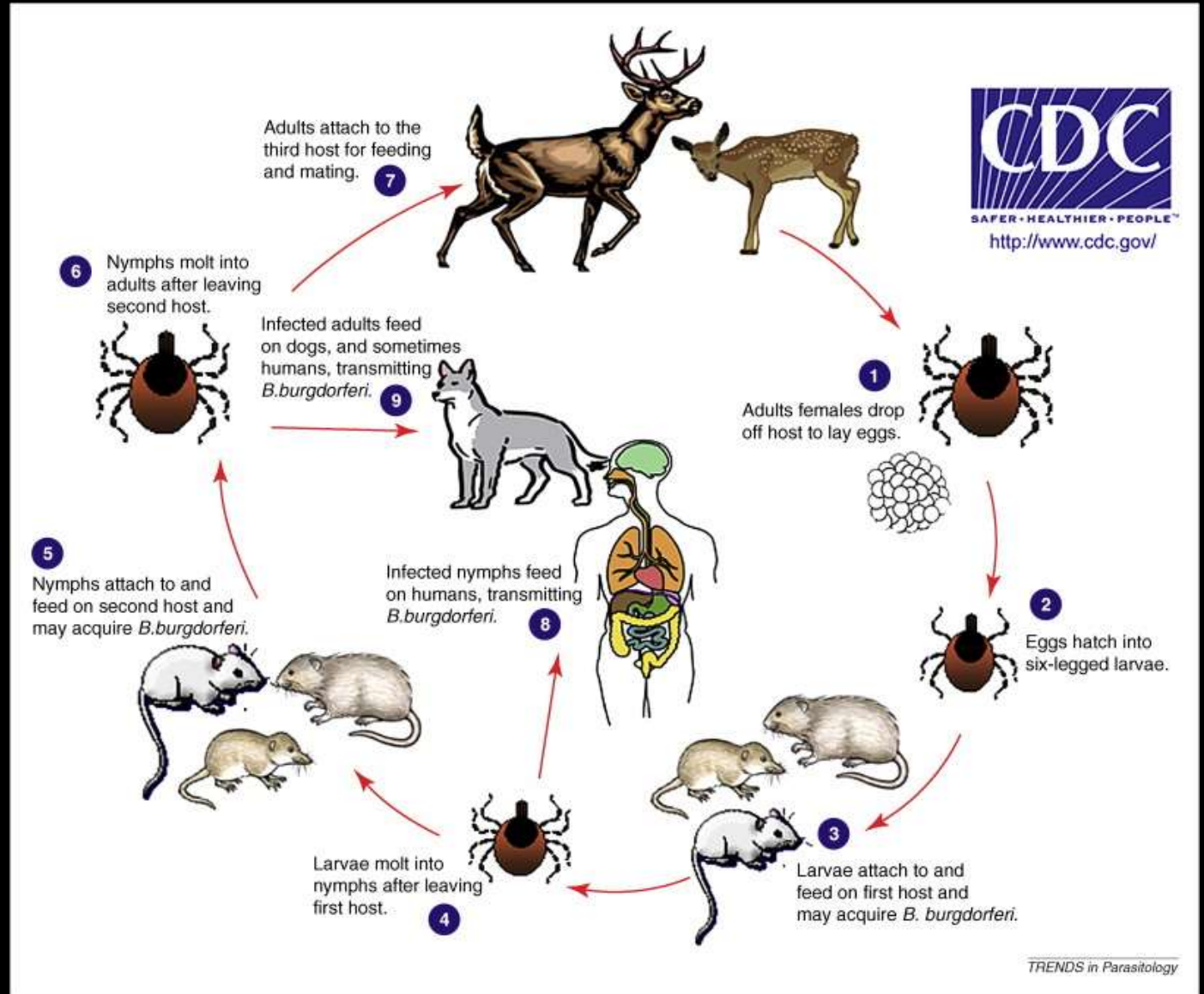
Tick microbiome analysis



Lyme disease in Canada (cbc.ca)



Ixodes scapularis (black-legged tick)



NGS @Queen's

- Biology Department has the only Illumina sequencer on campus (MiSeq)
- Other machines (IonTorrent, microarrays) spread across campus
- Queen's has no university-wide genomics core facility – very unusual for a research-intensive university

NGS @ Queen's: Typical WGS sequencing workflow

Library Prep

- Extract & Purify DNA
- Fragment and size-select
- Ligate sequencing primers

Sequencing

- Generate FASTQ file – includes sequence data (short-reads) and quality score

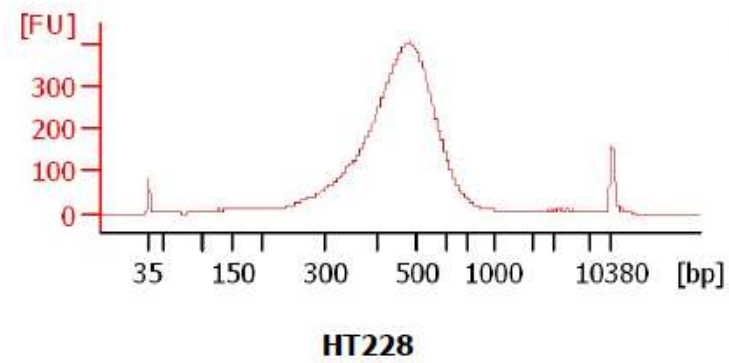
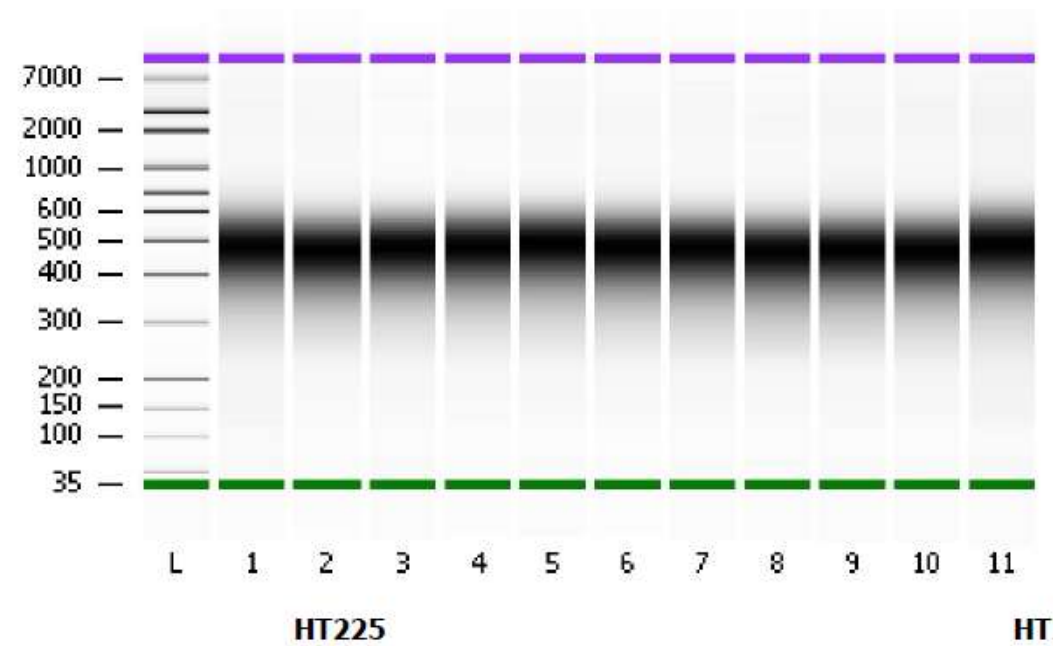
Assembly

- Assemble short-reads into contigs and scaffolds

Analysis

- e.g. genome annotation; comparative genomics; structural variant detection; taxonomy ID (metagenomics); align short reads for variant detection (e.g. for GWAS)

QC @ Queen's



Agilent Bioanalyzer

MiSeq @ Queen's

Miseq Kits	Running time (est. in hours)	Max read length (bp)	Number of Reads (million)	Data generated (giga-bases)	Plan to Stock at the core?	Price \$CDN (USD x 1.33)
MiSeq Reagent ver.3 2 x 300	65	600	25	15	Yes	2438
MiSeq Reagent ver.2 2 x 250	39	500	15	7.5	No	1850
MiSeq Reagent ver.2 2 x 150	24	300	15	4.5	Yes	1676
MiSeq Reagent ver.3 2 x 75	24	150	25	3.75	Yes	1470
MiSeq Micro 2 x 150	24	300	4	1.2	No	1432
MiSeq Reagent ver.2 2 x 25	6	50	15	0.75	No	1347
MiSeq Nano 2 x 250	39	500	1	0.5	No	1180
MiSeq Nano 2 x 150	24	300	1	0.3	No	1004

de novo
sequencing
Application



Ecological genomics of *Alliaria petiolata*



Loren Rieseberg



Oliver Bossdorf



Ecological genomics of *Alliaria petiolata*

Why *Alliaria petiolata*?

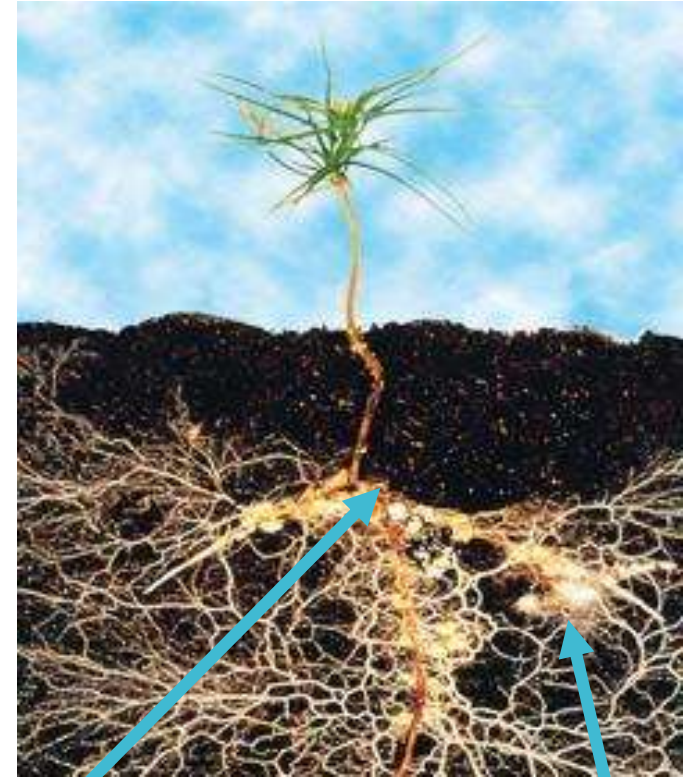
- Problematic invasive species in North America
- Easy to identify
- Simple lifetime fitness estimate
- Brassicaceae, self-pollinated
- Ecologically important chemistry
 - (Enemy release; allelopathy/soil biota)



Alliaria
petiolata
ecology:
plant-microbe
interactions



*Alliaria
petiolata*
ecology:
plant-microbe
interactions



Mycorrhizal fungus

Root

*Alliaria
petiolata*
ecology:
plant-microbe
interactions



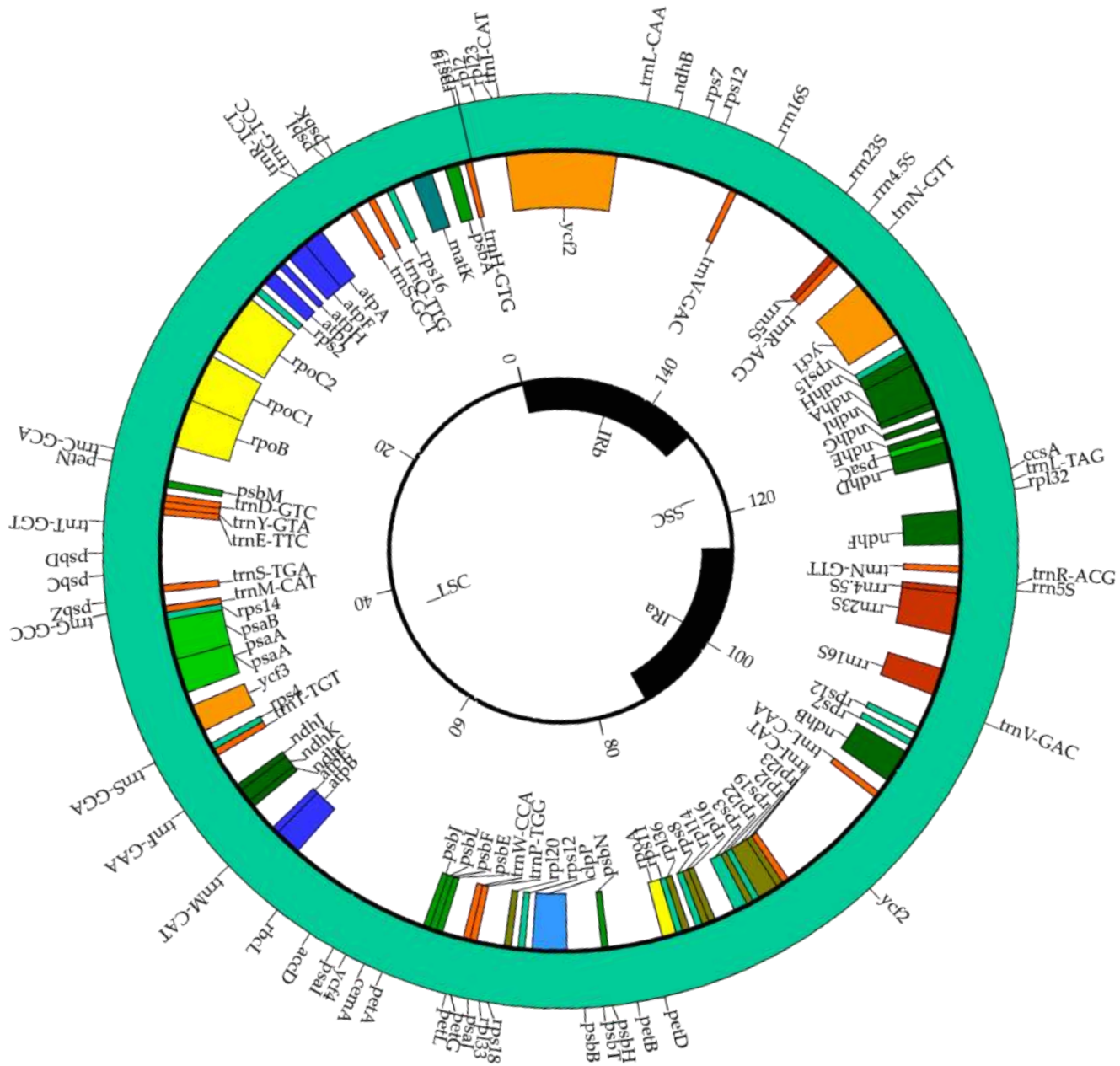
Global garlic mustard field survey

- International collaboration
 - 164 participants from 16 countries
- Field surveys
 - 45,000 field measurements
 - Plant performance traits
 - Herbivory & fungal damage
- Genetic resource
 - Seeds from >5,000 plants
 - 395 locations across Europe and North Am.

Population genomics: Some major questions

- How many introductions to North America, from which parts of Europe, and how has it been moved around?
- Which genes affect survival and reproduction in response to:
 - cold vs warm climates
 - wet vs dry conditions
 - native vs introduced range
 - herbivores and pathogens
 - competition with other plants
- Does selection act mainly on a few loci with large effects on phenotype, or many loci with smaller effects?
- How important is standing genetic variation vs new mutations?
- How important are SNPs vs genome rearrangements?

Alliaria
petiolata draft
chloroplast
genome
153,190 bp

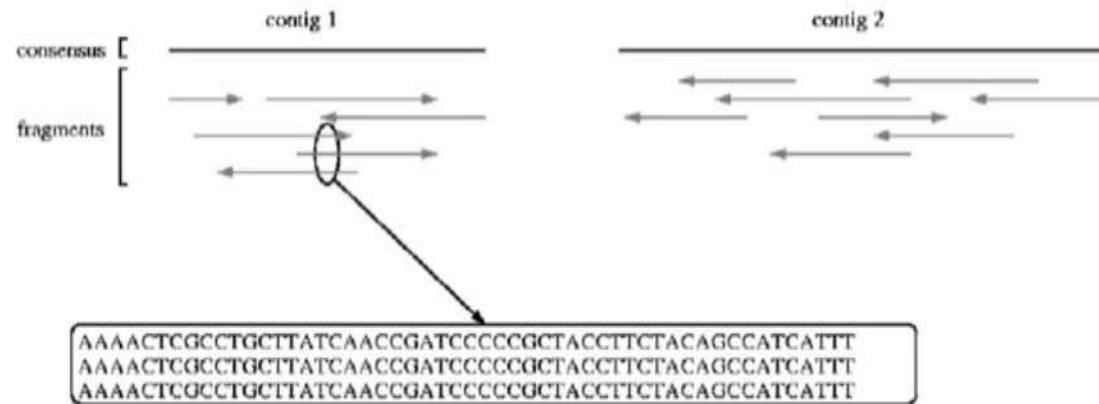


de novo assembly (Sanger)

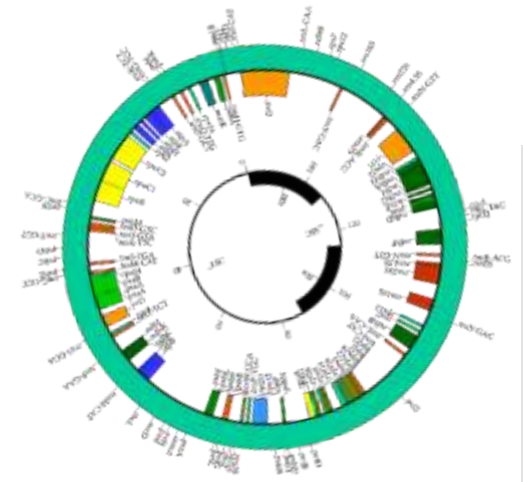
Overlap/layout/consensus

Essentially,

1. Calculate all overlaps
2. Cluster based on overlap.
3. Do a multiple sequence alignment



UMD assembly primer (cbcb.umd.edu)



de novo assembly (NGS)

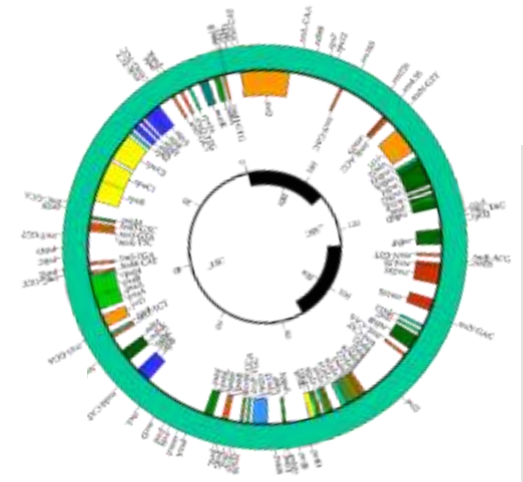
de Bruijn graph method based on k-mers

K-mers

Break reads (of any length) down into multiple overlapping words of fixed length k .

ATGGACCAGATGACAC ($k=12$) =>

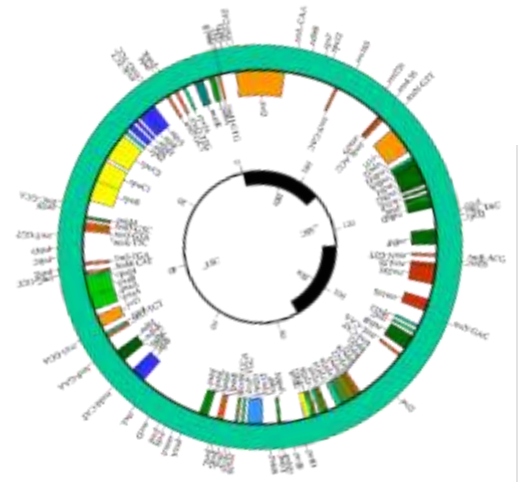
ATGGACCAGATG
TGGACCAGATGA
GGACCAGATGAC
GACCAGATGACA
ACCAGATGACAC



de novo assembly (NGS)

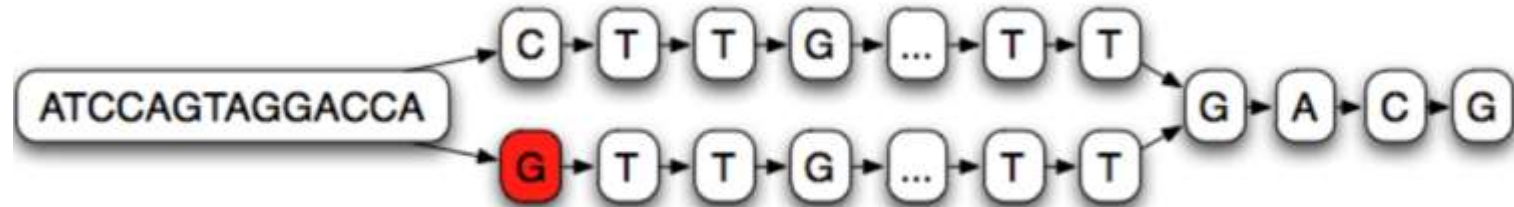
de Bruijn graph
'bubbles' in the assembly

K-mer graph (k=14)



ATCCAGTAGGACCACTTGACAGGCGATTGACG

ATCCAGTAGGACCA**G**TTGACAGGCGATTGACG



De novo assembly

In practice

```
$ tar -zxvf CAC_files.tar.gz /home/hpc####
```