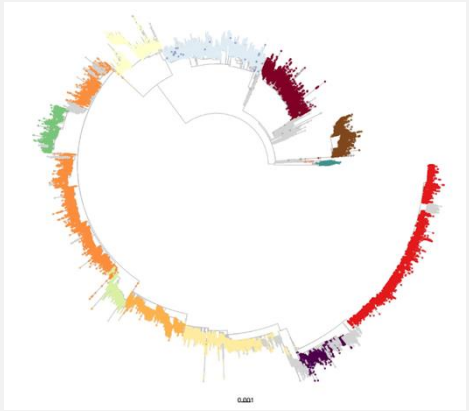
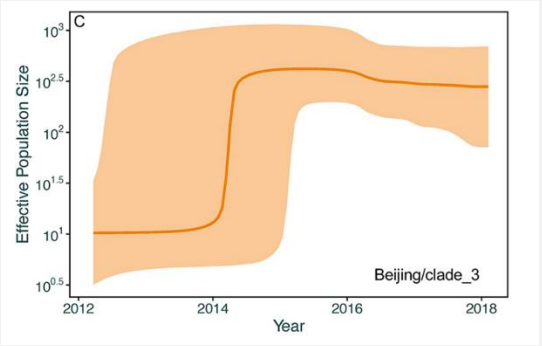
GENOMIC ANALYSIS AND PHYLODYNAMICS

Lecture 1: Introduction and Key Concepts





Instructor: Dr. Ben Sobkowiak
MRC Senior Research Fellow, University College London

PURPOSE OF THE WORKSHOP

- Familiarize participants with genomic sequence data and ‘demystify’ genomic epidemiology
- Process whole genome sequence data - from raw sequences to phylogenetic, phylodynamic and molecular evolution analyses
- Introduce the benefits of employing genomic data to public health and basic science research
- Gain confidence using command-line interface and R language tools

WORKSHOP OVERVIEW

https://bensobkowiak.github.io/BioinformaticsCourse/

Date	Time	Session	Modules
Saturday 3rd May	9:30–10:30	**Lecture 1: Introduction and Key Concepts**	• Course outline • Introduction to next generation sequencing and genomic epidemiology
	10:40–12:40	**Practical Session 1: Whole Genome Sequence Data Analysis**	• Obtaining sequencing data • Data manipulation and QC • Reference-based mapping and de novo assembly
	12:40–13:40	**Lunch Break**	
	13:40–14:15	**Practical Session 1 (cont.): Whole Genome Sequence Data Analysis**	• Catch-up, overview, QA
	14:15–15:15	**Lecture 2: Variant Detection and Phylogenetic Trees**	• What is a variant? How do we call variants? • Variant calling software and QC • What are phylogenetic trees? • Types of phylogenies, phylogenetic uncertainty (bootstrapping etc.)
	15:30–18:00	**Practical Session 2: Variant Calling and Maximum Likelihood Trees**	• Variant calling • SNP filtering and QC • Building SNP matrices • Aligning consensus sequences • Producing ML trees

WORKSHOP OVERVIEW

https://bensobkowiak.github.io/BioinformaticsCourse/

Sunday 4th May	9:30–11:00	**Lecture 3: Practical Applications of WGS and Phylogenetics**	• Species identification • Resistance and plasmid profiling • Transmission • Applications in real-world datasets
	11:00–12:00	**Practical Session 3: Timed Phylogenetic Trees**	• One-step timed phylogenetic tree with BEAST2 • Two-step timed phylogenies using ML + Bayesian frameworks
	12:00–13:00	**Lunch Break**	
	13:00–14:30	**Practical Session 3 (cont.): Timed Phylogenetic Trees**	• (cont.) One-step timed phylogenetic tree • Two-step timed phylogenies using ML + Bayesian frameworks
	14:45–16:45	**Practical Session 4: Transmission and Profiling**	• Identifying species, serotypes, and lineages from WGS • Inferring transmission networks/clusters
	16:45–17:00	**Closing Remarks - Short Course**	• Short course summary and feedback collection

WORKSHOP OVERVIEW

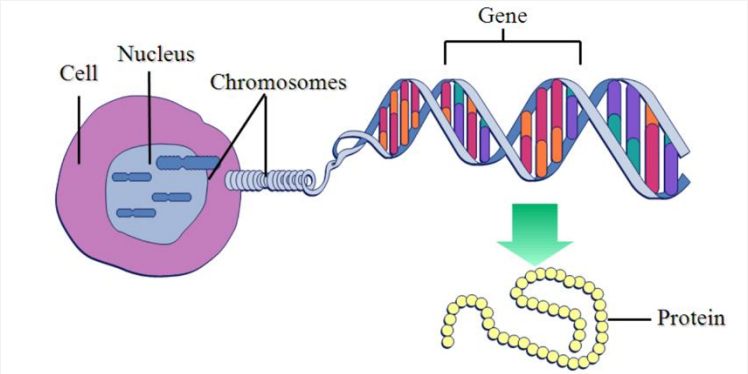
https://bensobkowiak.github.io/BioinformaticsCourse/

Monday 5th May (Advanced)	9:00–10:45	**Lecture 4: Advanced Applications of WGS**	• Phylogeography and phylodynamics • Recombination • Average Nucleotide Identity (ANI) • Mixed infection • Fitness and selection
	11:00–12:00	**Practical Session 5: Mixed Infection, Recombination and ANI**	• Identifying mixed infection • Calculating ANI • Testing for recombination
	12:00–13:00	**Lunch Break**	
	13:00–14:00	**Practical Session 5 (cont.): Mixed infection, Recombination and ANI**	• (cont.) Identifying mixed infection • Calculating ANI • Testing for recombination
	14:15–15:30	**Practical Session 6: Phylogeography and Phylodynamics**	• Phylogeography (ancestral state reconstruction) • Phylodynamic analysis with BEAST2 (Skyline analysis)
Tuesday 6th May (Advanced)	9:00–12:00	**Practical Session 7: Fitness and Selection**	• Strain-specific fitness (LBI) • Site-specific selection (homoplasy, dN/dS) • GWAS
	12:00–12:30	**Closing Remarks - Advanced Course**	• Full course summary and feedback collection

LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Genetics vs Genomics

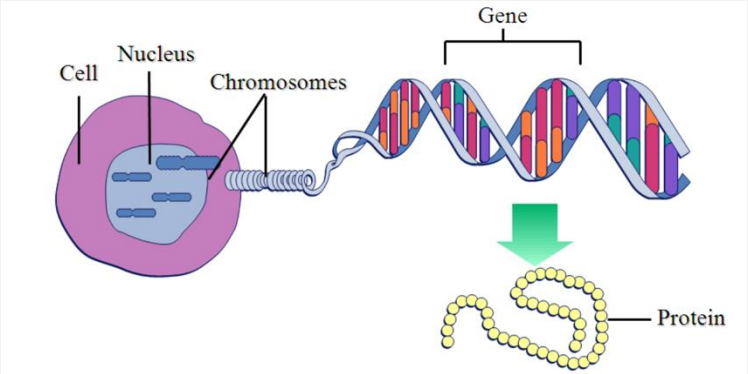
- Genetics is the study of single genes – inherited units of DNA or RNA
- Genes are coding - instructions to make proteins to inform cellular function
- Regions of non-coding DNA - can still be integral for activity within the cell – transcription, promotors, enhancers, DNA structure



LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Genetics vs Genomics

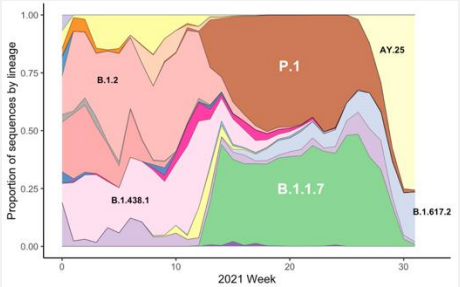
- Genomics takes all the genes of the organism, and intergenic regions, together – the whole genome.
- The majority of traits are not determined by single genes – multi-locus genes, epistatic interaction
- Can investigate the interaction between the multiple genes and the environment
- Also, more complex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



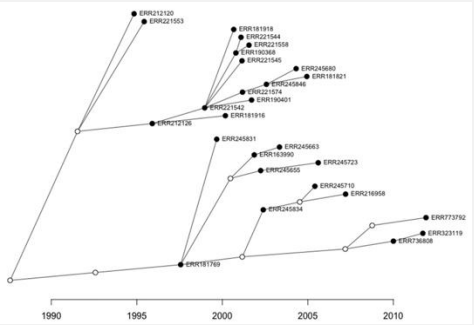
LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

The impact whole genome sequencing for investigating pathogens

- Genomic epidemiology - the use of genomic data to understand the patterns, causes, and effects of health and disease conditions in populations.
- Particularly crucial in studying the transmission and evolution of infectious pathogens.
- Incorporating whole-genome sequencing, phylogenetic analysis, and comparative genomics,
- Enables the tracking of pathogen transmission, identification of outbreak sources, and understanding pathogen evolution and resistance.



Sobkowiak, Colijn et al. 2022

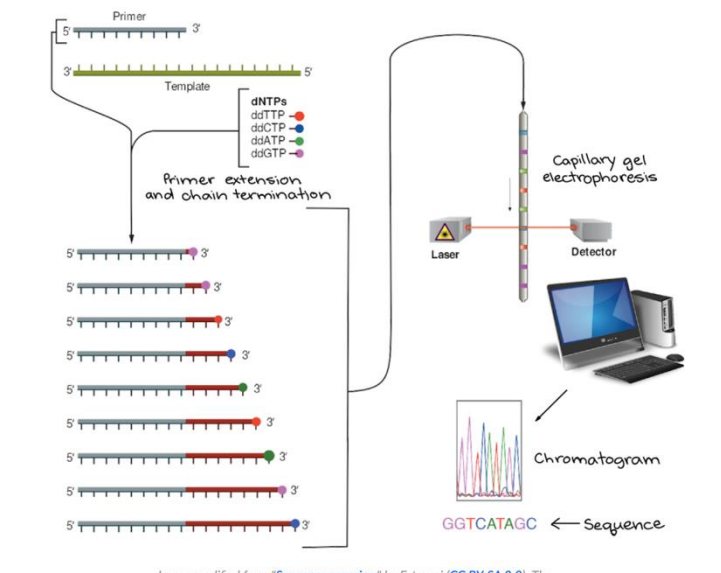


Sobkowiak et al 2020

LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Early approaches to sequencing

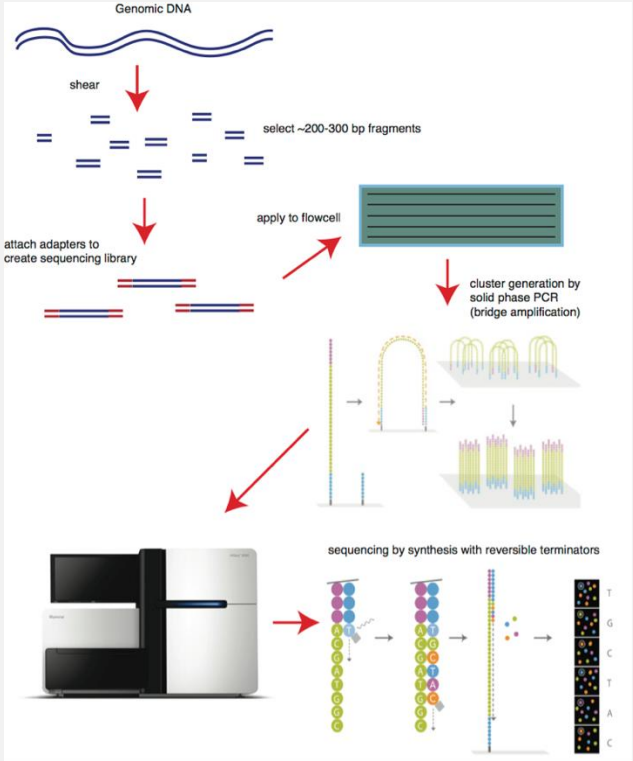
- Fred Sanger developed method in 1970s – "First-Generation" Sanger Sequencing
- Used in the Human Genome Project to sequence short stretches of DNA
  - Very time-consuming and expensive
- Although we now typically use other methods that are faster and cheaper, Sanger sequencing is still in wide use for the sequencing of individual pieces of DNA, or targeted sequencing



LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Next generation sequencing (NGS)

- Massively parallel, high-throughput sequencing – can sequence whole genomes quickly and deeply
  - Pivotal in large-scale genomics projects and complex genetic analyses
- "Second-Generation" (short-read) sequencing involves the preparation of amplified libraries – random fragments of cloned DNA or reverse transcribed RNA – usually sequenced on Illumina platforms
- Results in (hopefully) 100,000s or millions of short (~100 – 250bp) stretches of sequenced genome called 'reads'

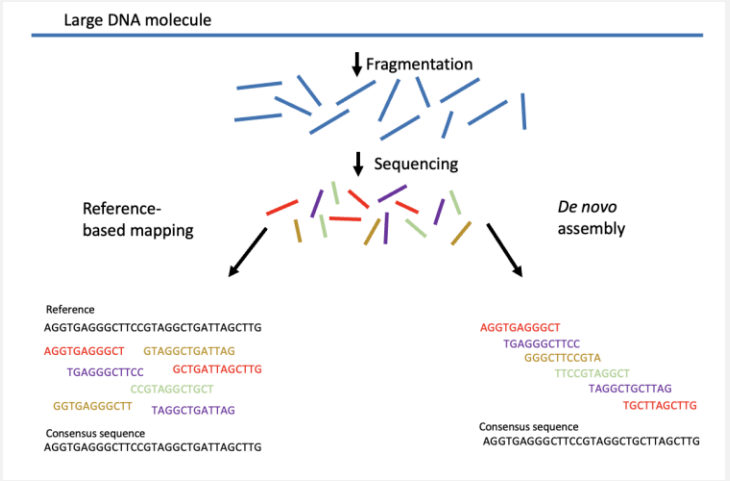


LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Next generation sequencing (NGS)

- Computationally intensive task to re-assemble these short 'reads' into full genomes
- The format of the files that are produced by the sequencer are called FASTQ
- Different approaches are available to reconstruct the genome from these reads, the choice depends on the data and research question
- Reference-based mapping/alignment or *de novo* assembly?

```
1 @M01637:250:000000000-BP8GK:1:1101:17234
2 GTCTAGAGACCGGGGACTTATCAGCCAACCTGTTACTAGA
3 +
4 CCCCCFFFFFDDGGGGGGGGGGHHHHHHGGHHHHHHHHHH
```



LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

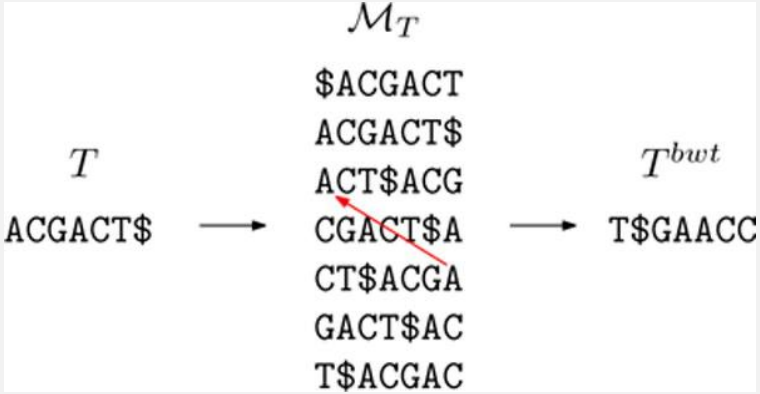
Reference-based mapping/alignment

- Most commonly-used method to reconstruct genomes from short-read sequence data
- The sequence in each read is aligned to a known reference genome
- There are different algorithms for refence-mapping
  - Trade-off between efficiency and sensitivity
  - Map sequence in reads to the reference whilst allowing for some error, mismatches etc.

LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Reference-based mapping/alignment

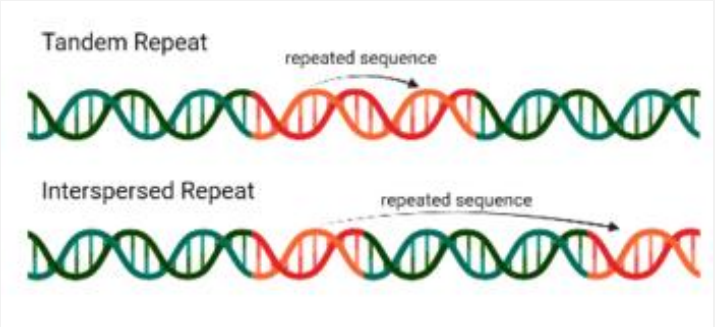
- Burrows-Wheeler Transform is a method widely used by software to map reads to a reference (e.g., BWA and Bowtie)
- It reorders the characters in a string (sequence) into runs of similar characters, allowing for compression of the data and efficient searching of matching sequences
- More information at: *Short Read Mapping: An Algorithmic Tour – Canzar & Salzburg 2015*



LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Reference-based mapping/alignment

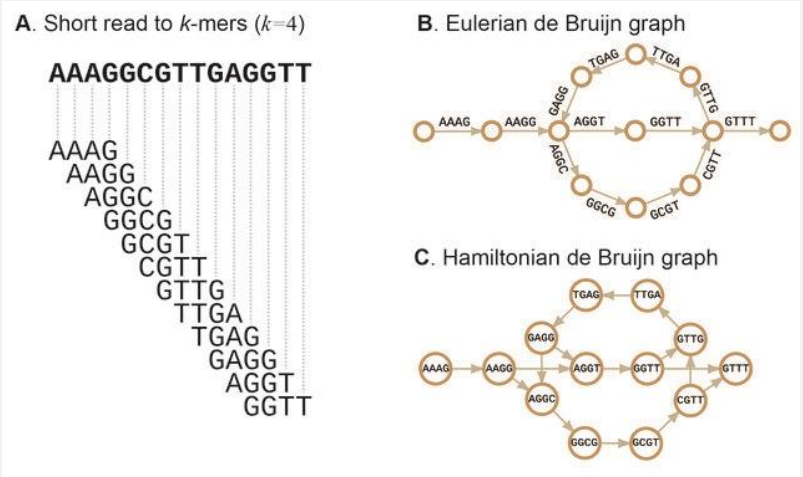
- Requires a well-characterized reference strain to be effective
- May not provide sufficient information to resolve ambiguous or repetitive regions of the genome
- Also, may miss novel genetic variants or full genes if not present in the reference sequence



LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

De novo assembly

- Most tools employ De Bruijn graph approach to *de novo* assemble genomes without a reference (e.g., SPAdes, Velvet)
- Transform short-read sequences into a graph structure where each node represents a k-mer).
- Edges connect overlapping k-mers, facilitating the reconstruction of the original sequence
- Can still be complex to resolve due to repetitive sequences and sequencing errors

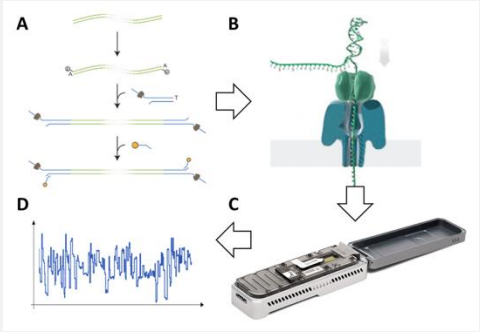


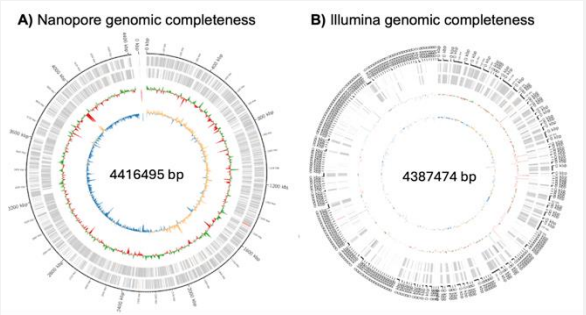
From Sohn & Nam, 2016

LECTURE 1: INTRODUCTION TO SEQUENCING AND GENOMIC SEQUENCE ANALYSIS

Third generation sequencing

- **Single molecule long-read sequencing** e.g. PacBio, Oxford Nanopore MinION/GridION/PromethION
  - Reads can be MBs or even GBs in length
- Even greater resolution (Identify rare variants and full complete the genome)
- Requires more genetic material as input and error rates typically higher than Illumina – though improving
- Potential for real-time outbreak analysis, drug susceptibility testing etc.





PRACTICAL 1: WHOLE GENOME SEQUENCE DATA ANALYSIS, MAPPING AND ASSEMBLY

1. Obtaining sequencing data
2. Viewing raw sequence data (FASTQ) files
3. Quality control (QC) of FASTQ files
4. Cleaning and filtering FASTQ files
5. Mapping/aligning sequence data to a reference genome
6. *De novo* assembly