TOPIC 1:
Introduction to BIOL525D

BIOL525D - Bioinformatics for Evolutionary Biology 2021

Instructors

Dr Tom Booker

booker@zoology.ubc.ca

Dr Julia Kreiner

julia.kreiner@botany.ubc.ca

WEBSITE: https://ubc-biol525d.github.io/

Overview of the week

1. Introduction: Scope of course and overview of technology [Tom]
2. Introduction to command line programming [Tom]
3. Fastq files and quality checking/trimming [Kay]
4. Alignment: algorithms and tools [Tom]
5. Assembly: transcriptome and genome assembly [Kay]
6. RNAseq + differential expression analysis [Kay]
7. SNP and variant calling [Julia]
8. Population genomics and plotting in R (Part 1) [Julia]
9. Population genomics and plotting in R (Part 2) [Julia]
10. Case studies [Tom/Julia]

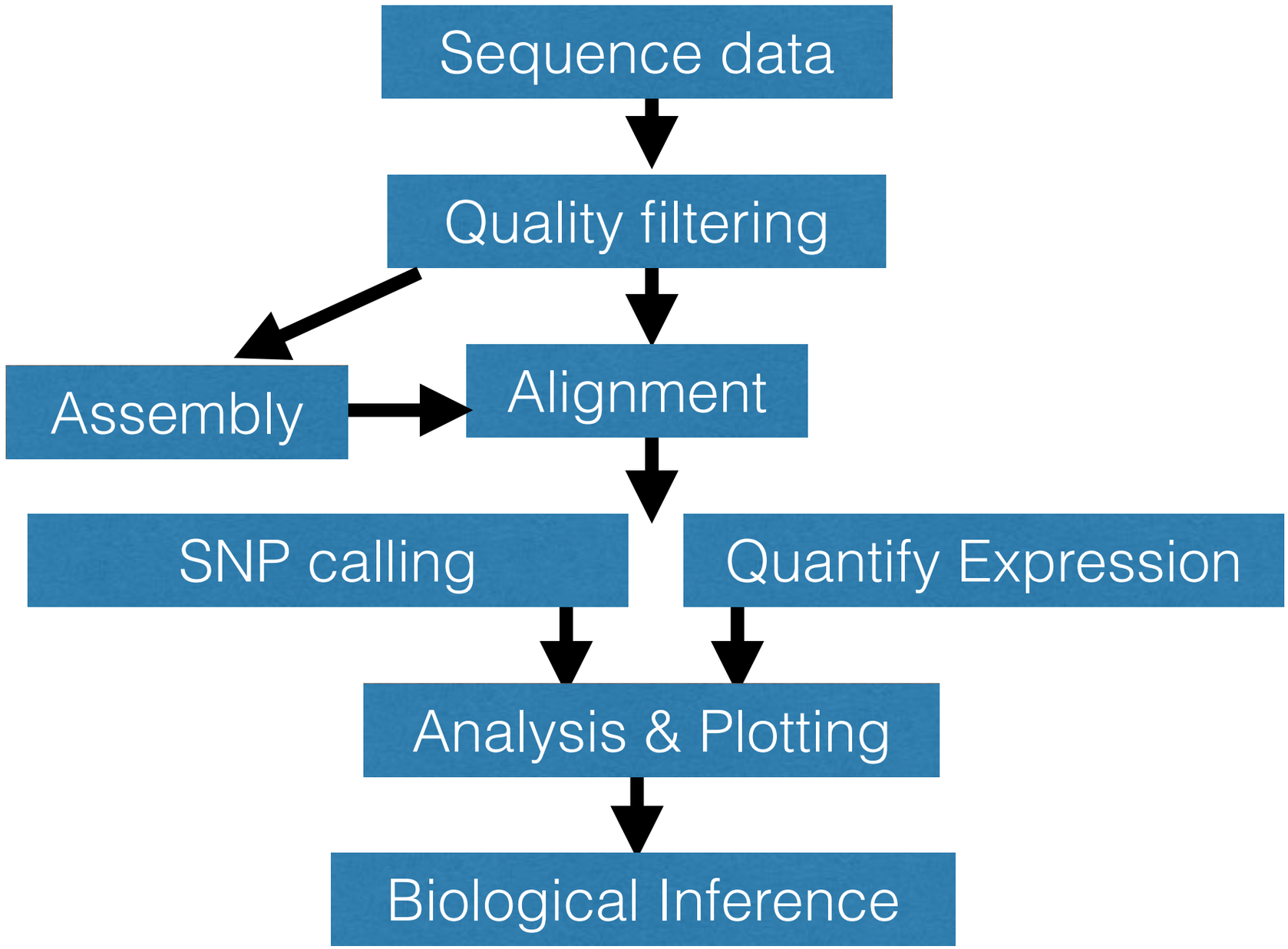
Goal of any sequencing project

Raw sequence data

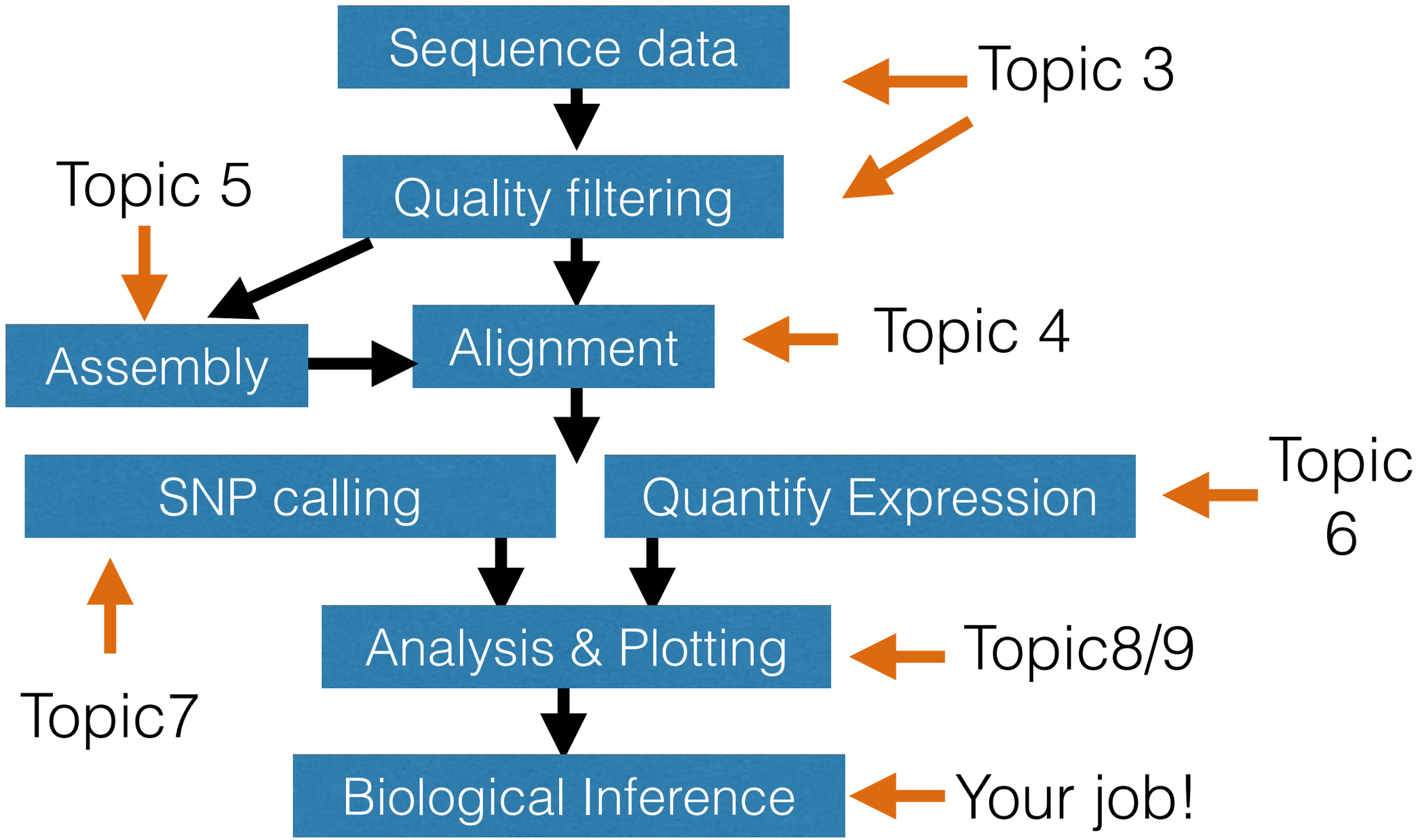
????

Biological inference

Rough outline



Rough outline



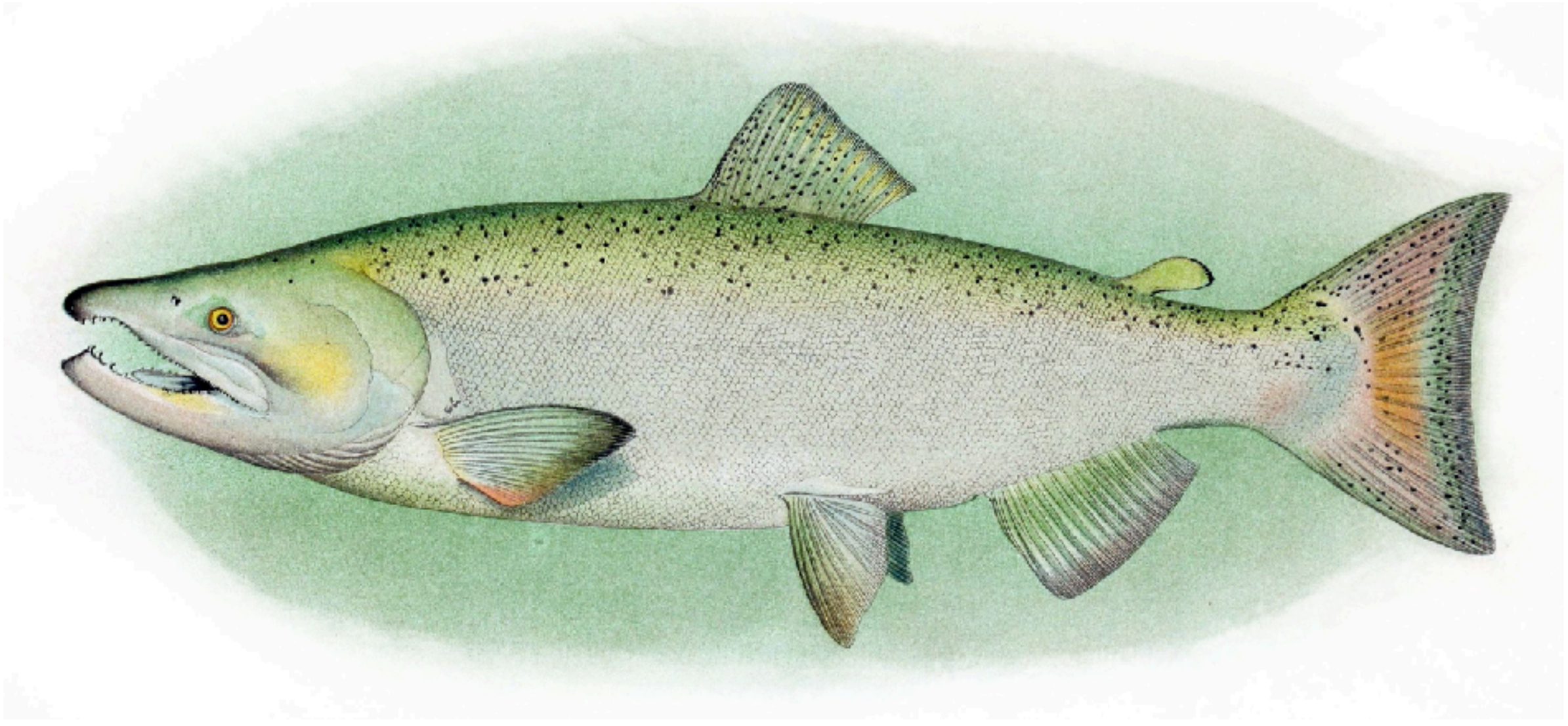
Learning outcomes

The field of bioinformatics is rapidly evolving

A general grasp of concepts and approaches is very valuable

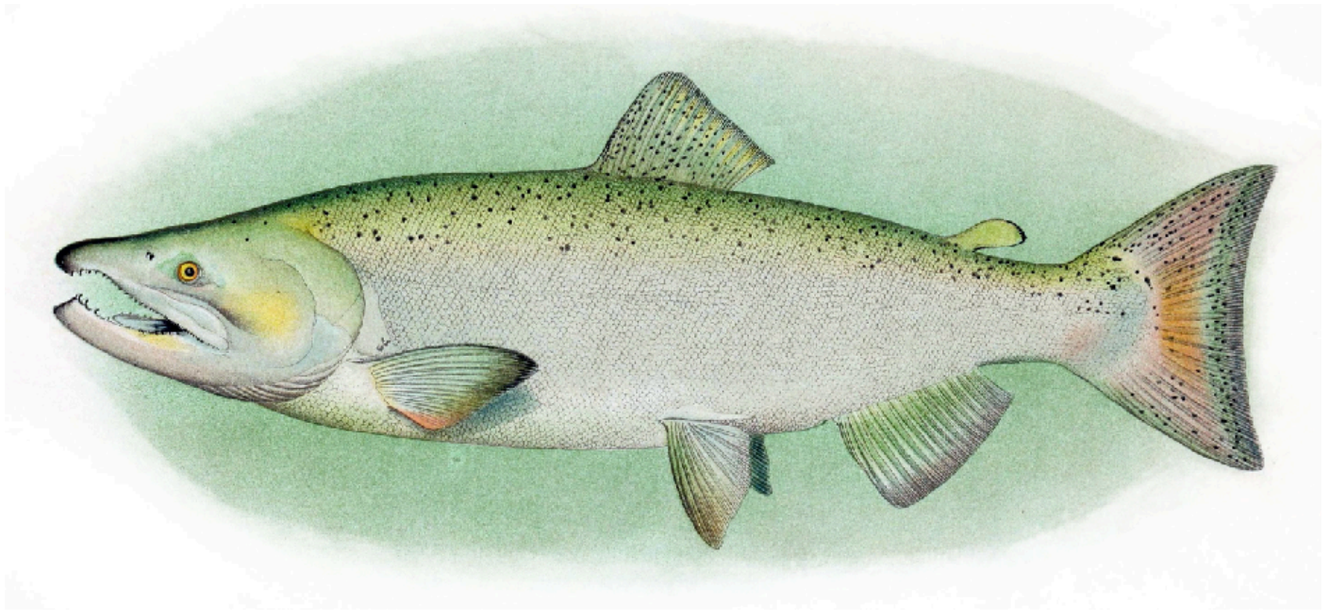
Common file formats

Chinook



Fish pic from Wikipedia

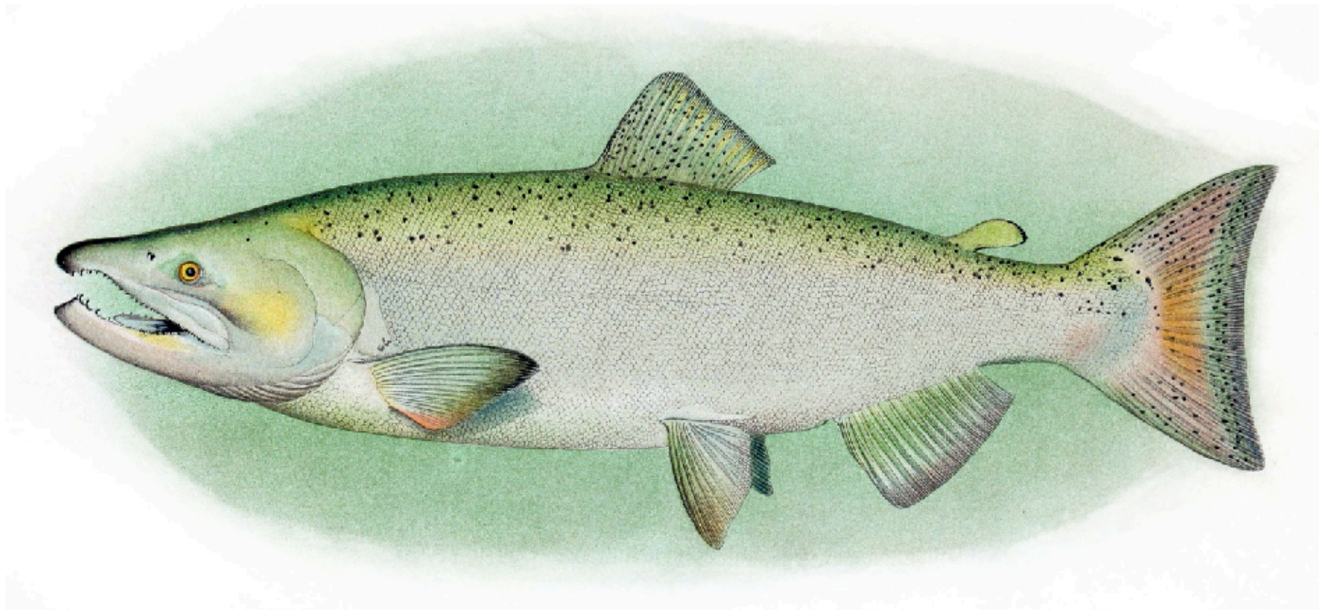
Chinook



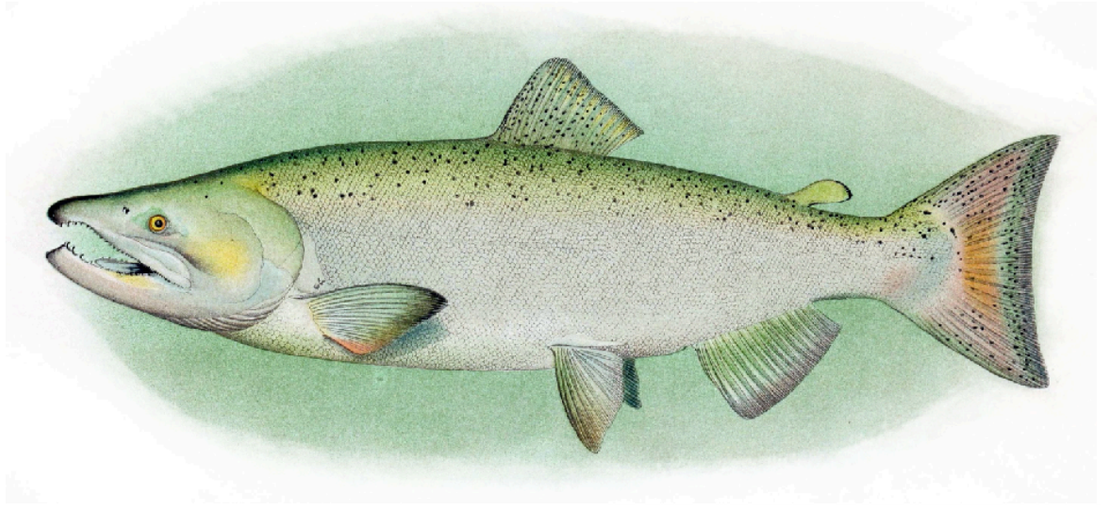
Why use simulated data?

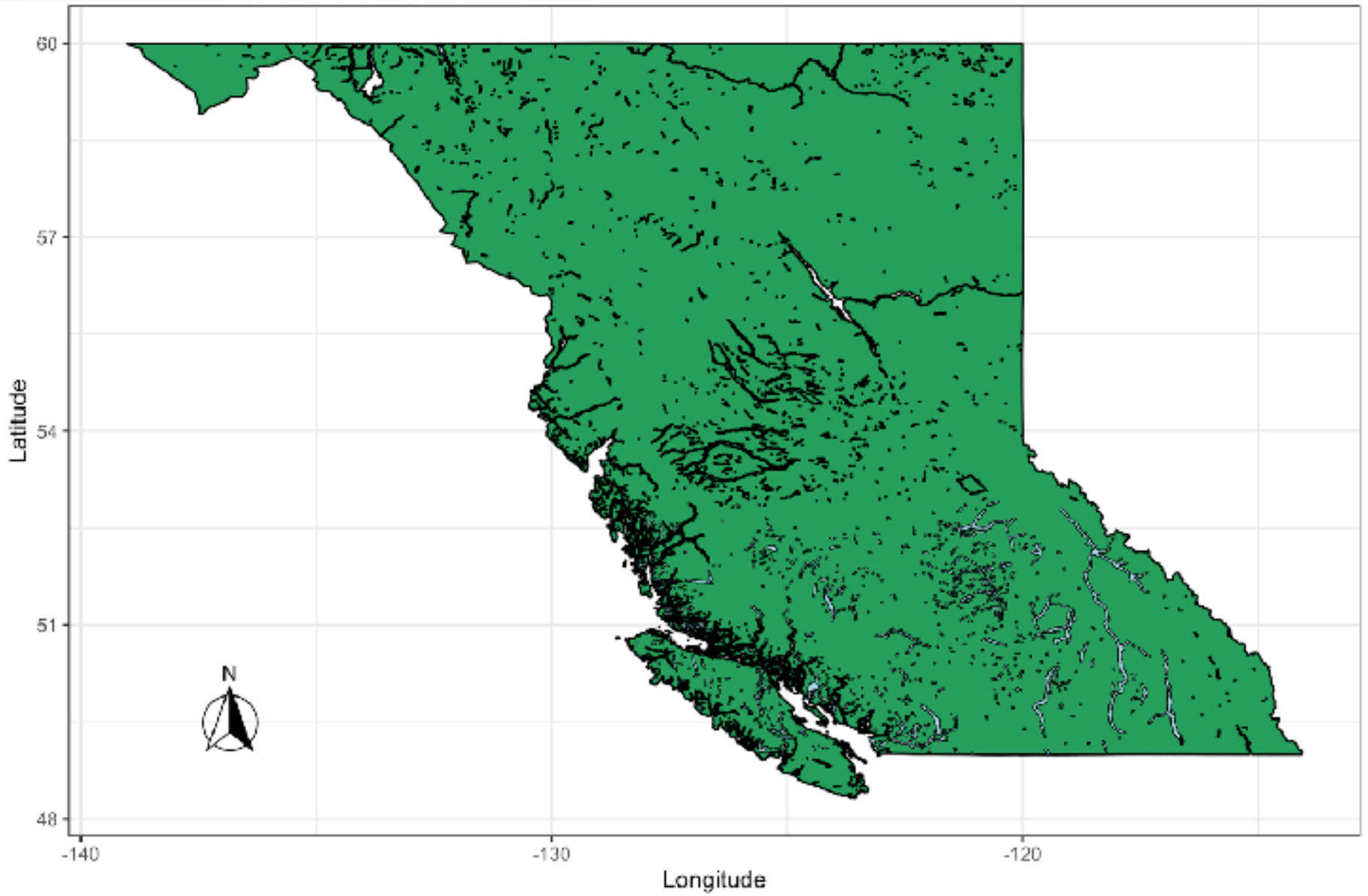
Chinook

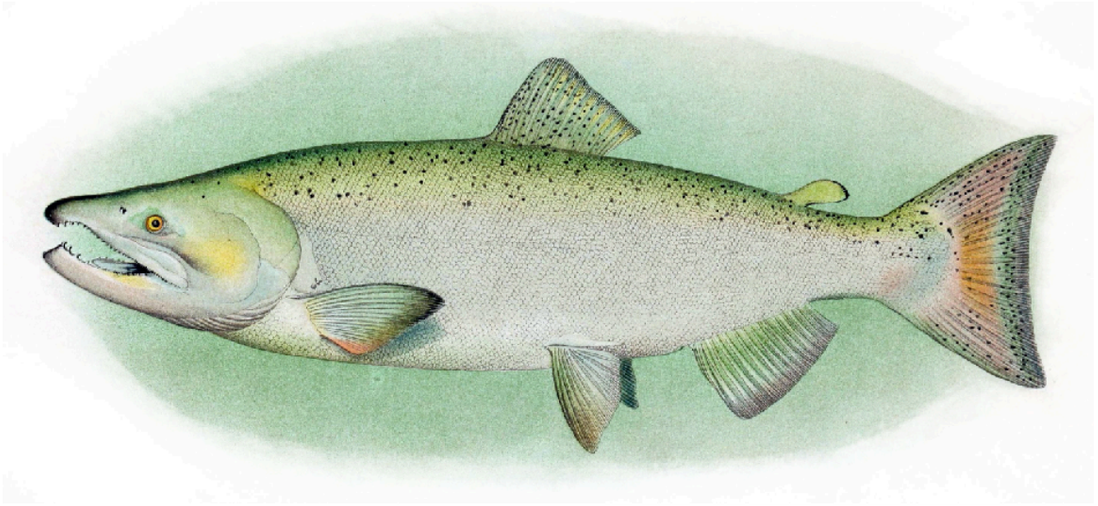
Chinook Salmon
Oncorhynchus tshawytscha

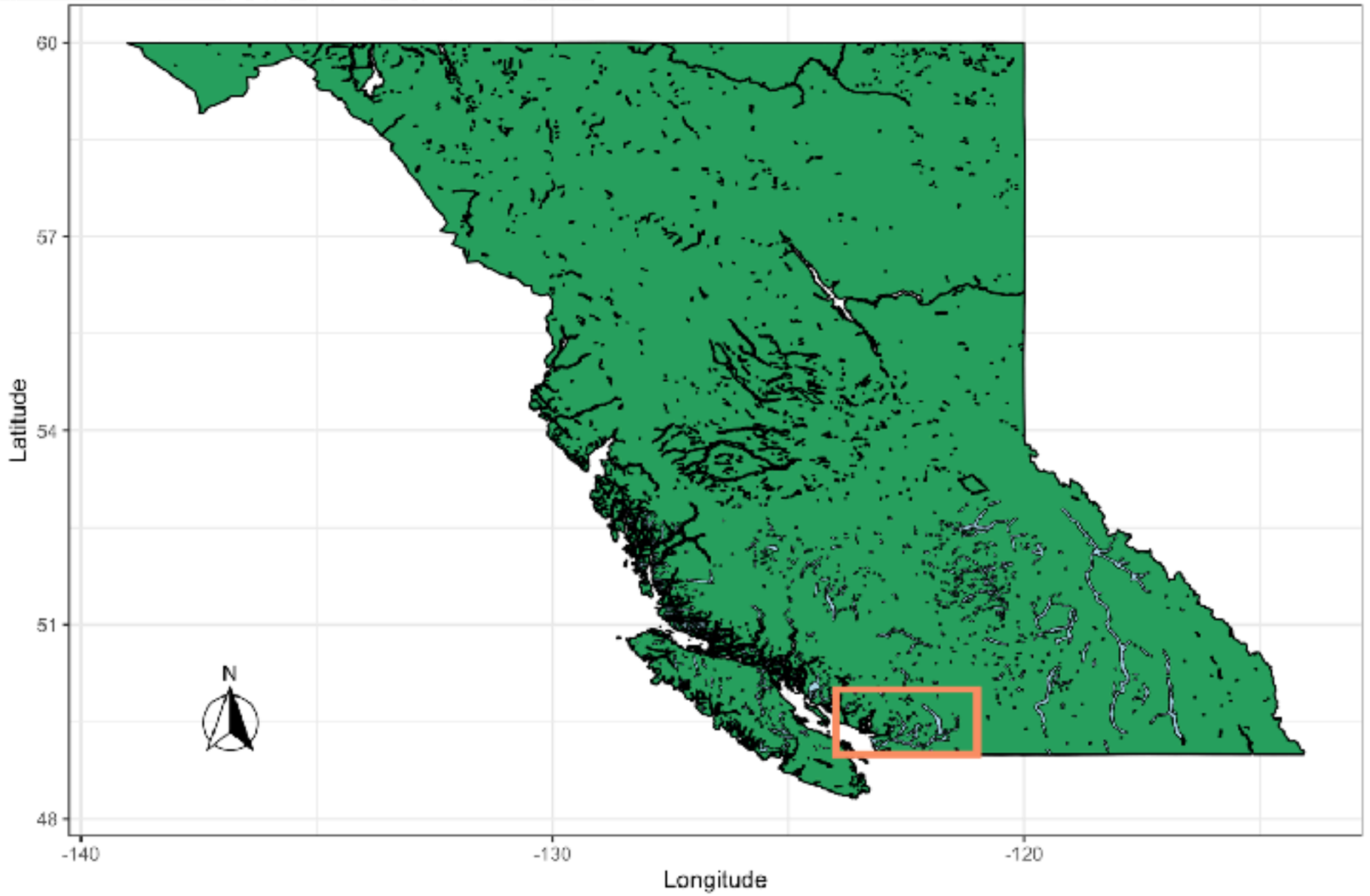


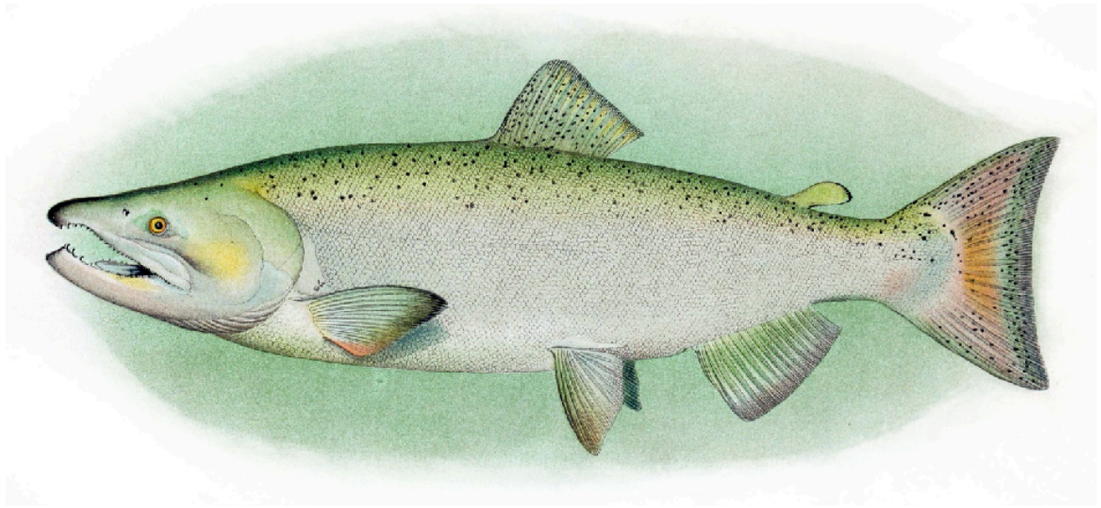
>2.4 Gbp Genome
32 chromosomes

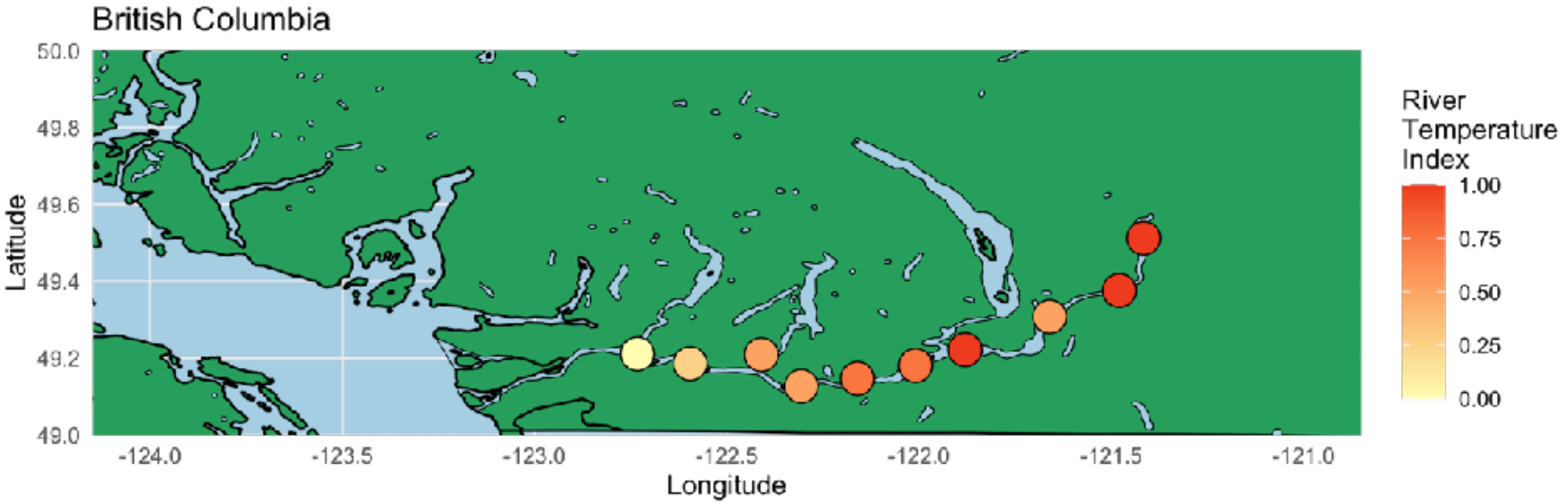


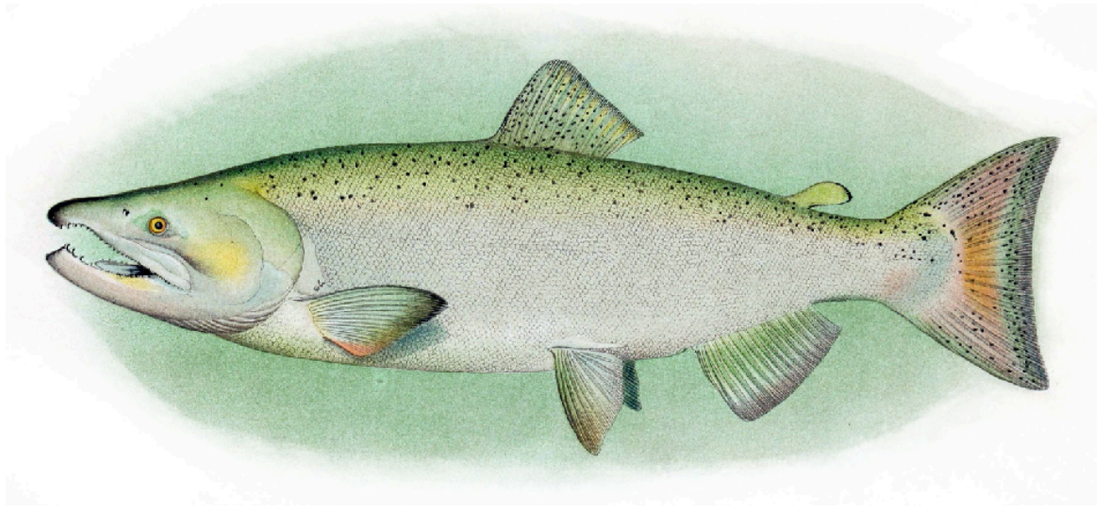

60
57
54
51
48
Latitude
-140
-130
-120
Longitude
N

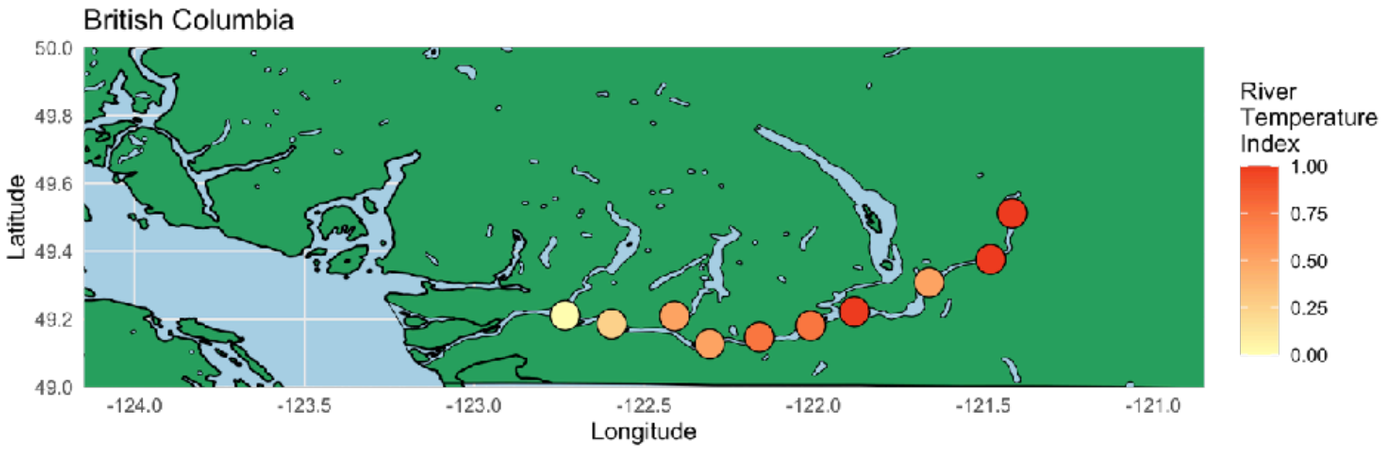



60
57
54
51
48
Latitude
-140
-130
-120
Longitude
N

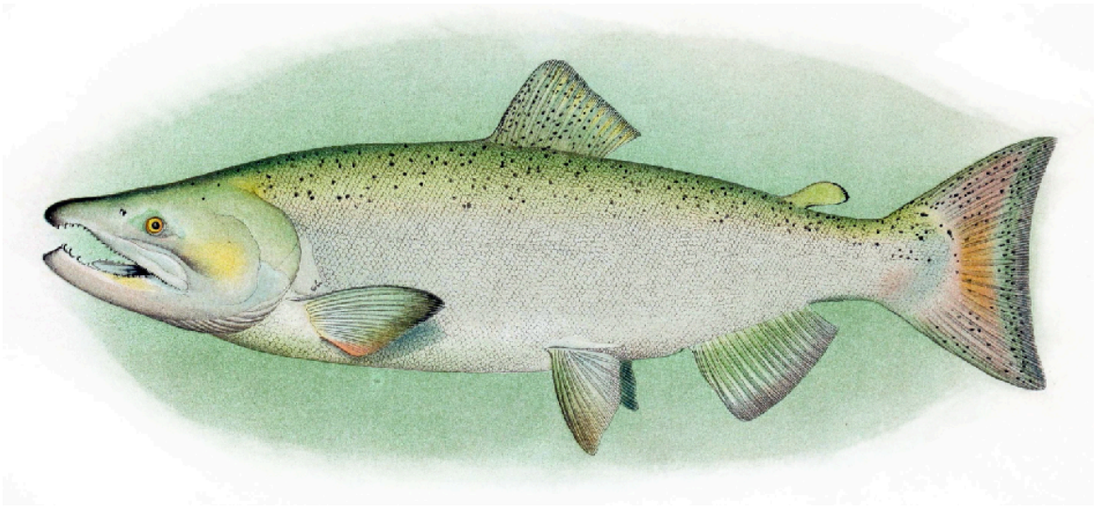


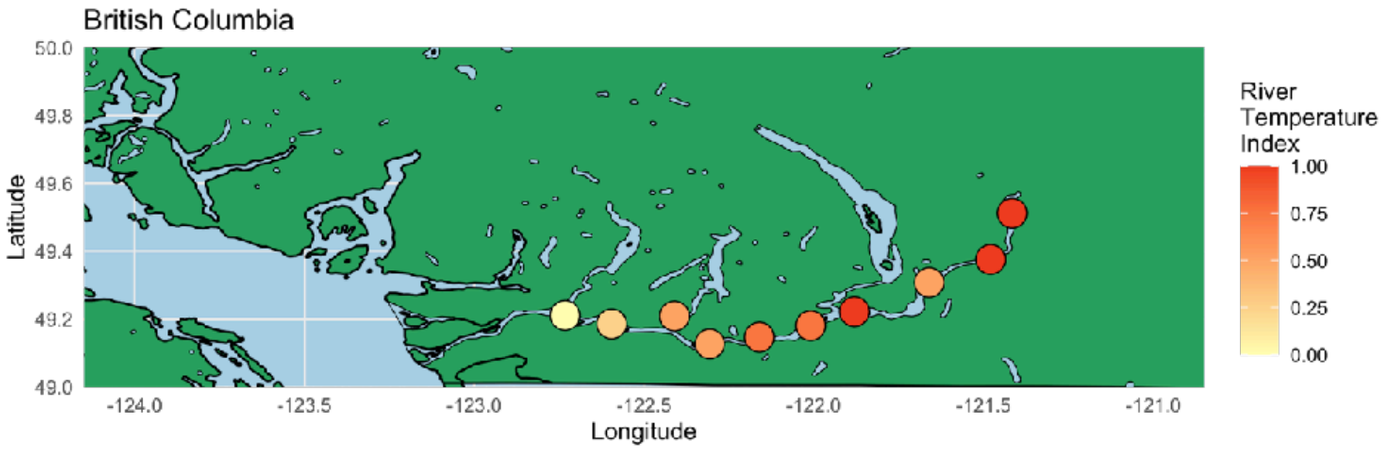

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

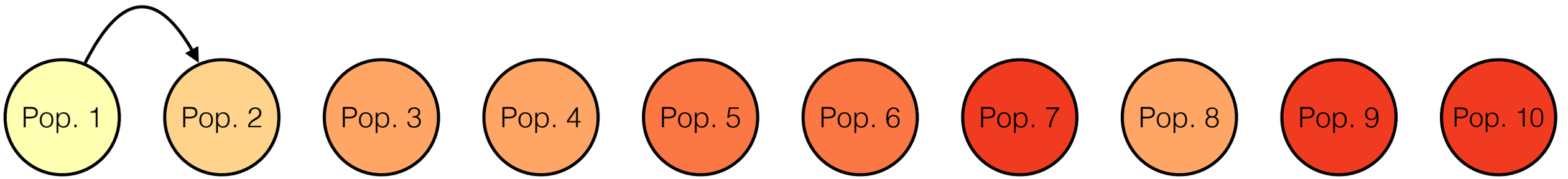


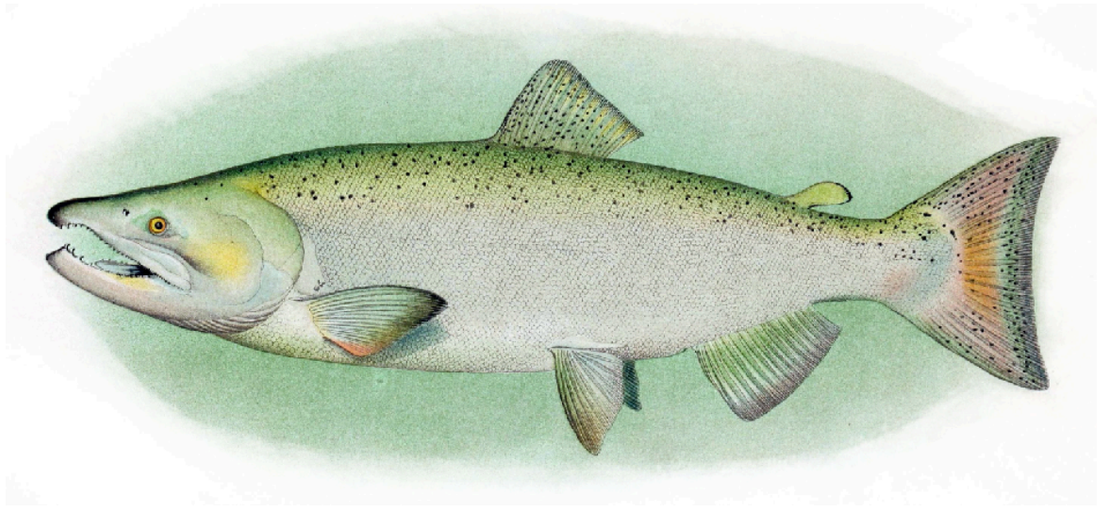


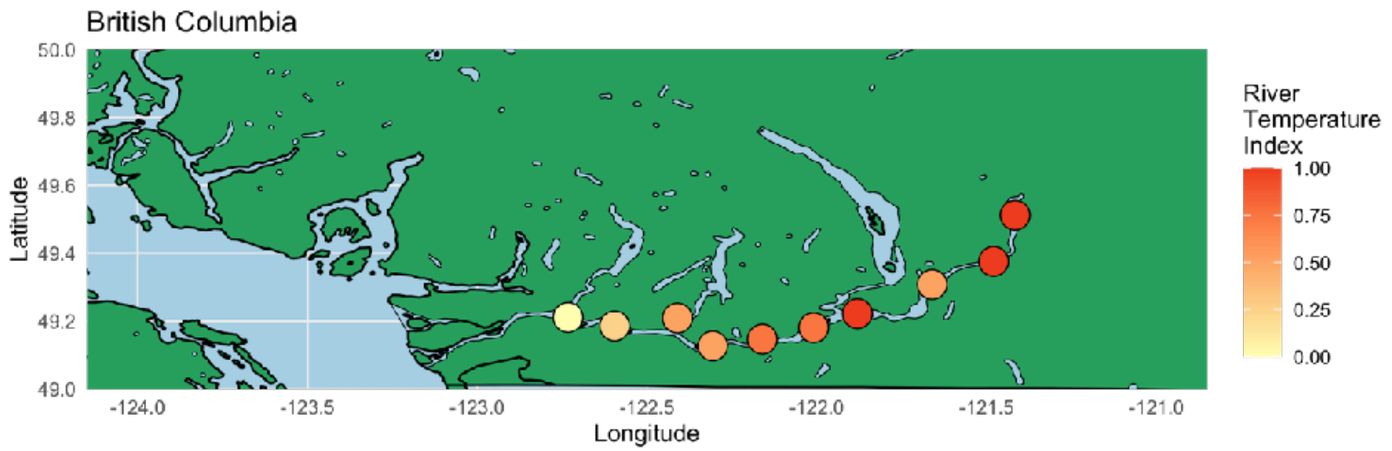
Pop. 1 Pop. 2 Pop. 3 Pop. 4 Pop. 5 Pop. 6 Pop. 7 Pop. 8 Pop. 9 Pop. 10

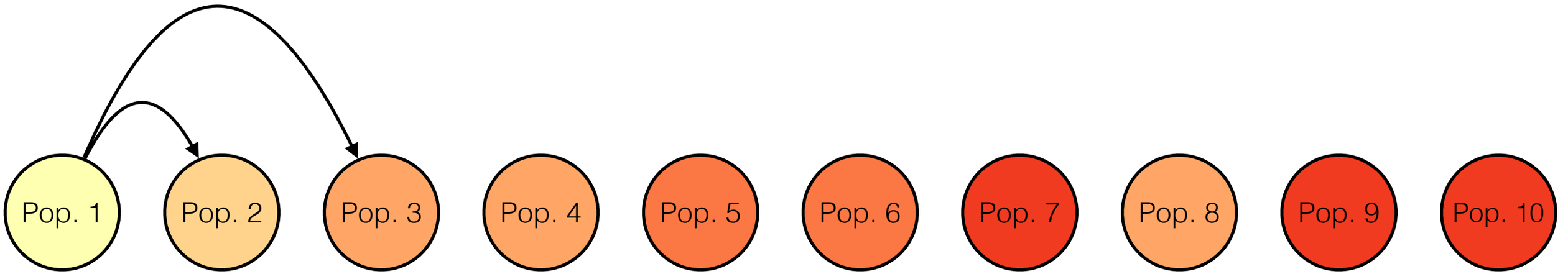


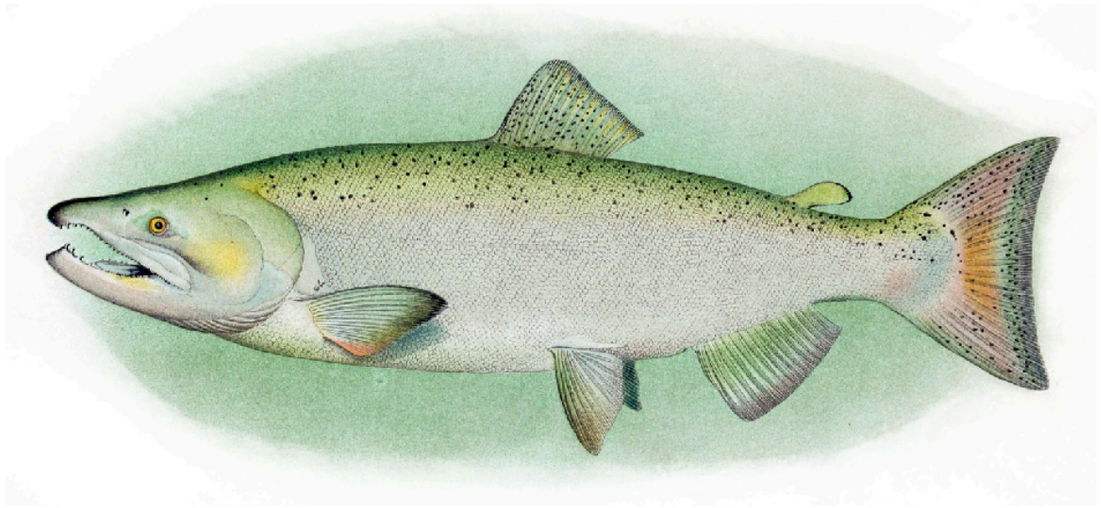

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

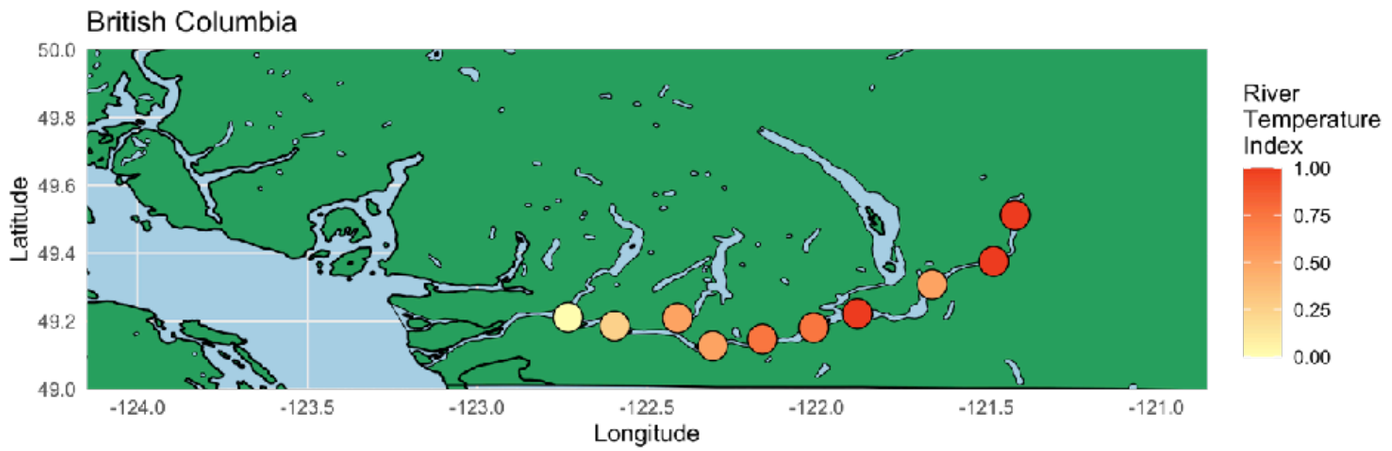

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10

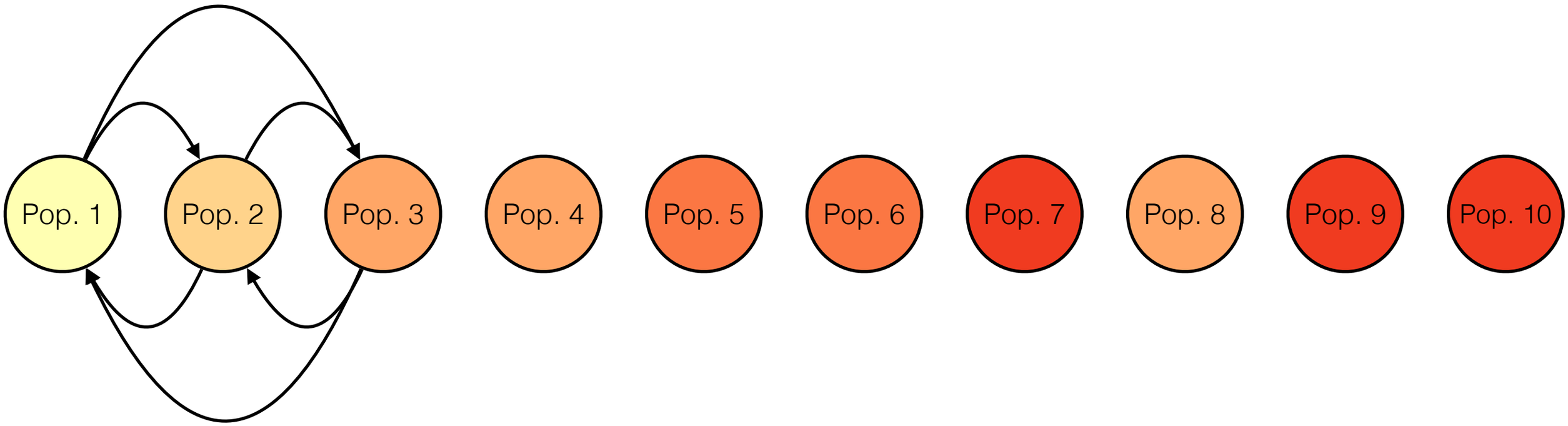


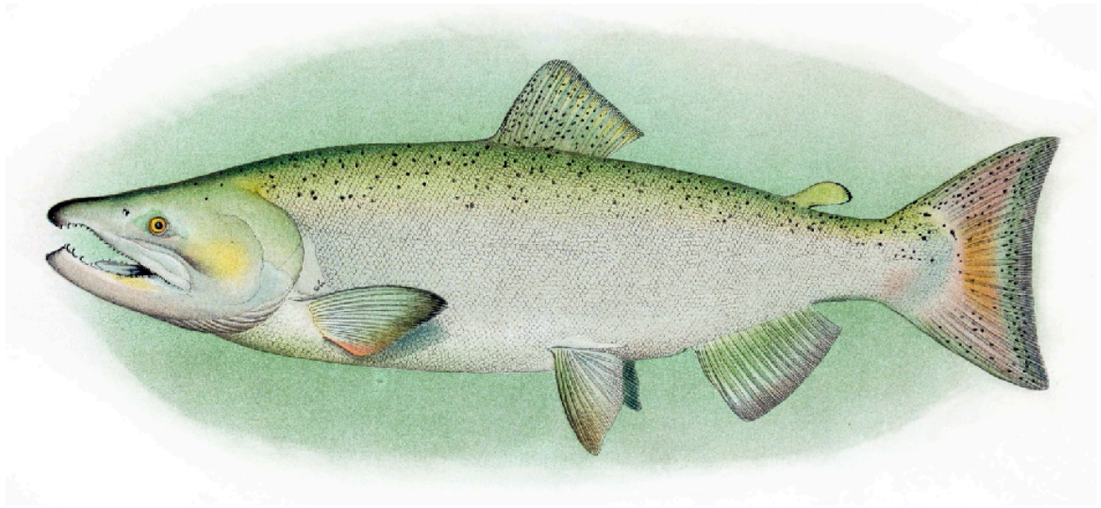

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

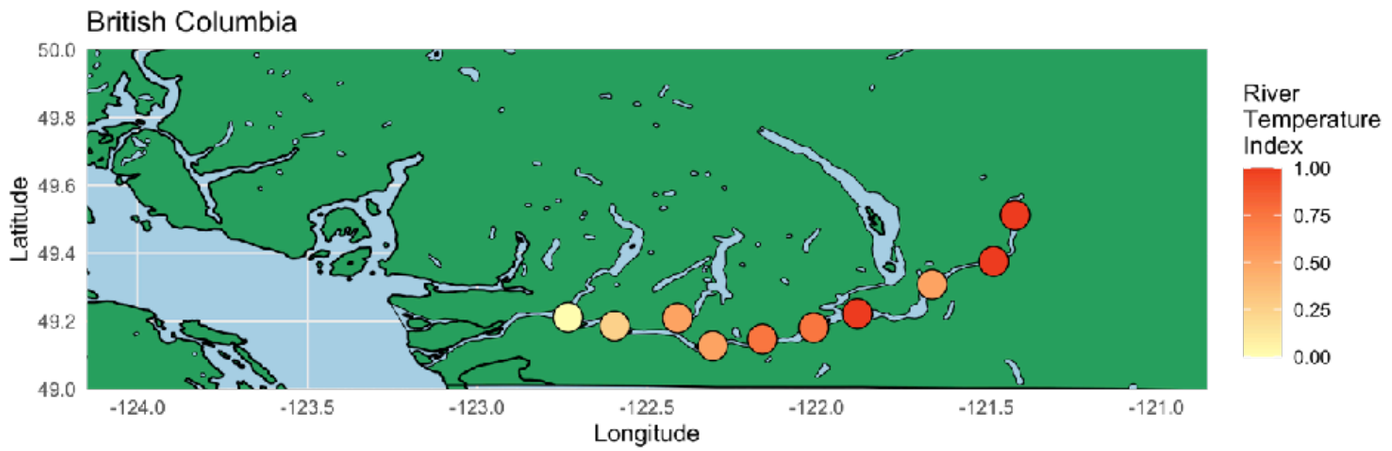

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10

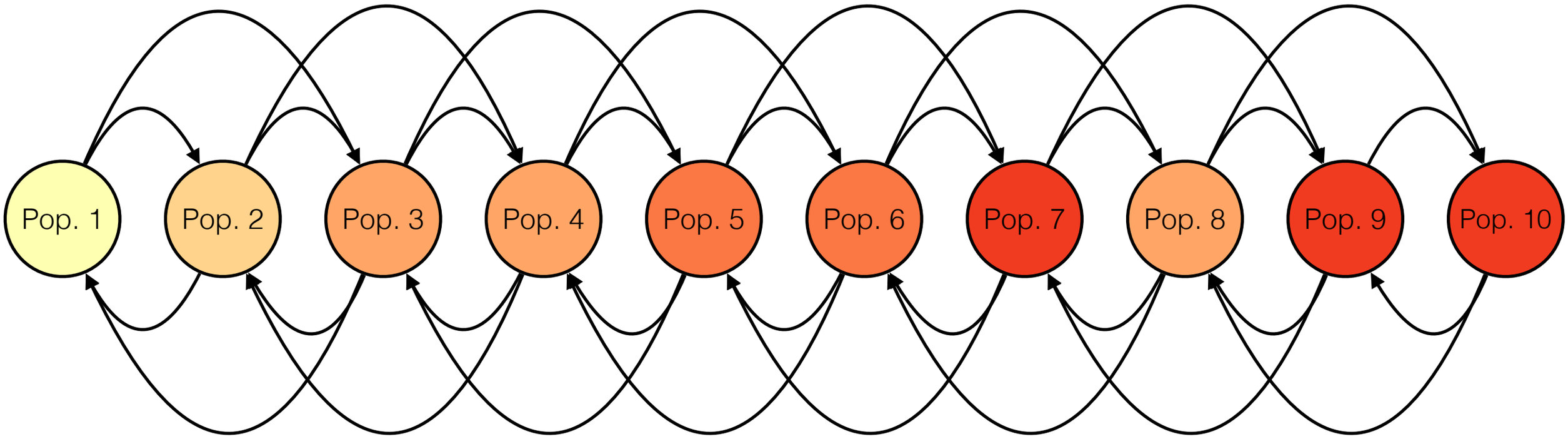


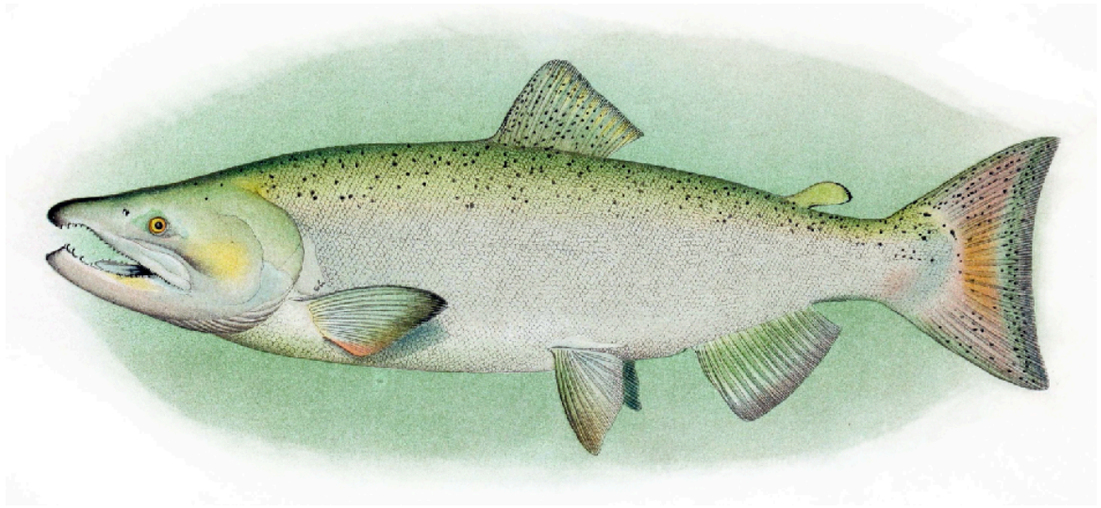

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

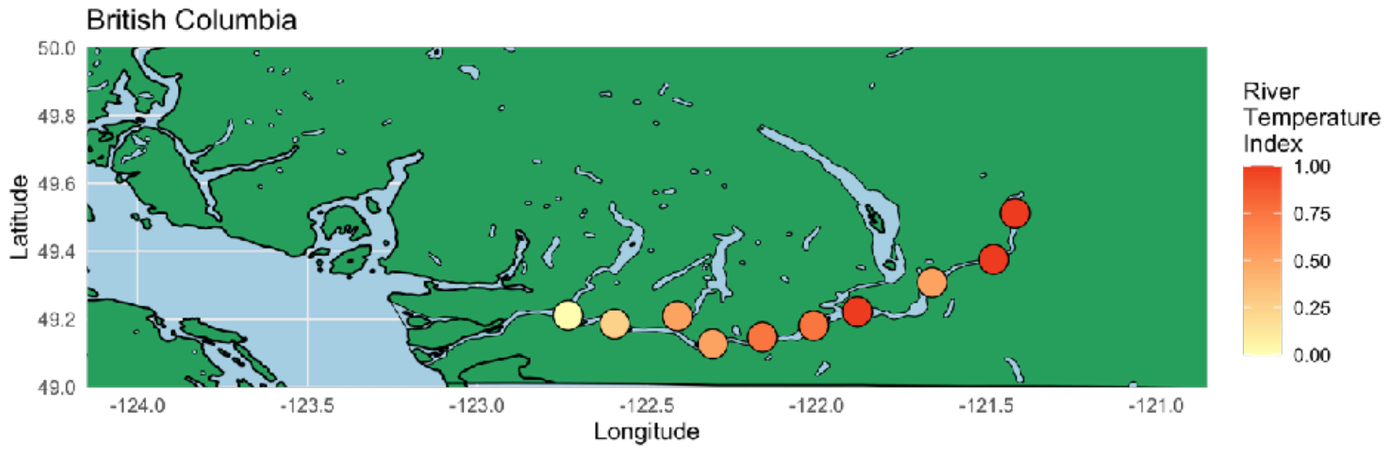

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10

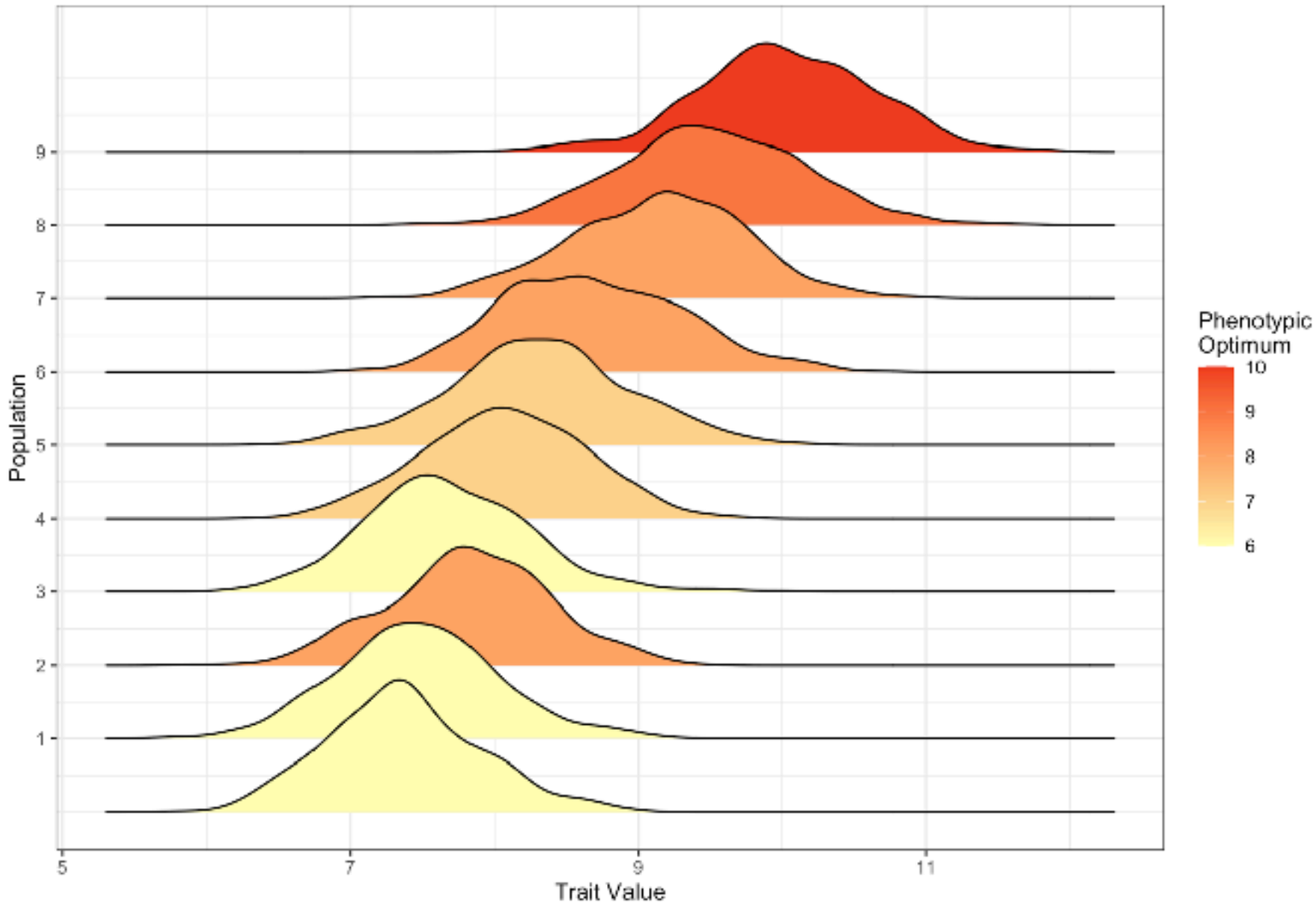


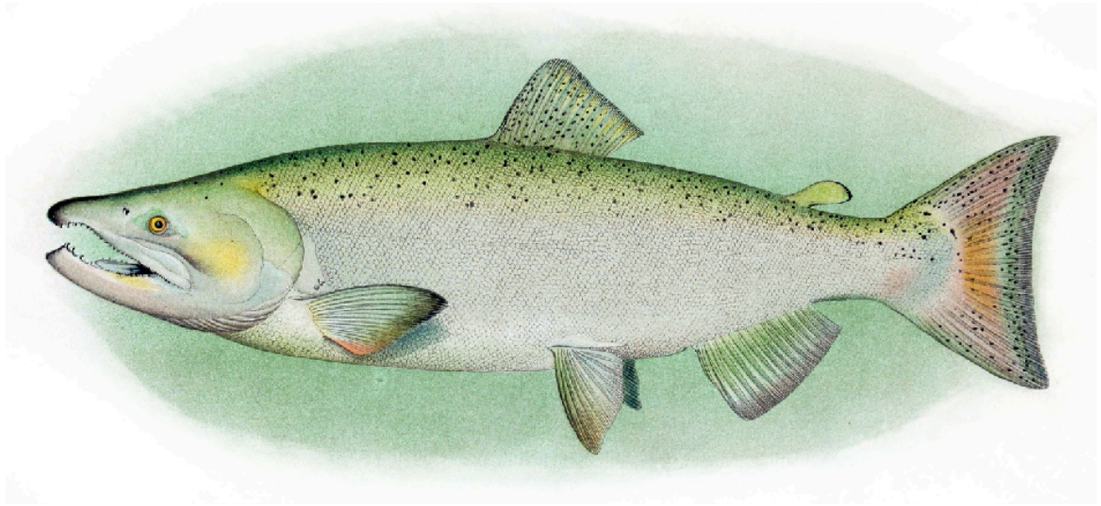

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

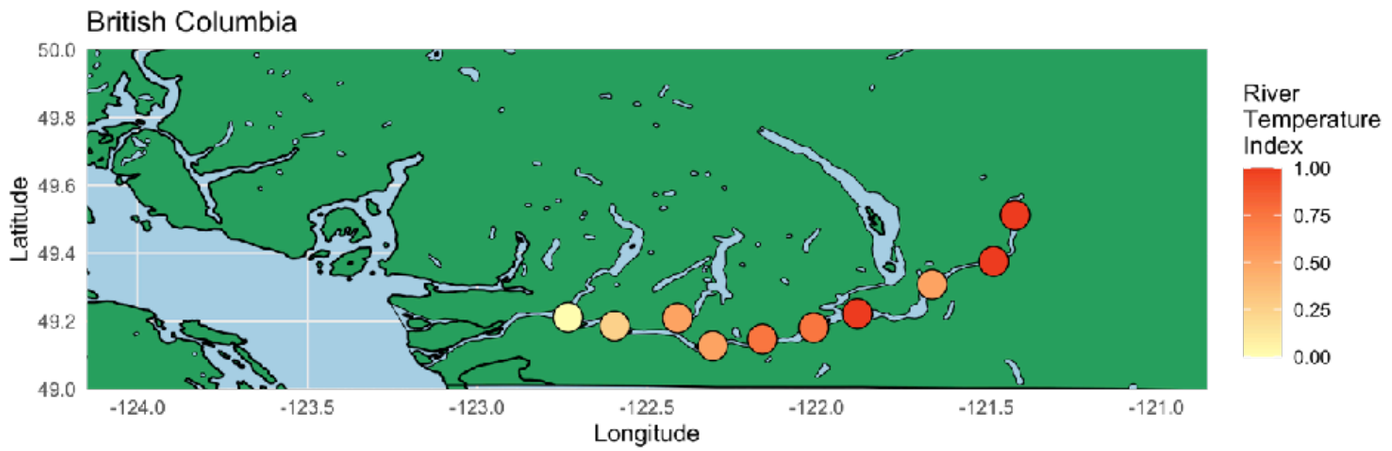

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10




British Columbia
Latitude
50.0
49.8
49.6
49.4
49.2
49.0
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

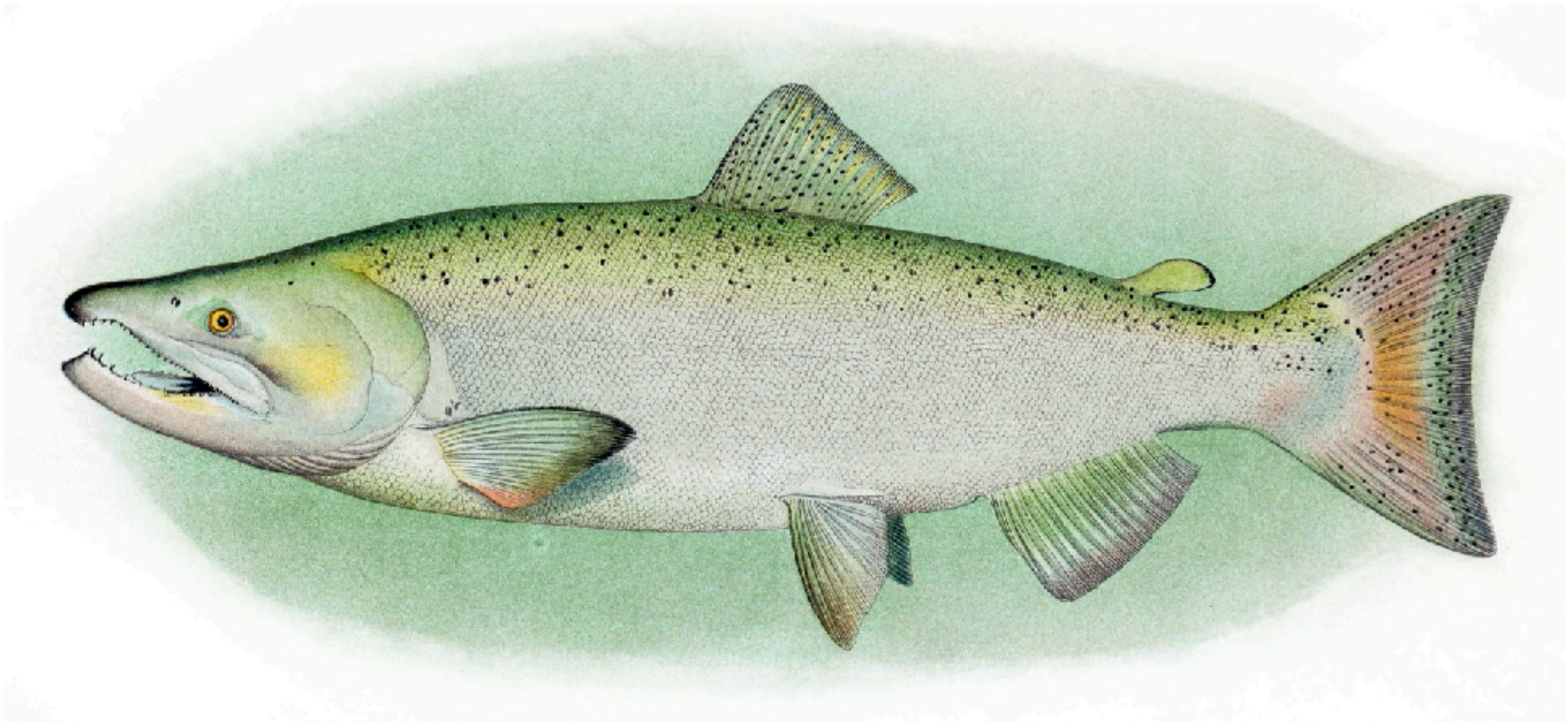

Population
9
8
7
6
5
4
3
2
1
5
7
9
11
Trait Value
Phenotypic Optimum
10
9
8
7
6





This week we will use the simulated data to:

- Assemble a genome (Tuesday afternoon)
- Align sequencing reads (Wednesday morning)
- Quantify gene expression (Wednesday afternoon)
- Call variants (Thursday morning)
- Examine population structure (Friday morning)
- Analyse the genetic basis of a quantitative trait (Friday afternoon)



Questions?

Topic 1 Tutorial

In this first tutorial, we will introduce a number of core concepts in bioinformatics.

By the end of this tutorial we will have done the following:

- Played with a reference genome and the FASTA format
- Explored high-throughput sequence alignments (SAM/BAM files)
- Examined genome annotations stored in the General Feature/ Transfer Format (GFF/GTF) and BED formats
- Assessed genetic variants called from genome sequencing data

A brief history of DNA sequencing

Genome milestones

1977: *Bacteriophage* ΦX174
1982: *Bacteriophage lambda*
1995: *Haemophilus influenzae*
1996: *Saccharomyces cerevisiae*
1998: *Caenorhabditis elegans*
2000: *Drosophila melanogaster*
2000: *Arabidopsis thaliana*
2001: *Homo sapiens*
2002: *Mus musculus*
2004: *Rattus norvegicus*
2005: *Pan troglodytes*
2005: *Oryza sativa*
2007: *Cyanidioschyzon merolae*
2009: *Zea mays*
2010: Neanderthal
2012: Denisovan
2013: The HeLa cell line
2013: *Danio rerio*
2017: *Xenopus laevis*

Excerpted and edited from Box 1 and 2 - Shendure et al 2017 Nature

A brief history of DNA sequencing

Technological milestones

1953: Sequencing of insulin protein

1965: Sequencing of alanine tRNA

1968: Sequencing of cohesive ends of phage lambda DNA

1977: Maxam–Gilbert sequencing

1977: Sanger sequencing

1990: Paired-end sequencing

2000: Massively parallel signature sequencing by ligation

2003: Single-molecule massively parallel sequencing-by-synthesis

2003: Zero-mode waveguides for single-molecule analysis

2003: Sequencing by synthesis of in vitro DNA colonies in gels

2005: Four-colour reversible terminators

2005: Sequencing by ligation of in vitro DNA colonies on beads

2007: Large-scale targeted sequence capture

2010: Direct detection of DNA methylation during single-molecule sequencing

2010: Single-base resolution electron tunnelling through a solid state detector

2011: Semiconductor sequencing by proton detection

2012: Reduction to practice of nanopore sequencing

2012: Single-stranded library preparation method for ancient DNA

Excerpted and edited from Box 1 and 2 - Shendure et al 2017 Nature

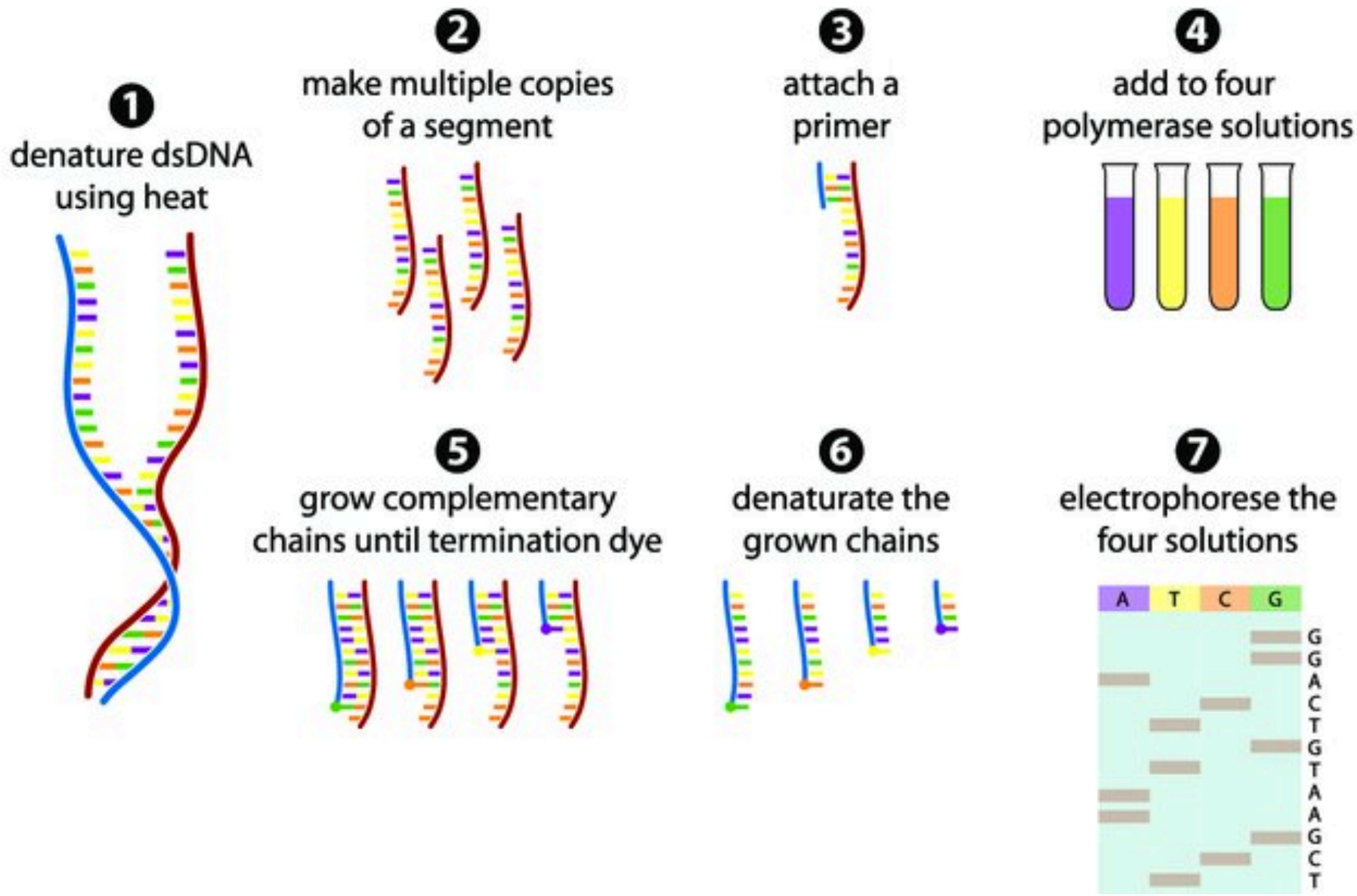
First Generation Sequencing

Maxam-Gilbert: Chemical modification and cleavage followed by gel electrophoresis

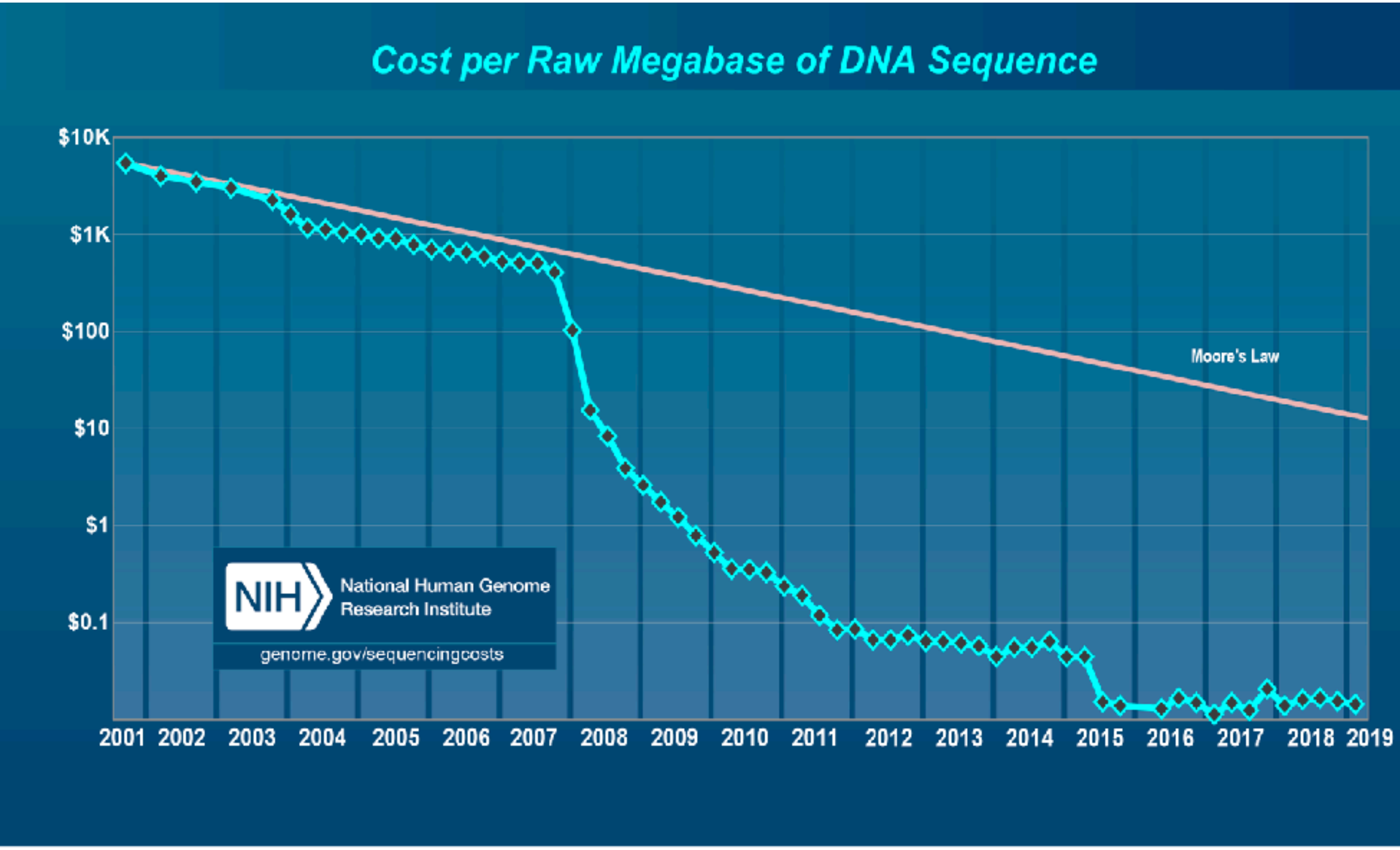
Sanger: Selective incorporation of chain-terminating dideoxynucleotides followed by gel electrophoresis

- Became full automated using flourescently labeled dideoxy bases
- Dominant sequencer up until 2007
- Only one fragment sequenced per reaction
- Still used for sequencing individual PCR products

Sanger sequencing



[26] From PhD thesis of Michel G. Gauthier



*Moore's law stated that the number of transistors on a microchip doubled every two years, while costs halved

Second (Next-gen) and third generation sequencing

Sequences many molecules in parallel

Don’t need to know anything about the sequence to start

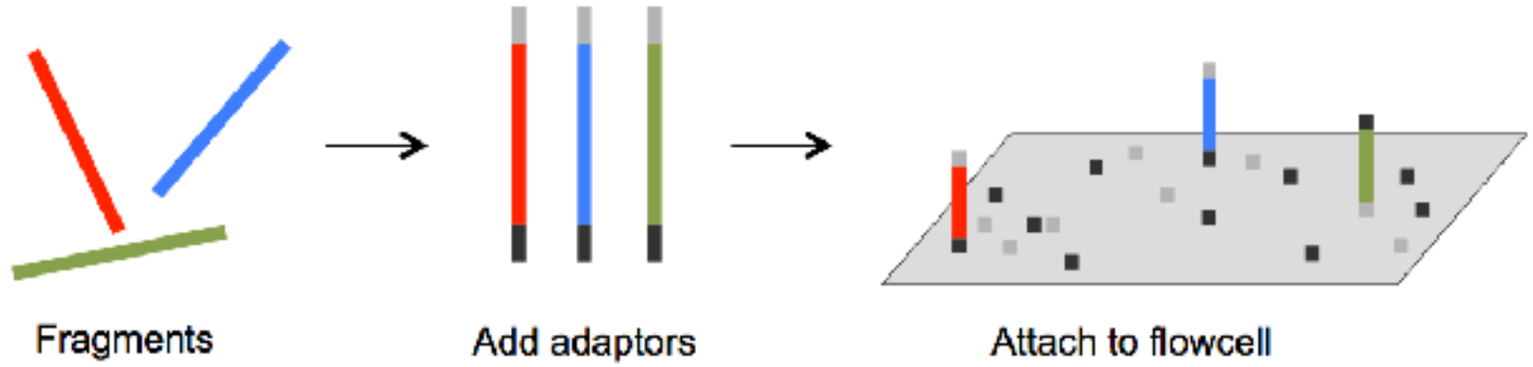
Main technologies:

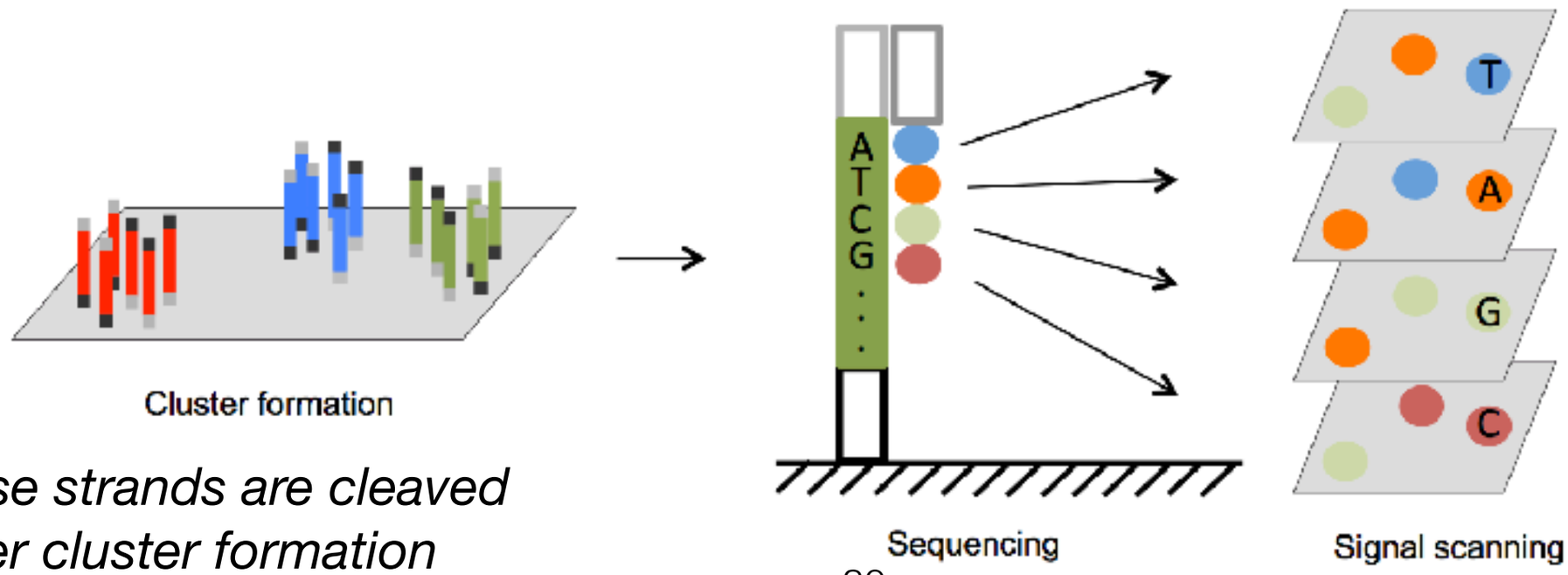
- Illumina
- Ion torrent
- 454 (Pyrosequencing)
- PacBio

Second generation sequencing

Technology	Read Length	Accuracy	Bases/run	Uses
Illumina	50-600bp	99.9%	500-600 GBase	Resequencing General depth
Oxford Nanopore	5kb-100kb	85-95%	10-30GBase	Microbial genomes Genome assembly
PacBio	10kb-40kb	85-90%	5-10Gbase	Genome assembly Structural variants

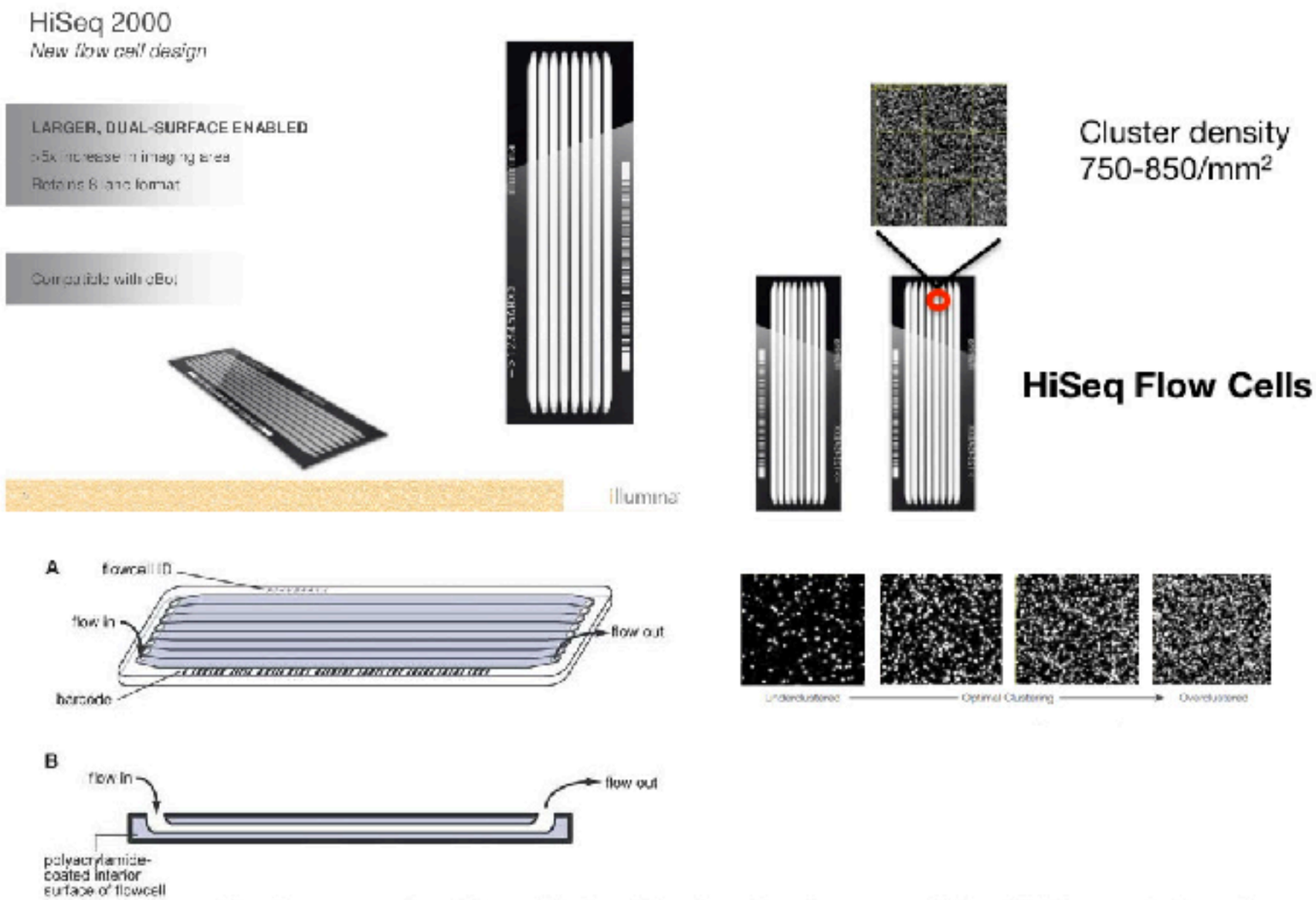
Illumina sequencing





Reverse strands are cleaved after cluster formation

Illumina sequencing



From hackteria.org
https://www.hackteria.org/wiki/File:FlowCell.jpg

Open up the website to the Topic 1 tutorial

Part 1: Setting up the tutorial

Install IGV

- Follow the links on the Topic 1 page to install the IGV from the Broad institute if you haven't done so already

Part 1: Setting up the tutorial

Download data for the tutorial

- Follow the links in the tutorial to obtain the data for this practical session

- Make sure you have all the items listed on the Topic 1 page

- ***OPTIONAL***: Check data integrity

Part 2: Reference genomes

A reference genome is a representation of the average genome for a species/population

Part 2: Reference genomes

A reference genome is a representation of the average genome for a species/population

Typically stored as a FASTA file (pronounced like pasta, or fast-a if you're fancy)

Part 2: Reference genomes

A reference genome is a representation of the average genome for a species/population

Typically stored as a FASTA file (pronounced like pasta, or fast-a if you're fancy)

What a FASTA file looks like:

```
>chr_1
TGGGCAAGGCTGATGAACAGCAGCTGCATAAATTCTCCCCTAATTATATTGTAAATAGCT
GCAGCACAACAATAAAGCTTTGTTAGAGACATCTAGAGAATCACACACTGCATCTGTTCT
GCCGCTCTCCCTCTTGCTCTGTTCTGAGAAGCACTTGTTCACTGATTCTGGGTTTGTATT
TGTGTTTTTCATGCTTAACATTGTTATTTGTTTGCCTAGAAAGTTCTTTGATTGGGCCAA
ATTAGTCGATTTTAAAGAGTGCACTTCTCTAGTGCATGTAATCTATGTGGACATCTCAAT
AGCTGCTTAATTTGTTTAGTGGTAATCTCCTCTGAACAGAGAGAAAGGCCTACATGCAGC
CCTCAGAGGAGAGGTGTCAATCTCTCTTTGATTATCTCTTTGTTTCCCTTCAGAAGAATC
ATTCTAATCTGGTATTGTACAAGAGGAAATAAATGGGACTAAAACCAGGCATGCACCATC
TGATAGATTCACATCCCTAGAAGACTTTTGTTGTGTTTGTTTCAAGTGGAGAGCCTGCTG
```

FASTAs are plain text files

Part 2: Reference genomes

A reference genome is a representation of the average genome for a species/population

Typically stored as a FASTA file (pronounced like pasta, or fast-a if you're fancy)

What a FASTA file looks like:

Sequence name

```
>chr_1
TGGGCAAGGCTGATGAACAGCAGCTGCATAAATTCTCCCCTAATTATATTGTAAATAGCT
GCAGCACAACAATAAAGCTTTGTTAGAGACATCTAGAGAATCACACACTGCATCTGTTCT
GCCGCTCTCCCTCTTGCTCTGTTCTGAGAAGCACTTGTTCACTGATTCTGGGTTTGTATT
TGTGTTTTTCATGCTTAACATTGTTATTTGTTTGCCTAGAAAGTTCTTTGATTGGGCCAA
ATTAGTCGATTTTAAAGAGTGCACTTCTCTAGTGCATGTAATCTATGTGGACATCTCAAT
AGCTGCTTAATTTGTTTAGTGGTAATCTCCTCTGAACAGAGAGAAAGGCCTACATGCAGC
CCTCAGAGGAGAGGTGTCAATCTCTCTTTGATTATCTCTTTGTTTCCCTTCAGAAGAATC
ATTCTAATCTGGTATTGTACAAGAGGAAATAAATGGGACTAAAACCAGGCATGCACCATC
TGATAGATTCACATCCCTAGAAGACTTTTGTTGTGTTTGTTTCAAGTGGAGAGCCTGCTG
```

Nucleotide sequence

FASTAs are plain text files

Part 2: Reference genomes

Load the reference genome into IGV and explore it a bit

Part 2: Reference genomes

Load the reference genome into IGV and explore it a bit

1. How many chromosomes do our Salmon have?

Part 2: Reference genomes

Load the reference genome into IGV and explore it a bit

1. How many chromosomes do our Salmon have?

2. What is the length of each chromosome?

Part 2: Reference genomes

Load the reference genome into IGV and explore it a bit

1. How many chromosomes do our Salmon have?

2. What is the length of each chromosome?

3. What is the nucleotide sequence corresponding to chr_1:666-670?

Flavours of DNA sequencing

- Whole Genome Sequencing
- Pool Seq
- RNAseq
- Amplicon Sequencing (GT-seq)
- Sequence Capture
- Reduced-Representation Sequencing (RADseq/GBS/ RADcapture)

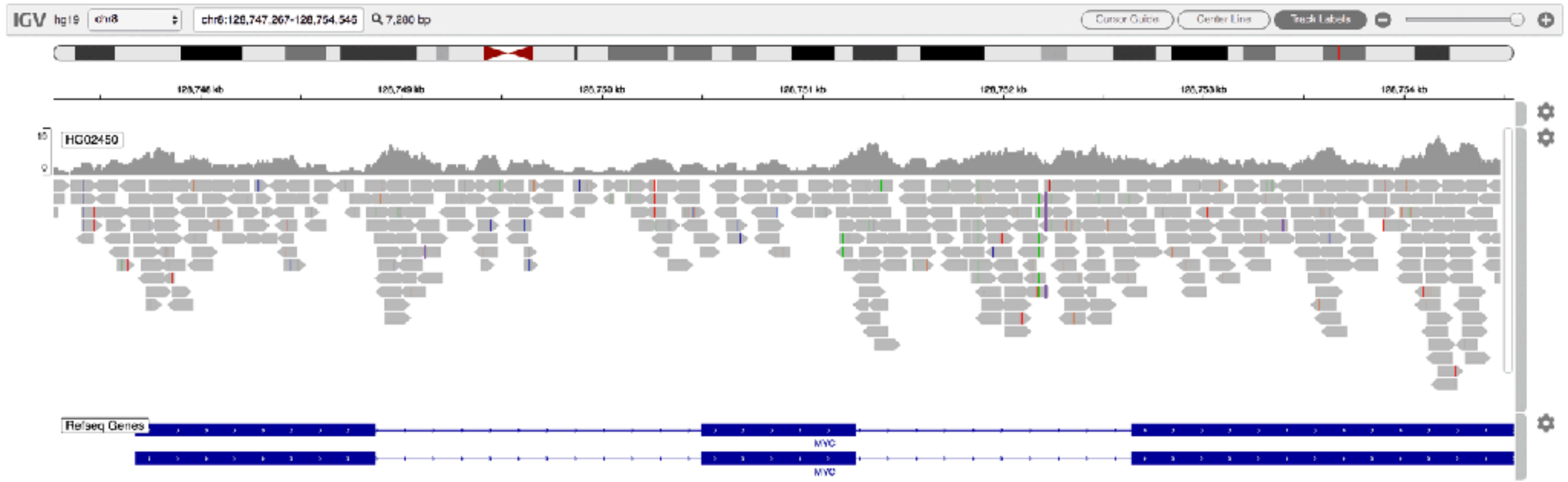
Different approaches have various pros and cons

This is not an exhaustive list, but based on what people in the BRC do

Whole Genome Sequencing

Randomly sheer DNA and sequence all fragments

May use double-stranded nuclease treatment to reduce repetitive elements

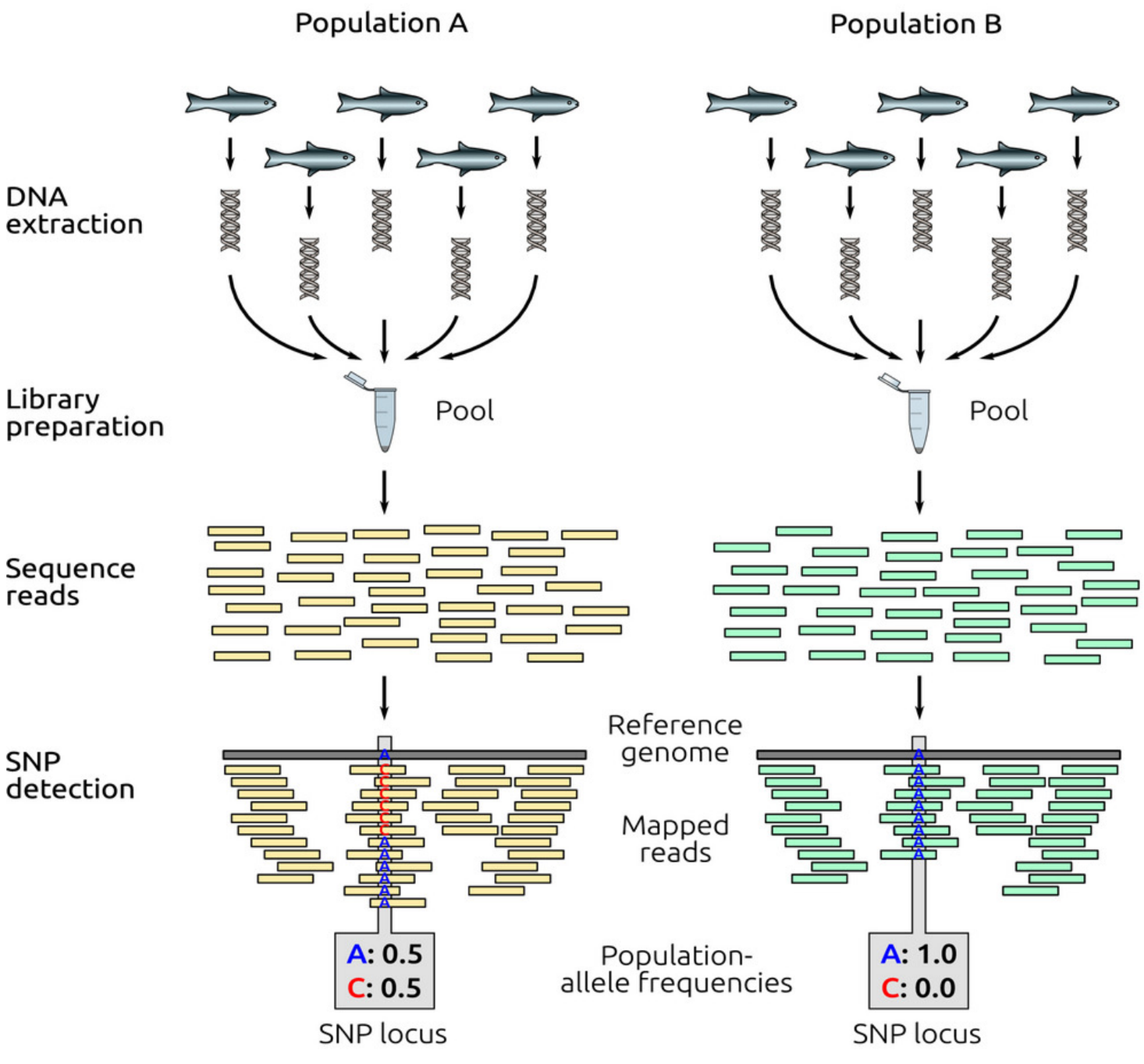


Screen shot from the Integrated Genomics Viewer

Whole Genome Sequencing

Pros	Cons
All sites possible	Comparatively expensive per sample
Simple library prep	Storage and bioinformatics challenging with lots of samples

Pool Seq

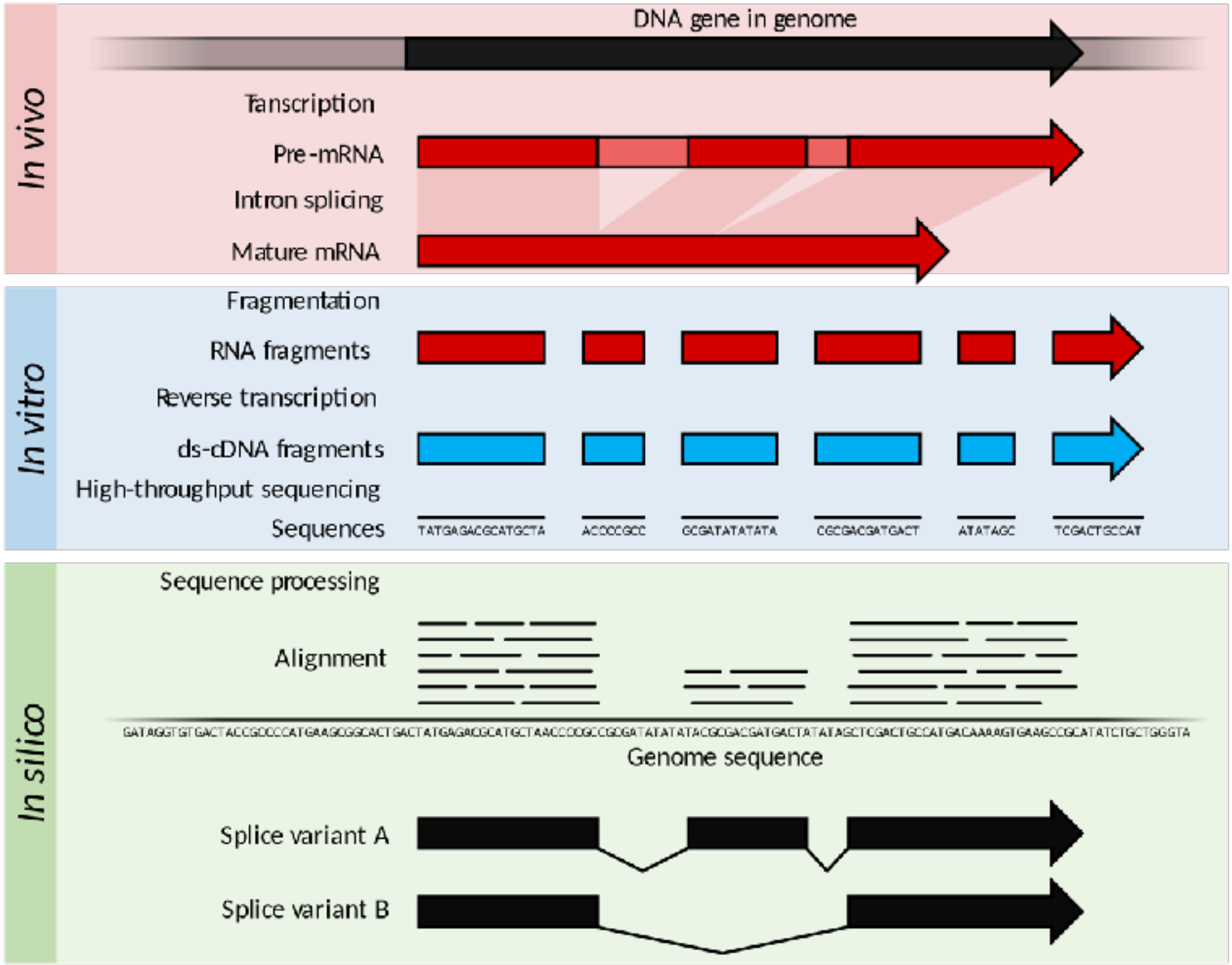


Adapted from Fuentes-Pardo & Ruzzante 2017 Mol. Ecol

Pool Seq

Pros	Cons
All sites possible	Limited analysis options
Simple library prep	No haplotype information
Cheaper than individual WGS	Best in cases where # samples > # reads

RNAseq



From Wikipedia

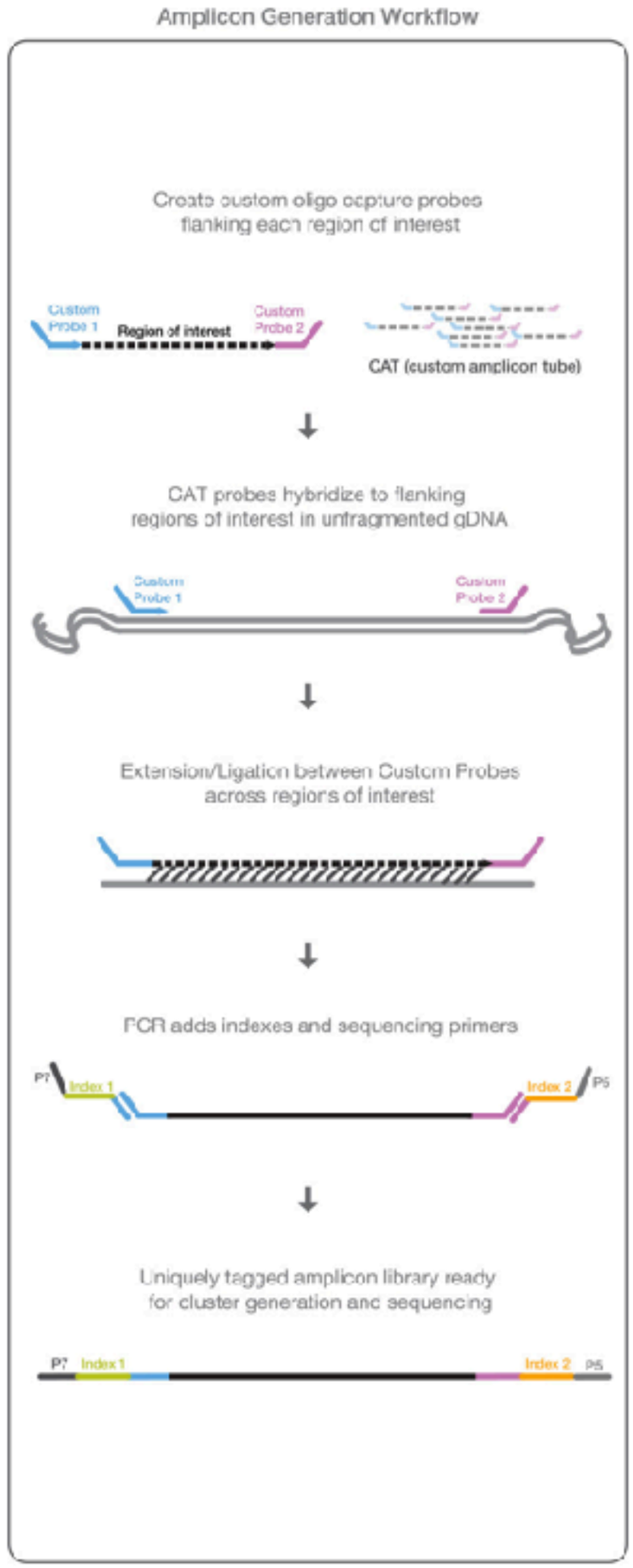
RNAseq

Pros	Cons
Many sites and only in genes	Expression differences complicate SNP calling
Also get expression information	Expensive for pop gen level sampling
Relatively easy to assemble	Difficult library prep (or so I'm told!)

Amplicon Sequencing

- Use PCR to amplify target DNA. Sequence many barcoded samples in one lane.

- Used to characterise microbiome by sequencing 16s rRNA



Amplicon Sequencing

Pros	Cons
Get incredible depth at single locus	Limited to one or few loci
Simple bioinformatics.	Mutations in primer site don't sequence

GT-seq

- Genotyping by Thousands
- Based on Amplicon sequencing
- Multiplex PCR amplify ~200 known SNPs and then sequence pooled PCR products.
- Very cheap ($1/sample), and bioinformatically simple.
- Useful for genotyping thousands or tens of thousands of samples.
- Complicated initial set-up.

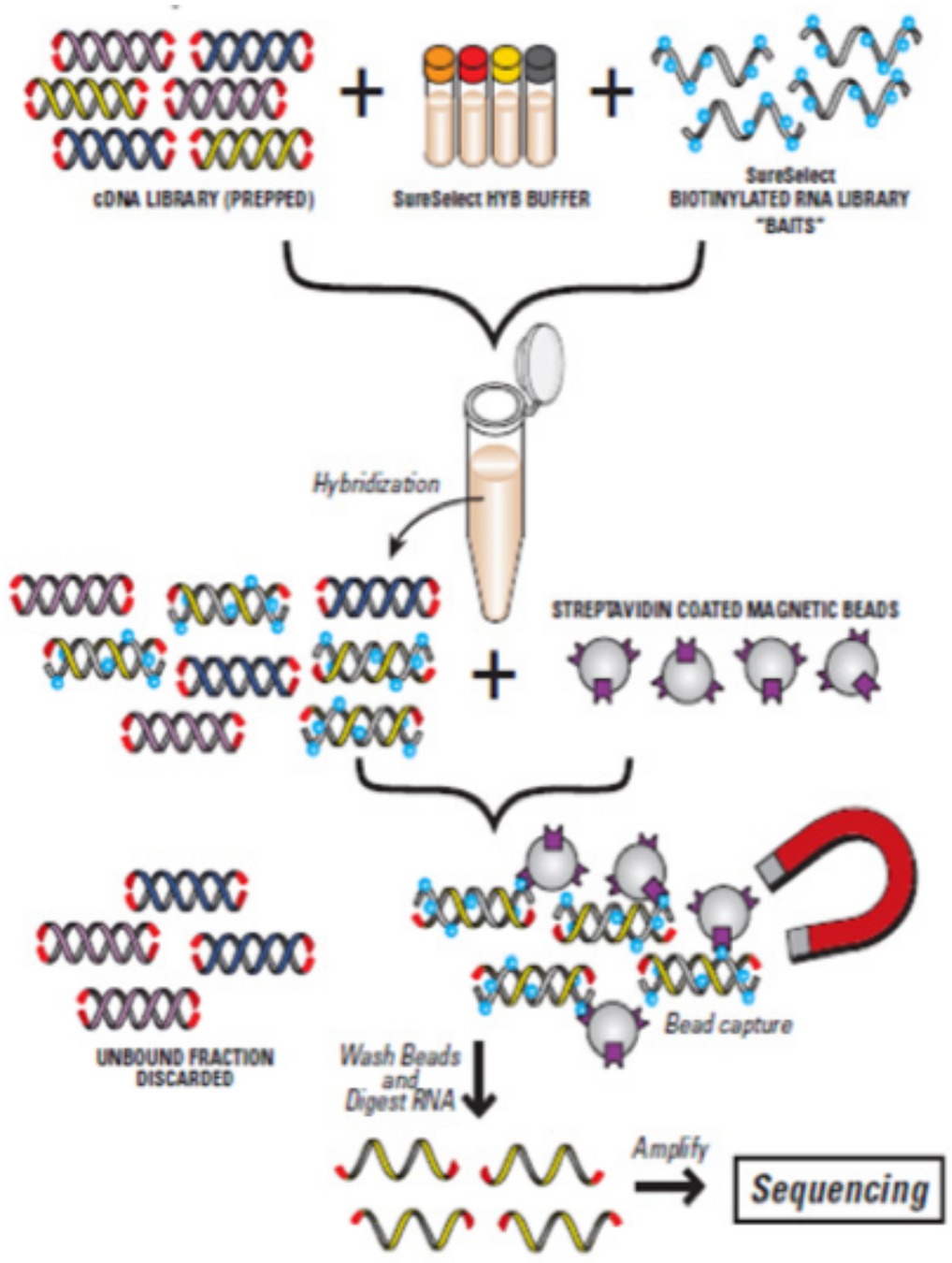
Campbell et al. 2015 *Mol Ecol Resour*.

Sequence Capture

- Design probe sequences from genome resources, synthesis attached to beads

- Make WGS library, hybridize with probe set. Matching sequence will be captured, all others washed away

- Collect capture sequence, amplify and sequence



Sequence Capture

Pros	Cons
Relatively cheap per sample	Requires designing probes
Good depth at targeted sites	Long library prep

Reduced Representation Sequencing

Instead of sequencing the whole genome, it can be sufficient to sequence just a part of it

Figure from Peterson et al PLoS One 2012

Reduced Representation Sequencing

Pros	Cons
Quick library prep for hundreds of samples	Relatively sparse SNPs compared to other methods - limiting analysis options
Comparatively cheap per sample cost	Can have problems overlapping different library preps

Part 3: Sequence alignments

Using any of the methods above, most projects will involve some form of sequence alignment

Part 3: Sequence alignments

Using any of the methods above, most projects will involve some form of sequence alignment

In this part of the tutorial, we will explore several sequence alignments obtained using different methods

Part 3: Sequence alignments

Using any of the methods above, most projects will involve some form of sequence alignment

In this part of the tutorial, we will explore several sequence alignments obtained using different methods

Flavours of DNA/RNA sequencing

- **Whole Genome Sequencing**
- Pool Seq
- **RNAseq**
- Amplicon Sequencing (GT-seq)
- Sequence Capture
- **Reduced-Representation Sequencing (RADseq/GBS/ RADcapture)**

Part 3: Sequence alignments
Whole genome resequencing

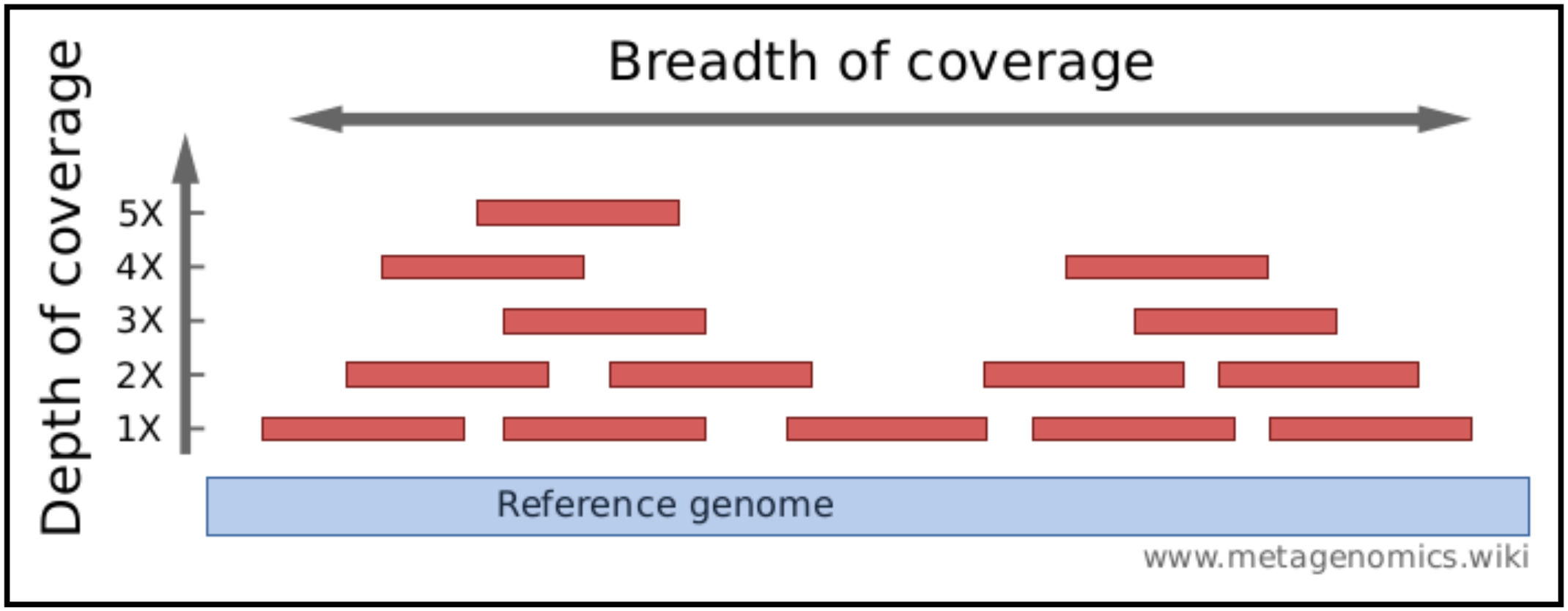
Follow the instructions under Part 3.1 of tutorial on the website

Load the the alignment files (BAM files) as indicated on the tutorial page and explore them using IGV

Start by comparing Salmon.HiSeq.10x.bam to Salmon.HiSeq.5x.bam

Part 3: Sequence alignments

Whole genome resequencing



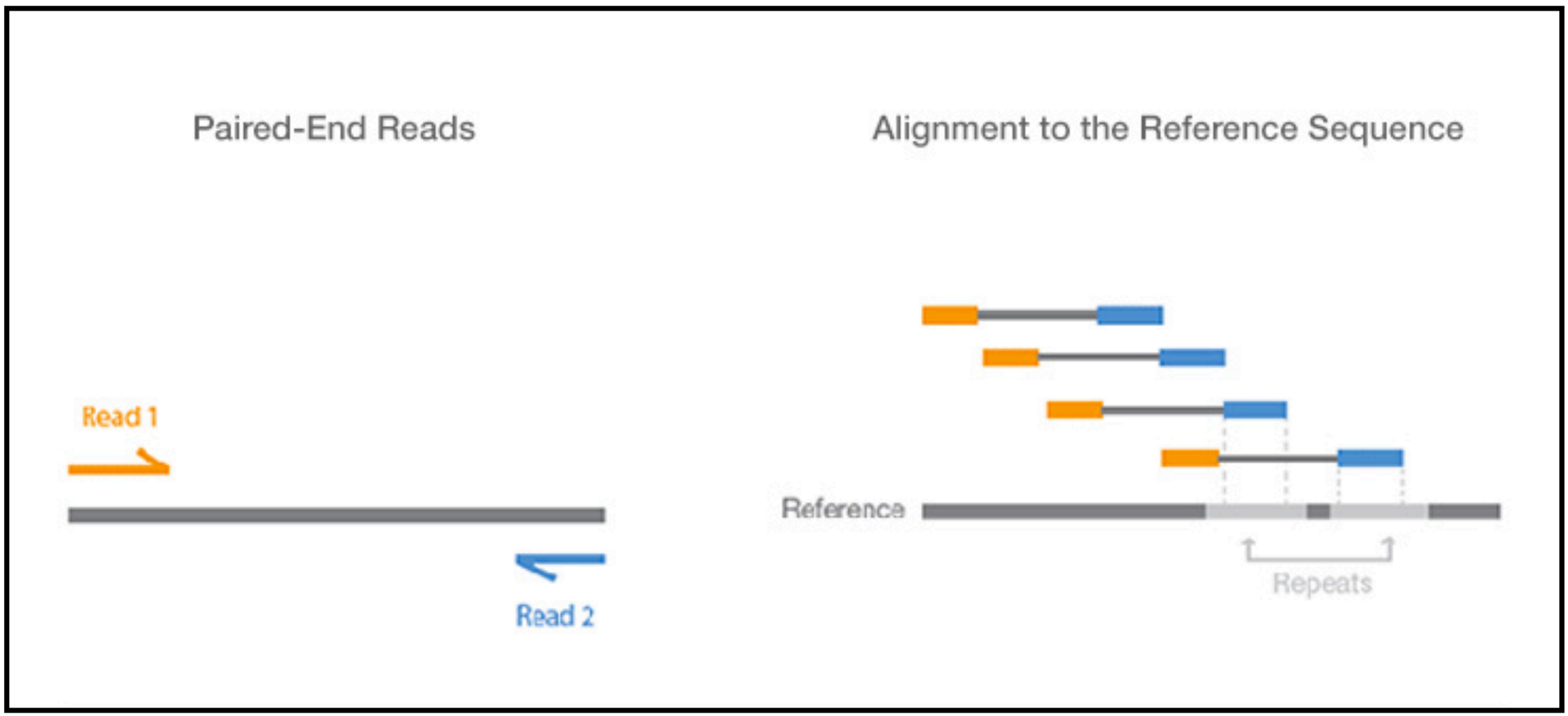
What does increasing sequencing depth give you?

What does increasing breadth give you?

Part 3: Sequence alignments

Whole genome resequencing

Why paired ends?



Picture from Illumina website

Illumina Machines

Name	MiSeq	HiSeq 4000	NovaSeq 6000
Sequencing Capacity	8Gbp	50Gbp	500-600Gbp
Cost (/lane)	~$1,500	~$3,000	~$8,000

We've been looking at HiSeq data

Long read sequencing

Two dominant companies are PacBio and Oxford Nanopore

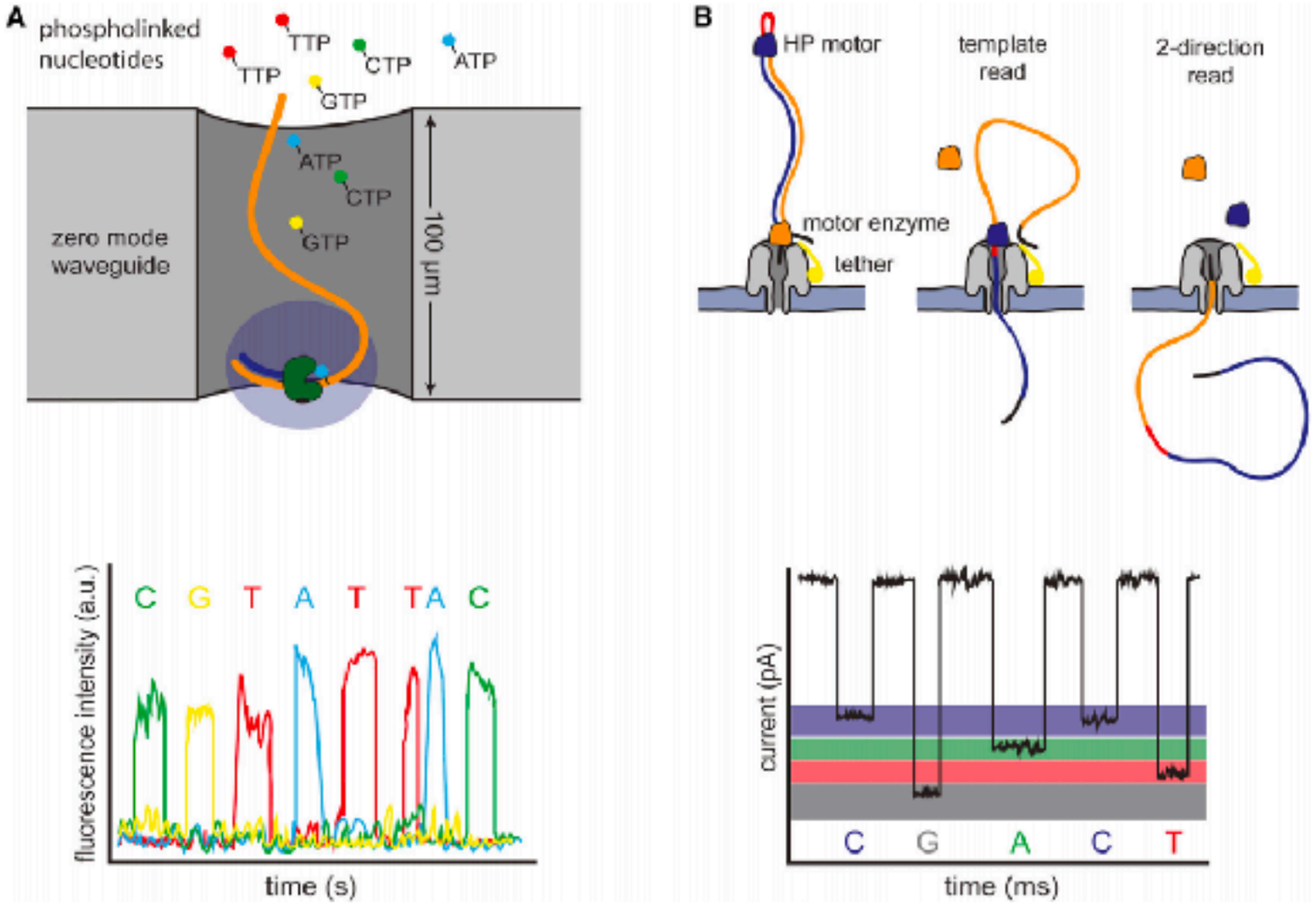


Figure 3. Single Molecule Sequencing Platforms

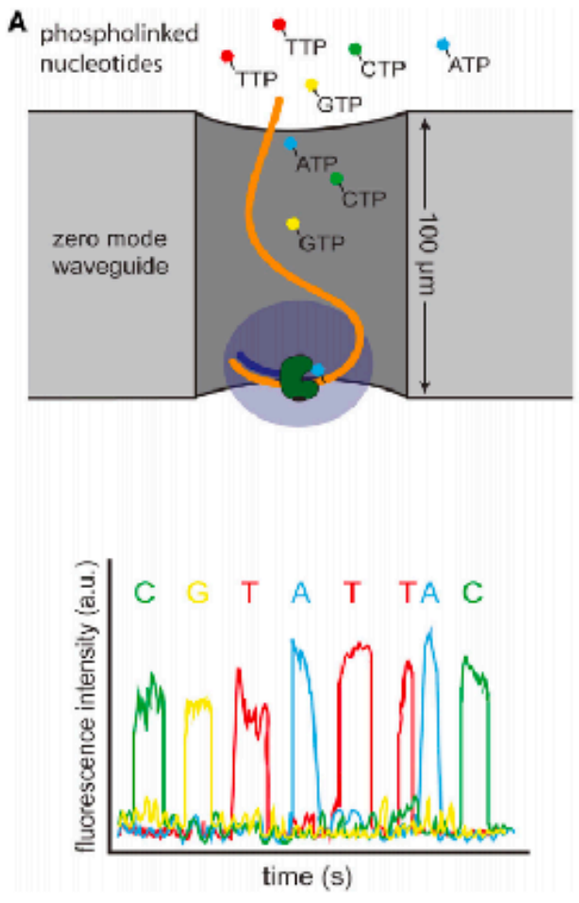
(A) Pacific Bioscience's SMRT sequencing. A single polymerase is positioned at the bottom of a ZMW. Phosphate-labeled versions of all four nucleotides a present, allowing continuous polymerization of a DNA template. Base incorporation increases the residence time of the nucleotide in the ZMW, resulting i detectable fluorescent signal that is captured in a video.

(B) Oxford Nanopore's sequencing strategy. DNA templates are ligated with two adapters. The first adaptor is bound with a motor enzyme as well as a teth whereas the second adaptor is a hairpin oligo that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the p are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads).

Excerpted from Reuter et al 2015 - Molecular Cell

Long read sequencing

PacBio - Pacific Biosciences



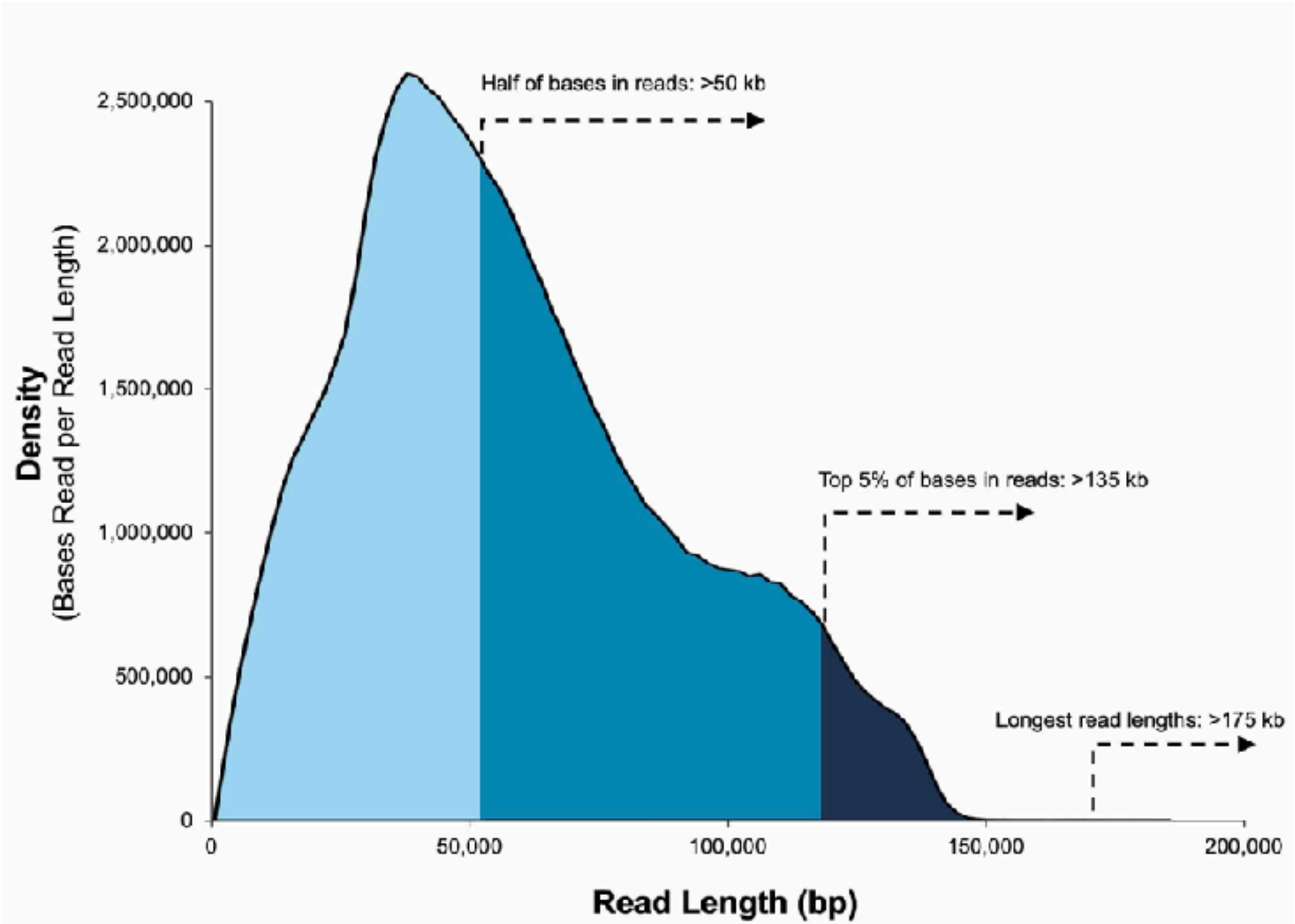
Sequel II

1-10Gb/flowcell

~$500/flowcell

13% error rate

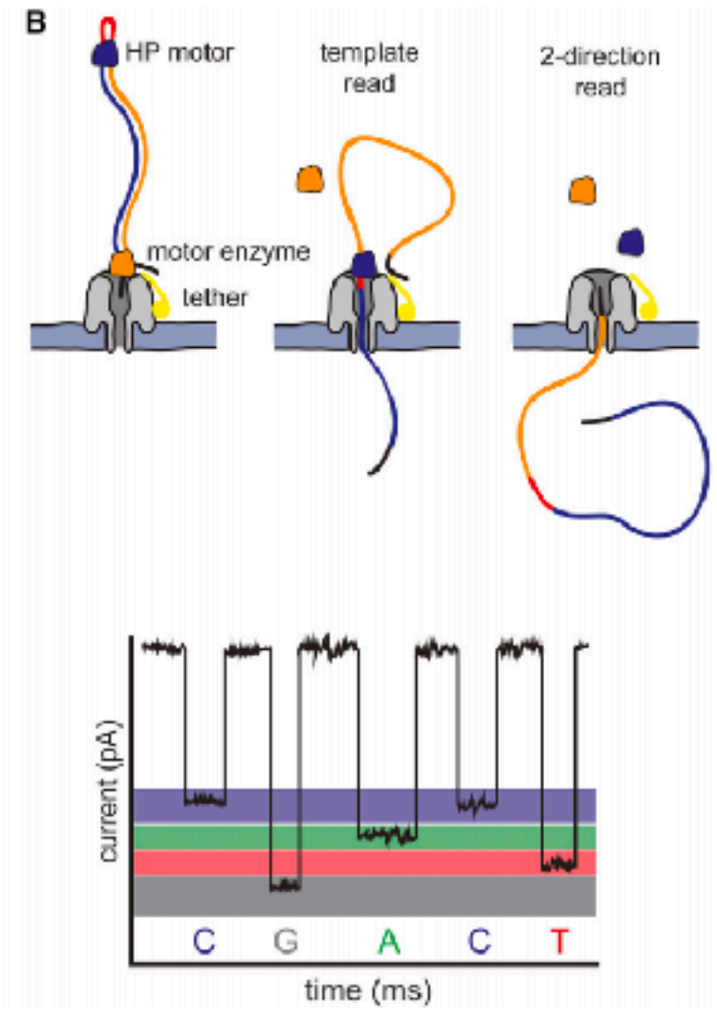
Pacific Biosciences



From the PacBio website

Long read sequencing

Oxford Nanopore



MinION

15-30Gb/flowcell

~$1000/flowcell

PromethION 24

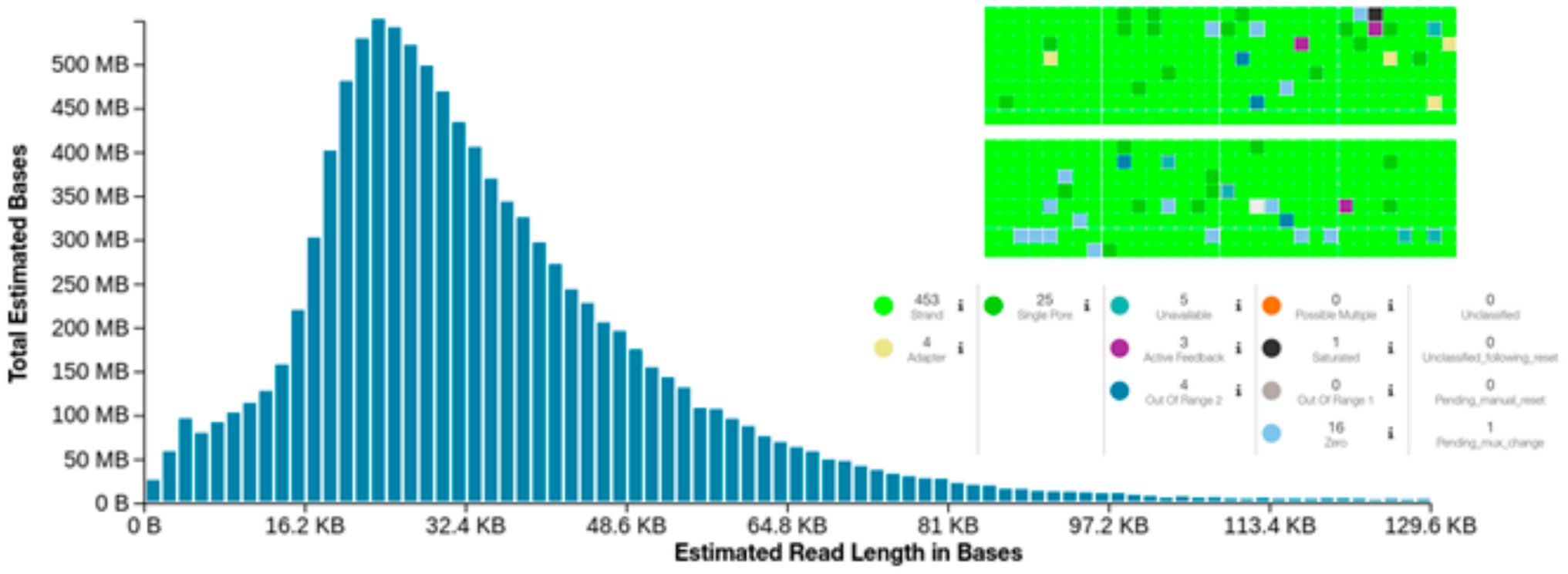
100-180Gb/flowcell

~$2000/flowcell

2-13% error rate

Oxford Nanopore

(C) *Eucalyptus albens;* end ligation library prep (SQK-LSK109). Output: 12.50 Gb.



Part 3: Sequence alignments
Whole genome resequencing

Follow the instructions under Part 3.2 of tutorial on the website

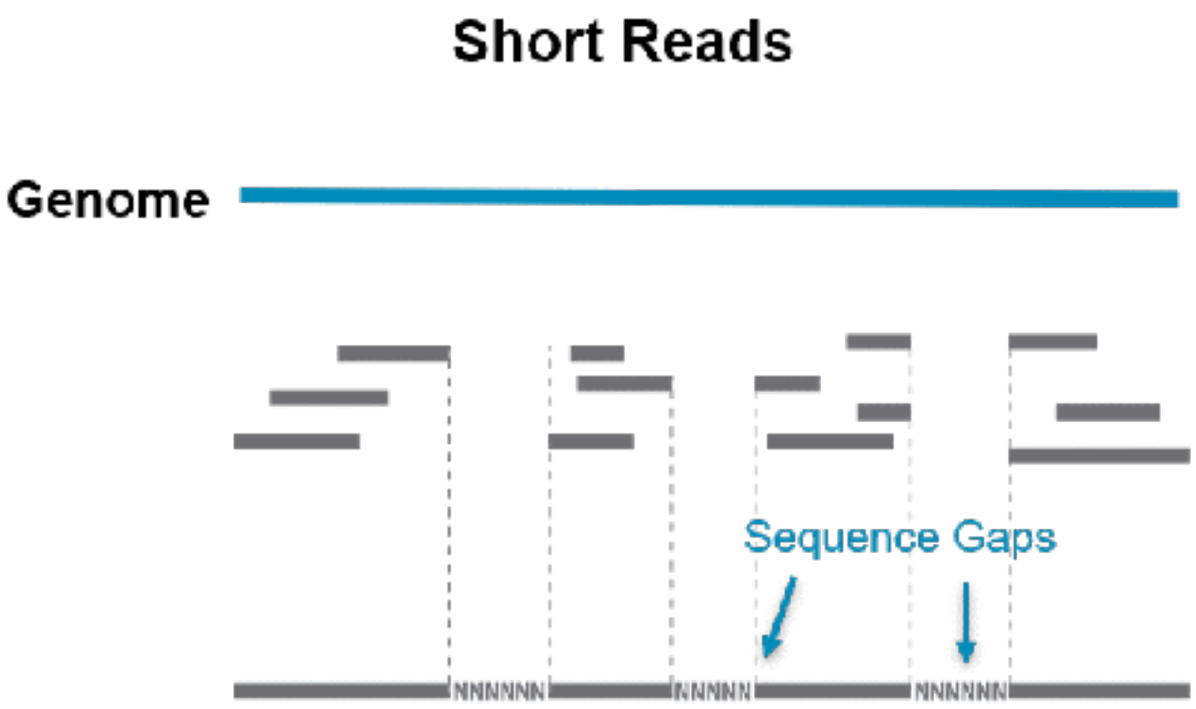
Load up Salmon.16x.PacBio.bam and explore it using IGV

Part 3: Sequence alignments
Whole genome resequencing

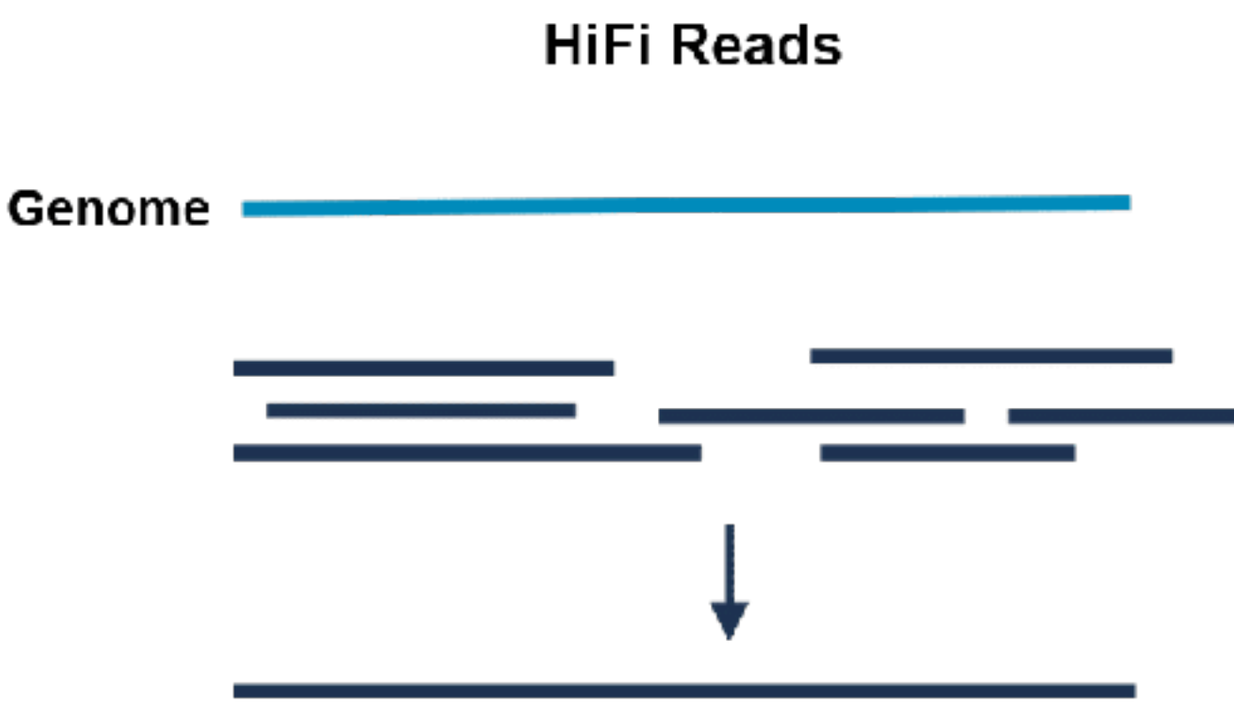
What do you think the longer reads would be useful for?

What are long reads bad at?

Part 3: Sequence alignments
Whole genome resequencing



Missing sequencing leads to missed genes and limits biological interpretation



A comprehensive structural, functional and organizational picture of the genome

Part 3: Sequence alignments
Whole genome resequencing

Short Reads		Long Reads	
Pros	**Cons**	**Pros**	**Cons**
Extremely accurate for complex regions	Rely on amplification, which can introduce errors (at a rate of around 10^{-6}-10^{-7}/bp).	Great for genome assembly • 30-60X coverage from ion torrent or PacBio will produce a nice draft genome.	More difficult library prep
Allele frequencies can be scored at many sites across the genome	Assembling and aligning short reads in repetitive regions is very challenging -> impossible	Can characterise alternate splicing of genes.	May be too expensive to be used for population level sequencing.
Very cost-effective	Both large and small structural variants pose difficulties	Structural rearrangement discovery and genotyping.	High error rate.

Part 3: Sequence alignments

Reduced representation sequencing

Follow the instructions under Part 3.3 of tutorial on the website

Load up Salmon.ddRAD.bam and explore it using IGV

Part 3: Sequence alignments

Reduced representation sequencing

What do you get with the ddRAD approach?

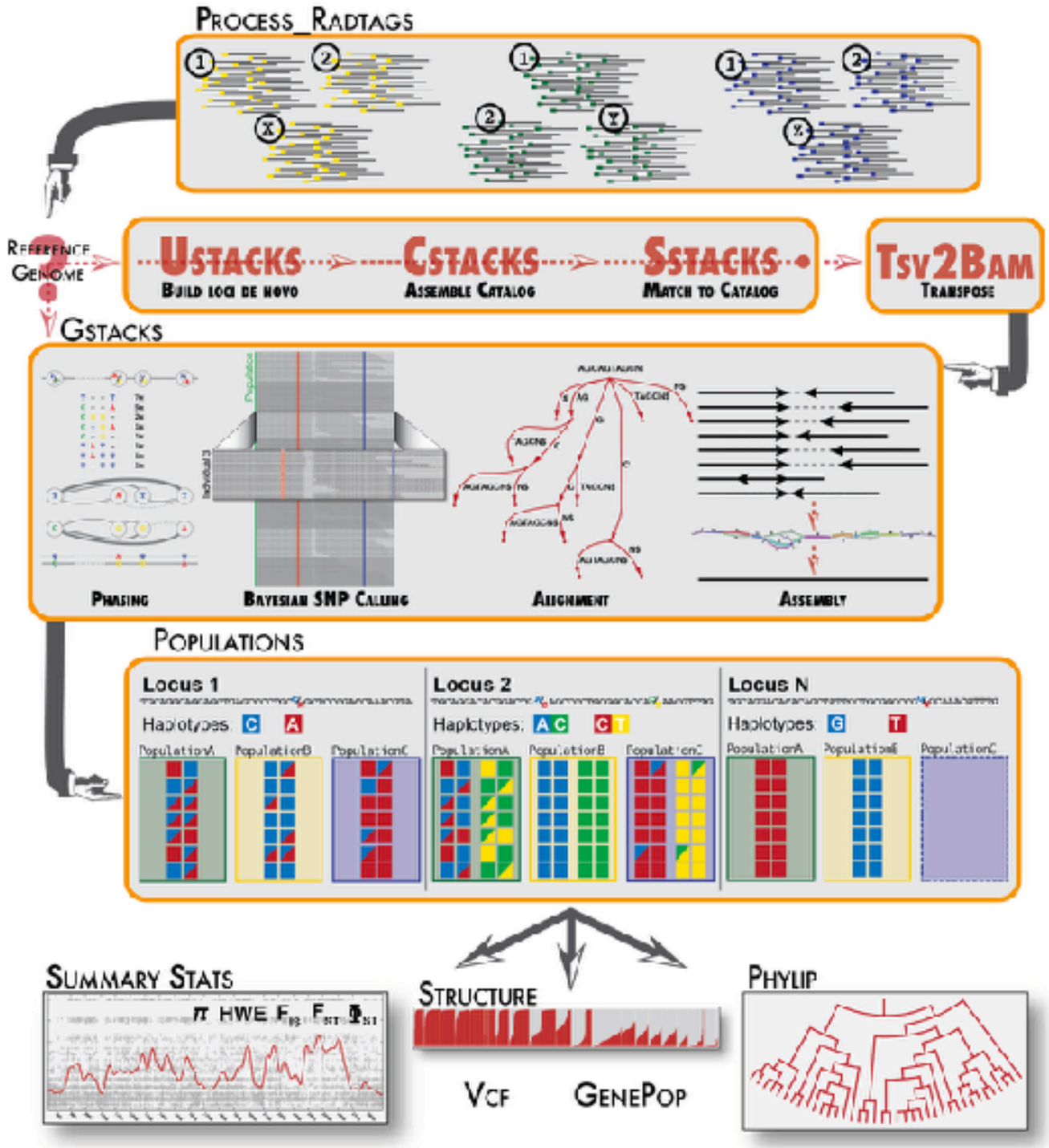
Can you think of some uses for this kind of data?

Part 3: Sequence alignments

Reduced representation sequencing

STACKs is an established pipeline for analysing reduced representation data (>2000 citations)

A real benefit is that you do not need a reference genome at all



Part 3: Sequence alignments

RNA-seq

Follow the instructions under Part 3.4 of tutorial on the website

Load up coldWaterSalmon.RNA.bam and explore it using IGV

Part 3: Sequence alignments

RNA-seq

You'll immediately notice the splice junctions that are inferred by STAR (the alignment software)

Can you think of any difficulties that might arise when trying to align RNA reads to a reference genome?

Part 5: Identifying variants

Follow the instructions under Part 4 of tutorial on the website

Re-load Salmon.HiSeq.5x.bam and Salmon.HiSeq.10x.bam as well as Salmon.HiSeq.20x.vcf.gz

Part 5: Identifying variants

Scroll around and inspect some variants

What are some features of sequencing that you would think would be useful when identifying variants?

How to choose?

The different technologies and methodologies have different pros and cons

What you use will obviously be informed by budget, but the biological question should also drive your choice

Further reading

PDFs are available on the GitHub page for this topic:

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, *7*(5), e37135.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345-353.

Extra stuff

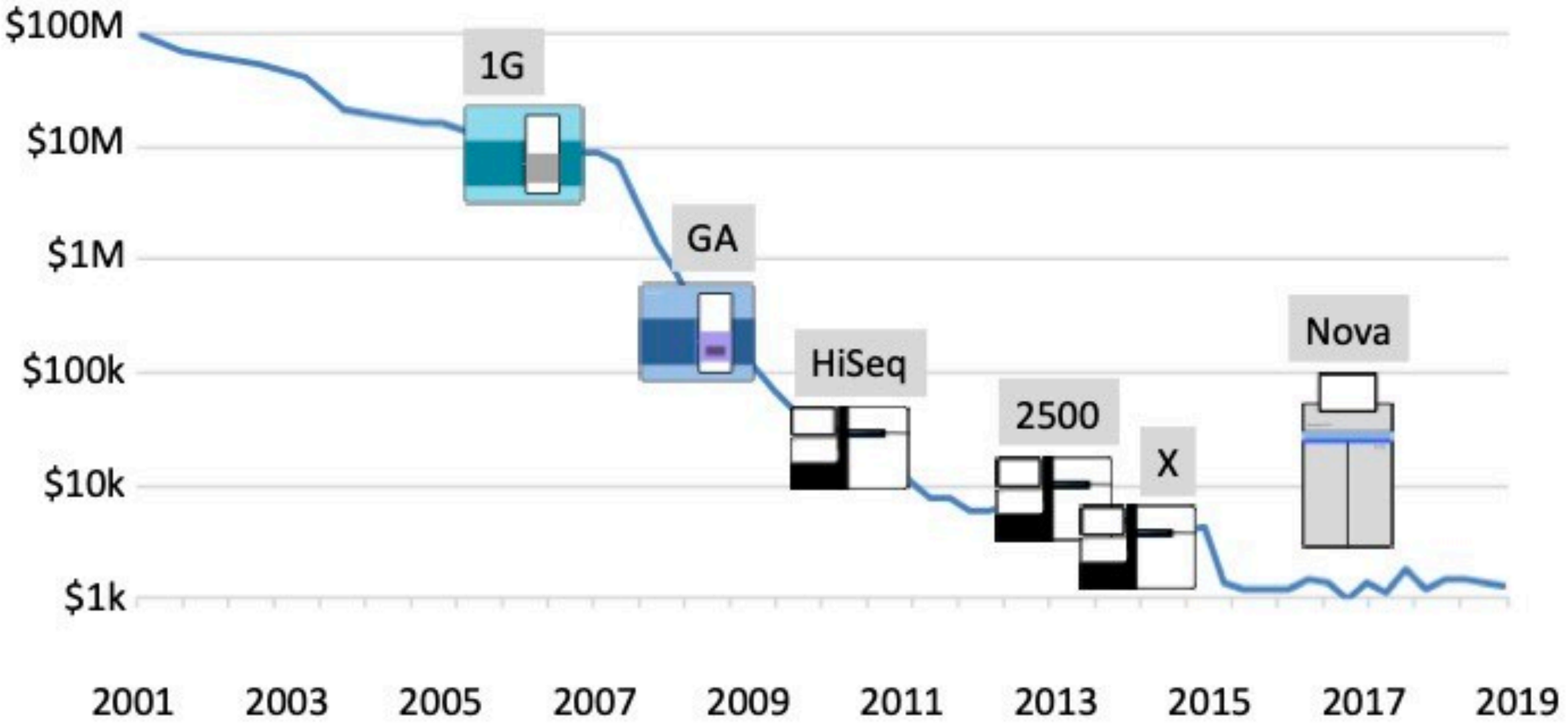
How to choose?

For example,

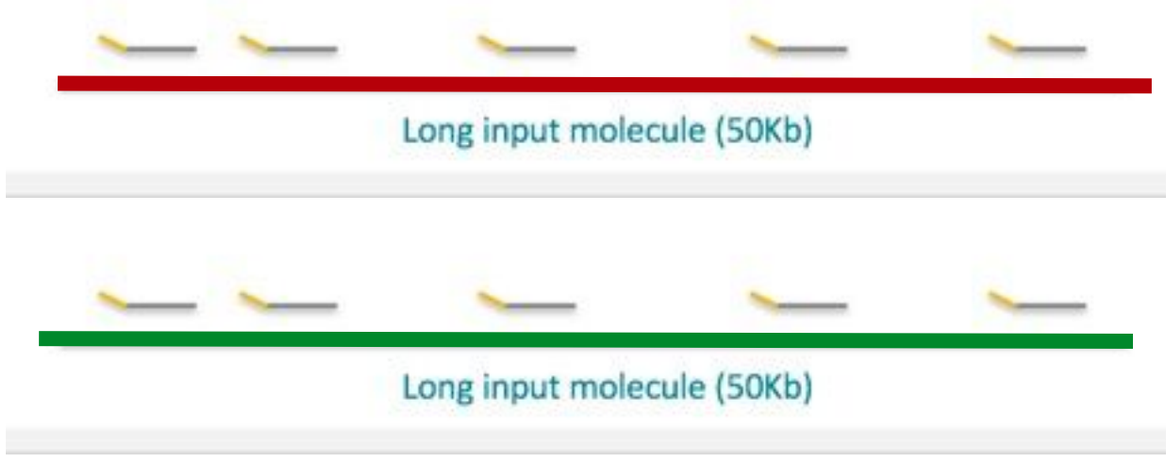
If you wanted to estimate demographic history from the distribution of allele frequencies, a reduced representation method might suffice to obtain an estimate of the site frequency spectrum

Or, if you want to perform a genome scan, looking at how haplotype frequencies varied among populations, you'd probably need deeper, whole genome information - it all depends on the questions you are tackling

Production cost per 30x Human genome over 18 years



Synthetic long reads



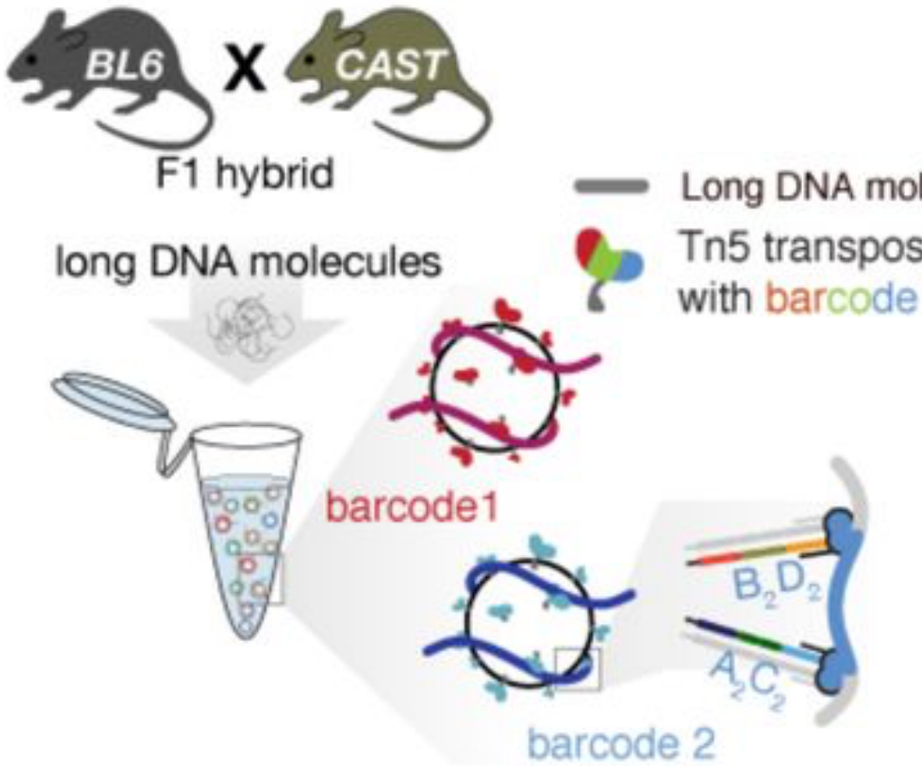
Barcodes read originating from individual DNA molecules

Sequence with Illumina reads

Original molecule can be reconstructed using the barcodes

Potentially very useful for genome assembly and phasing

Synthetic long reads



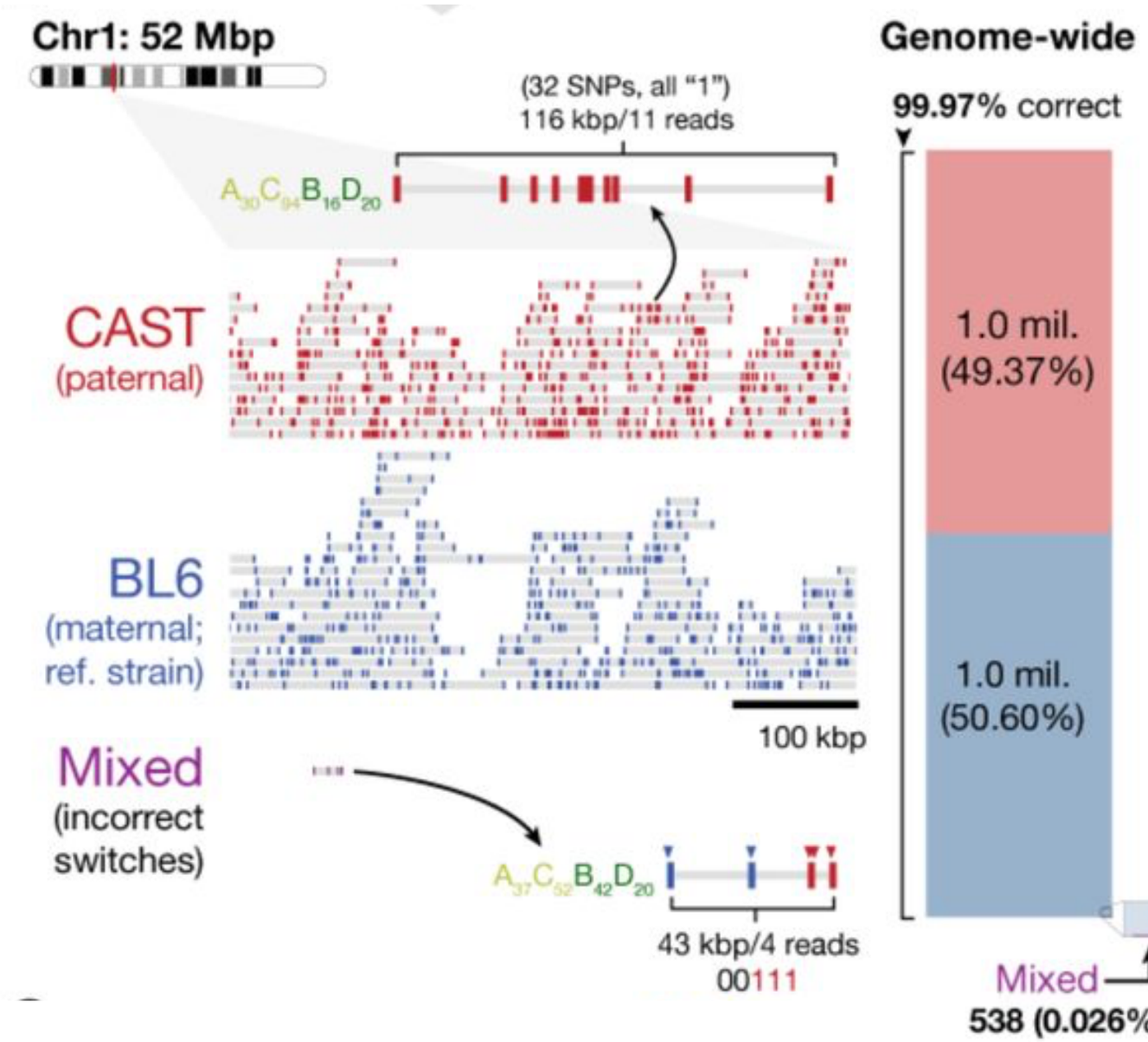


Figure adapted from Meier et al *Preprint* https://www.biorxiv.org/content/10.1101/2020.05.25.113688v2.full

Bisulphite Sequencing

Unmethylated cytosines are converted to **U**racil

Methylated **C**pG sites are unchanged and are detected as polymorphisms