BC Centre for Disease Control
Provincial Health Services Authority

Introductory Workshops on Whole Genome Sequencing Data Analysis

by Dr. Ben Sobkowiak

Series of three hands-on workshops in March 2020
Room 2264, Diamond Health Care Centre, Vancouver

Mar 10
1:30 - 4:30pm

Sequencing platforms & data quality control

Mar 13
1:30 - 4:30pm

Reference-based mapping & variant discovery

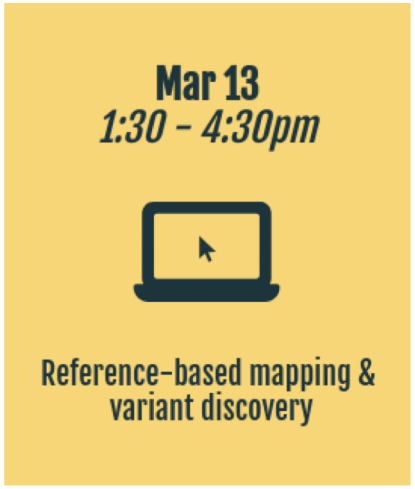
Mar 17
9:30 - 11:30am

De novo assembly & downstream analysis

Course Overview

- Introduction to genomic data
- Sequencing technologies
- Sequence data – first look
- Activity – visualization of data and quality control

Course Overview



- Reconstructing genomes using sequencing data
- SNPs and SNVs
- Other genomic features
- Activity – reference-based alignment and variant calling
- VCF files

Course Overview

- VCF files continued
- Summarizing variant calls for downstream analysis
- Activity – Building phylogenetic trees
- *De novo* assembly – introduction and short activity

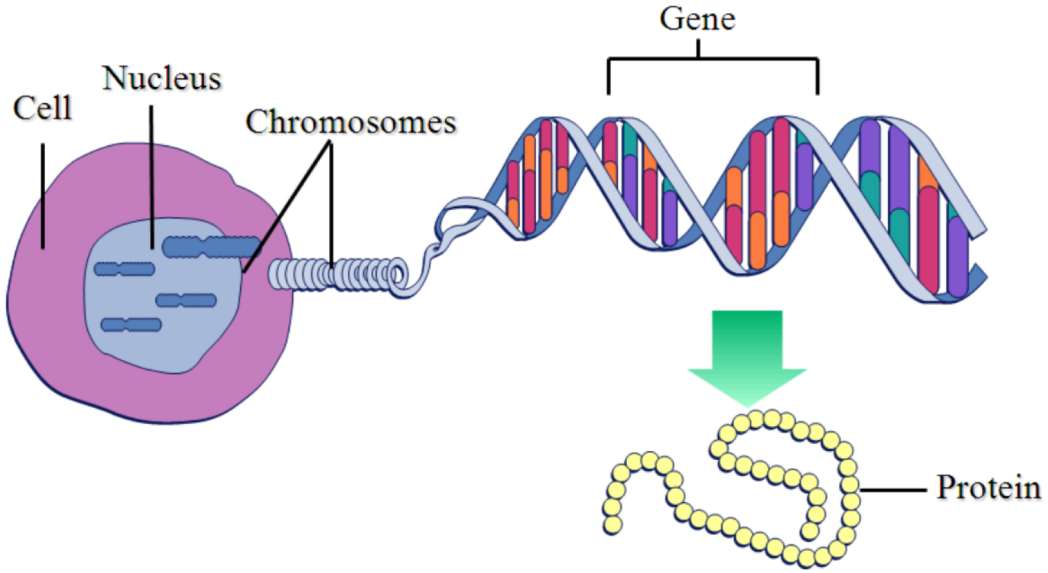
Purpose of the course

- Familiarize users with genomic sequence data and ‘demystify’ whole genome sequencing
- Process whole genome sequence data from raw sequences to variant data
- Introduce the benefits of employing genomic data to public health and basic science research
- Gain confidence using command-line interface and R language tools

What is genomic data and why do we use it?

Genetic vs Genomic sequencing

- Genetics is the study of single genes – inherited units of DNA
- Genes can be coding - instructions to make proteins to inform cellular function
- Or, non-coding - can still be integral for activity within the cell – transcription, promotors, enhancers, DNA structure



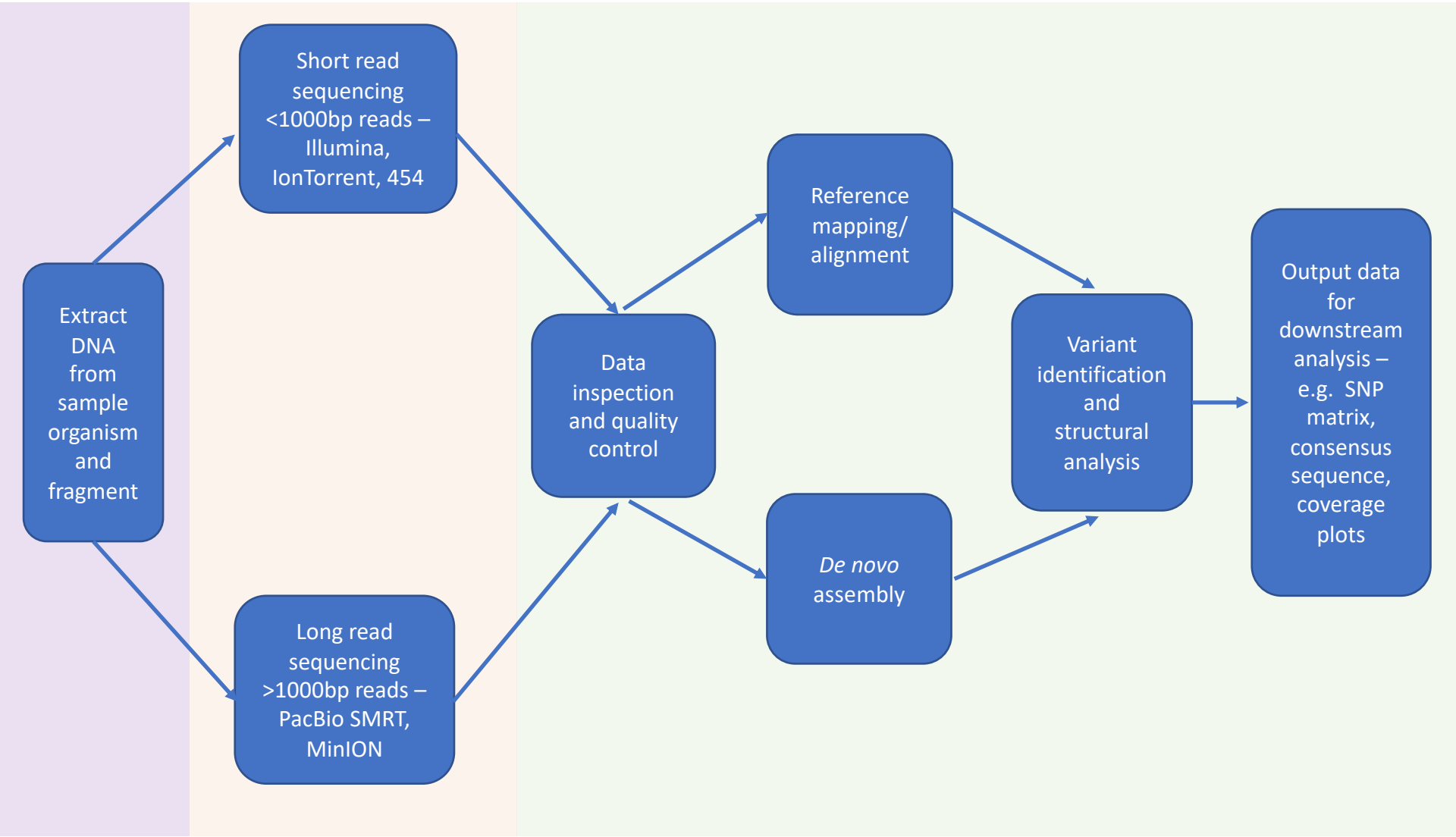
What is genomic data and why do we use it?

Genetic vs Genomic sequencing

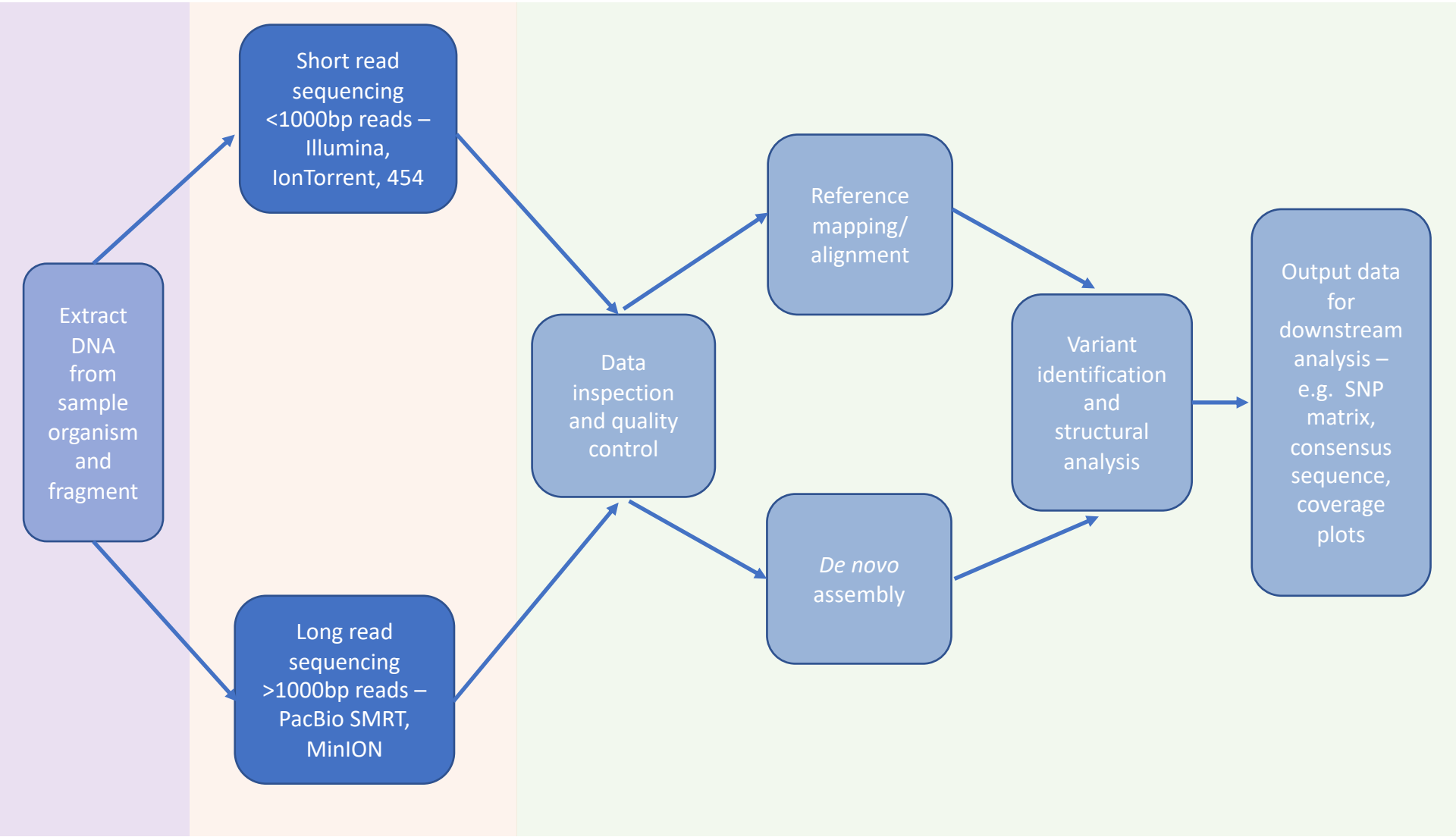
- Genomics takes all the genes of the organism, and intergenic regions, together – the whole genome.
- The majority of traits is not determined by a single gene – multi-locus genes, epistatic interaction
- Can investigate the interaction between the multiple genes and the environment
- Can investigate more complex characteristics, population effects, novel variation and environmental changes

Note: There are other ‘omics (transcriptomics, proteomics, metabolomics, systeomics) but we will just be focusing on genomics in this course

Genome sequencing and bioinformatics workflow



Genome sequencing and bioinformatics workflow



History of DNA sequencing

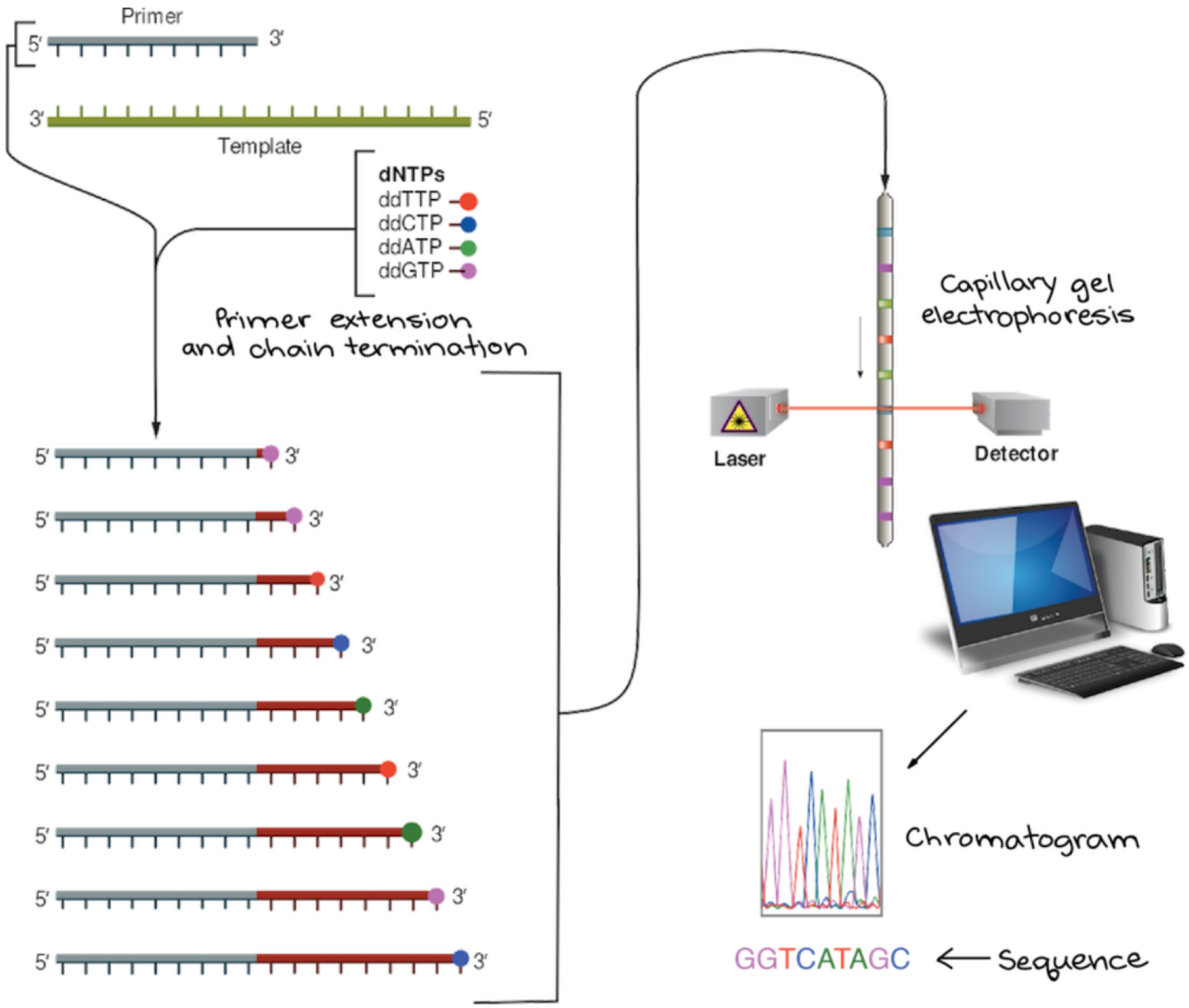
- Fred Sanger developed method in 1970s – "First-Generation" Sanger Sequencing

- Used in the Human Genome Project to sequence short stretches of DNA
  – Very time-consuming and expensive

- Although we now typically use other methods that are faster and cheaper, Sanger sequencing is still in wide use for the sequencing of individual pieces of DNA, or targeted sequencing

- Materials needed
  - A DNA polymerase enzyme
  - A **primer**, which is a short piece of single-stranded DNA that binds to the template DNA to start the process
  - The four DNA nucleotides (dATP, dTTP, dCTP, dGTP)
  - The target DNA to be sequenced

History of DNA sequencing

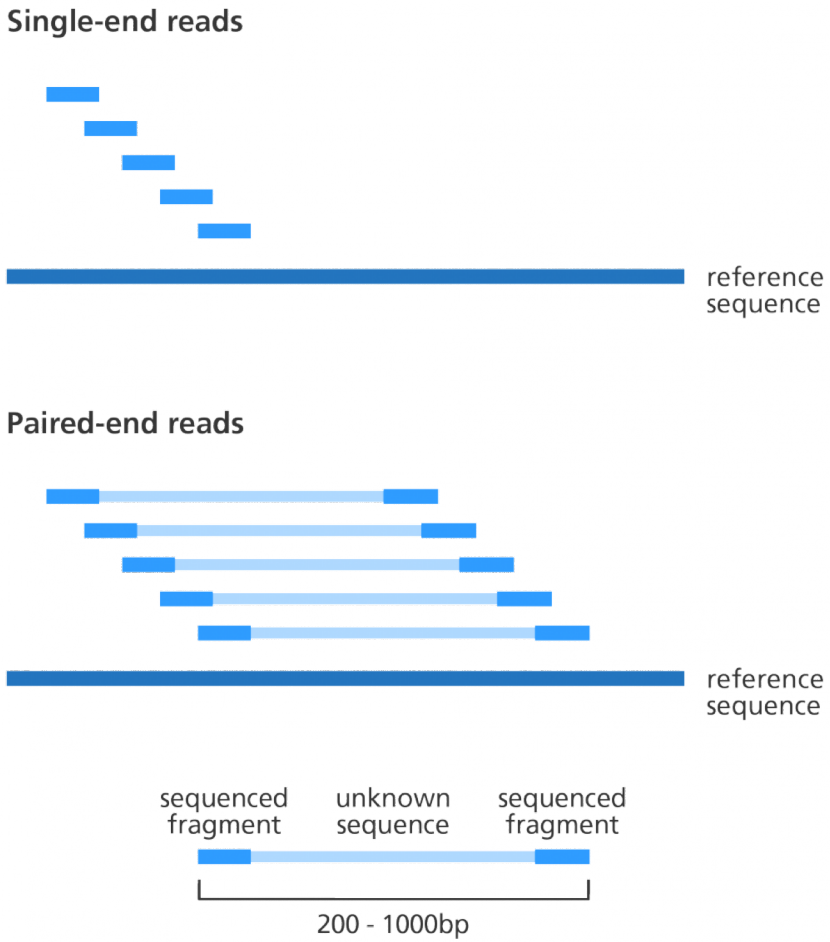


'Next-Generation' High-Throughput Sequencing

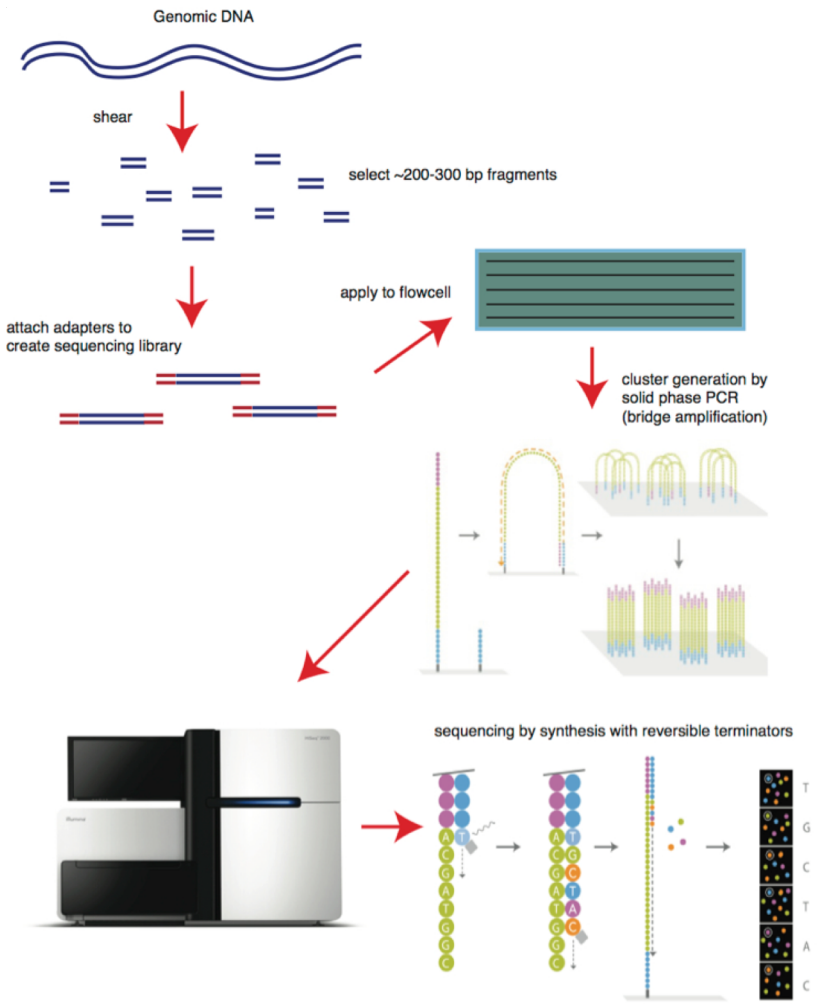
- Massively parallel sequencing – can sequence whole genomes quickly and deeply
- Millions to billions of DNA nucleotides can be sequenced in parallel
- For "Second-Generation" (short-read) sequencing - need to prepare amplified sequencing libraries – random fragments of cloned DNA
- Can also use cDNA from reverse transcribed RNA
- Nucleotide incorporation by synthesis is recorded directly by luminescence detection or by changes in electrical charge during sequencing
- The read types generated by NGS are digital and therefore enable direct quantitative comparisons

Short-read sequence reads

Single-end reads vs paired-end reads



Illumina short-read sequencing

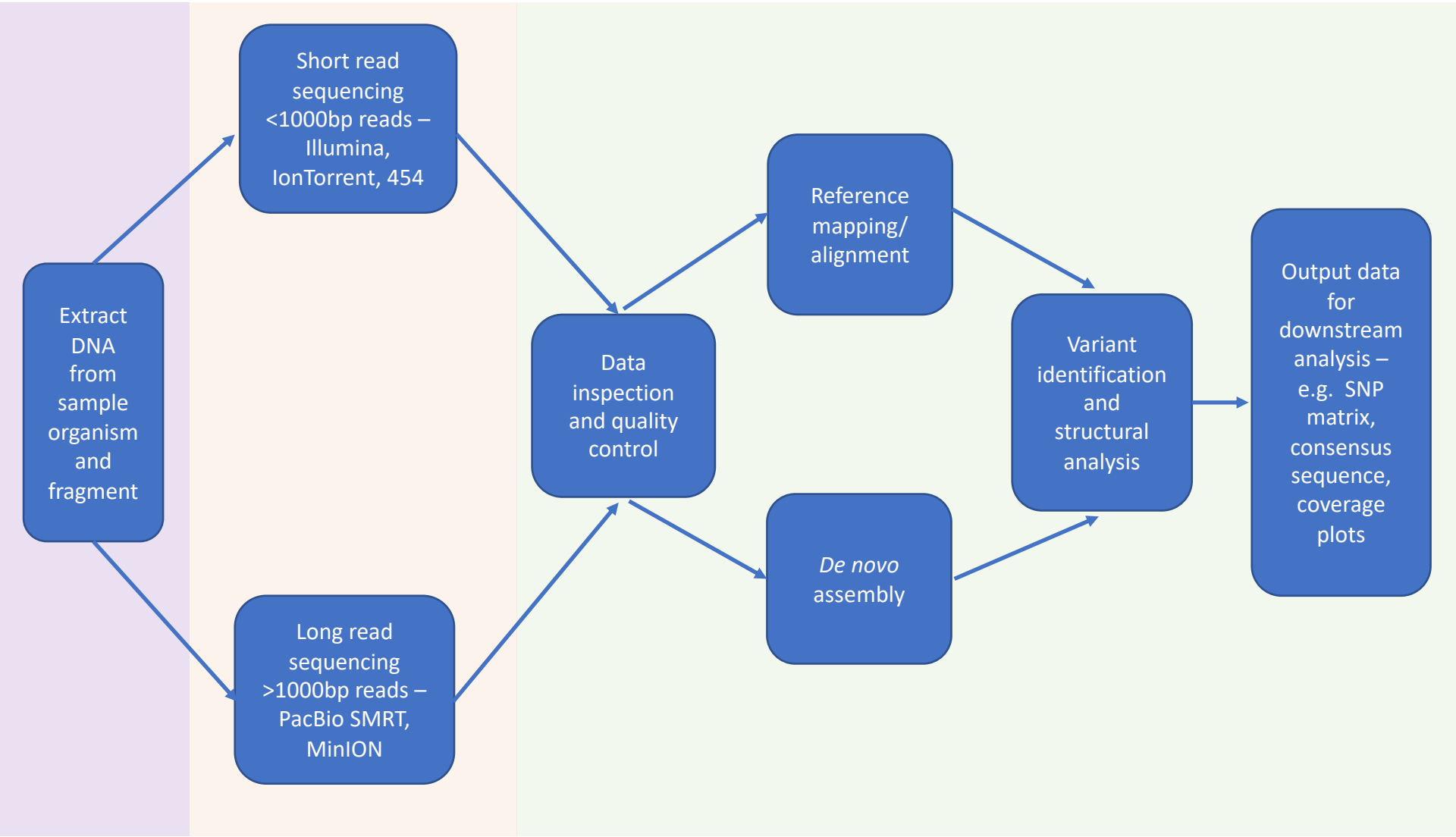


https://www.youtube.com/embed/HMyCqWhwB8E?iframe&rel=0&autoplay=1

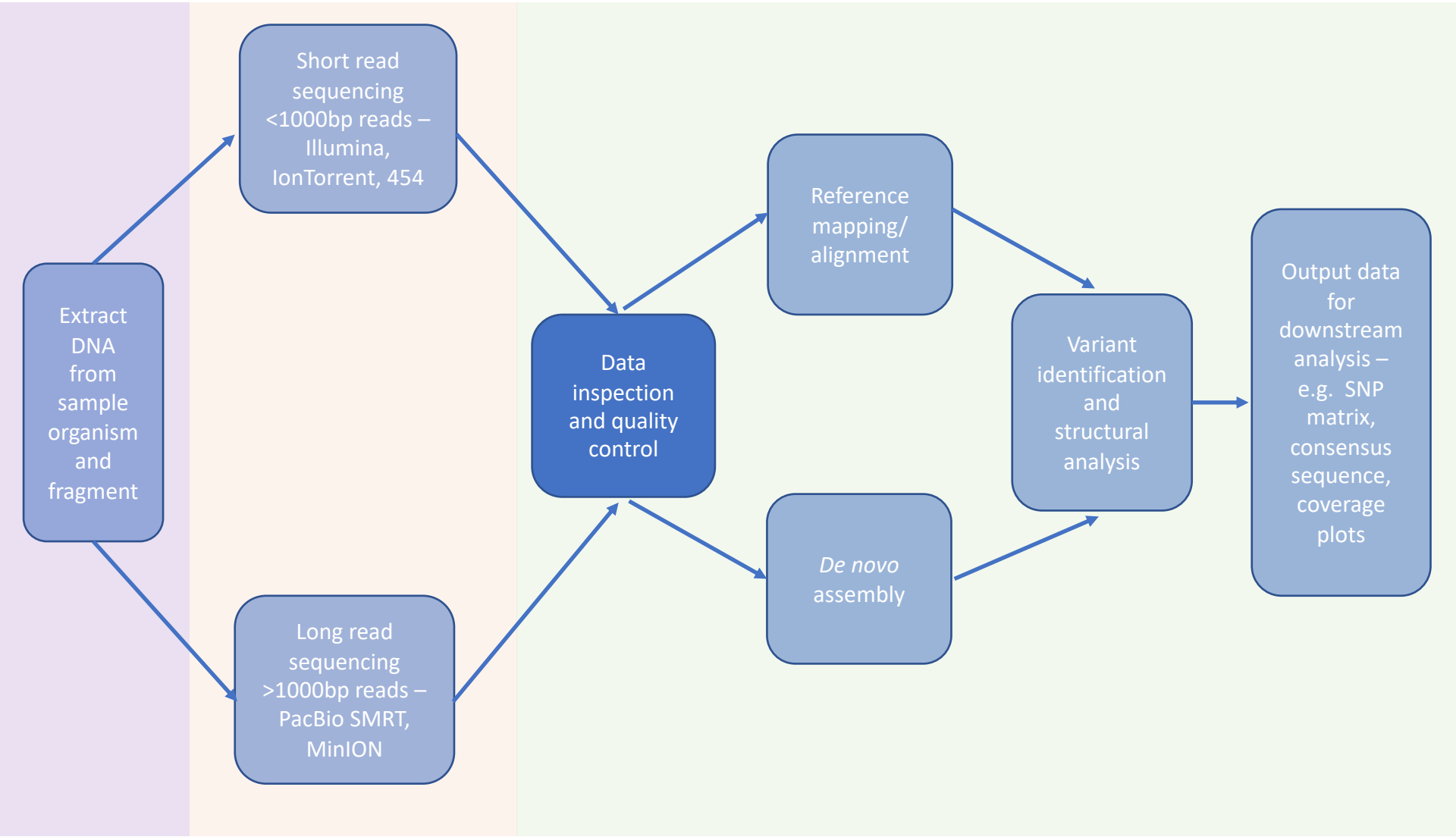
Sequencing platforms comparison

	Sanger	Roche 454	Thermo-Fisher Ion Torrent	Illumina MiSeq	Illumina HiSeq	Illumnia NextSeq	PacBio SMRT	Oxford Nanopore MinION
Generation	First	Second	Second	Second	Second	Second	Third	Third
Average read length	300-1000bp	200-700bp	200-600bp	300bp	150bp	150bp	~10-30Kb	~5-50Kb
Average yield	300-1000bp	~500MB	Up to 80GB	~15GB	300-2500GB	120GB	~4GB	15-30GB
Single or paired?	Single	Single	Single	Both	Both	Both	Single	Single
Cost per GB (USD)	~$2,500	~$5000	~$100-1000	~$100-300	~$5-500	~$50-100	~$10-50	~$7-100
Run time	20 mins - 20hrs	~10-24hrs	2-4hrs	4-55hrs	12-100hrs	12-30hrs	Up to 30hrs	1min-48hrs
Output file type	AB1/SCF	FASTQ	FASTQ	FASTQ	FASTQ	FASTQ	HDF5/ FASTQ	HDF5

Genome sequencing and bioinformatics workflow



Genome sequencing and bioinformatics workflow



Activity

Introduction to Whole Genome Sequence Data and Quality Control

- Inspect typical raw short-read sequence data, such as from Illumina platforms, in FASTQ format
- Quality control data and identify errors/issues
- Use command-line tools to clean data and fix issues

Please open the activity document - Day1-activity.docx

If you haven't already this can be downloaded from the GitHub – https://github.com/bensobkowiak/Genomics_workshop_2020_commands