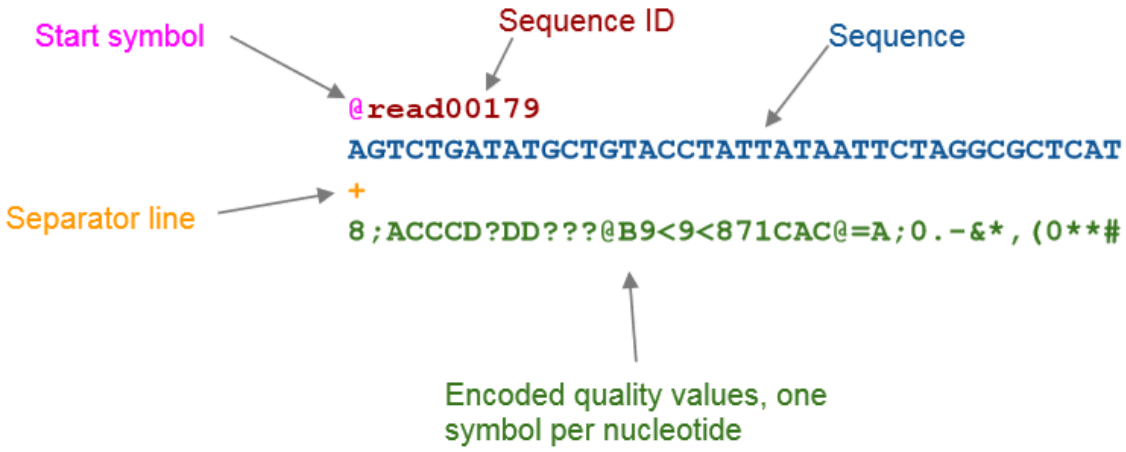
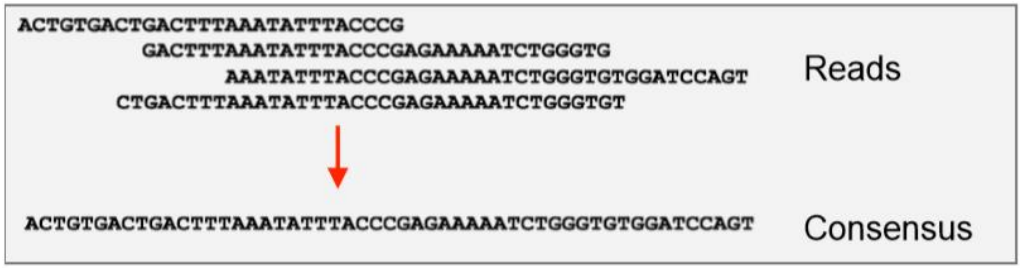
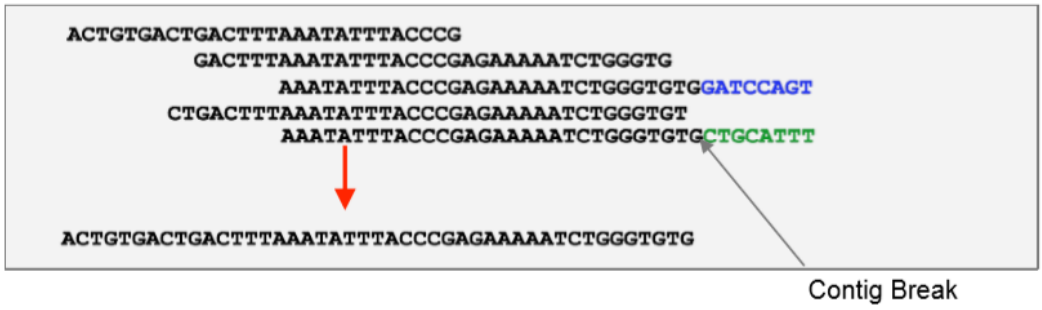


de novo Assembly

- *de novo* transcriptome assembly is the method of creating a transcriptome without the aid of a reference genome
  → transcriptome = sum total of all mRNA molecules expressed from the genes of an organism
- Overlapping reads can be summarised to a consensus sequence
  → an unbroken consensus sequence is called a "contig"

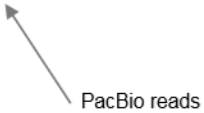


- Only a draft sequence
  → sequence of DNA with less accuracy than a finished sequence
  → some segments are missing
  → some segments are in the wrong order/oriented incorrectly
  → there are many contigs as there are several copies of the same sequence on the genome
  ➔ e.g. rRNA gene operon

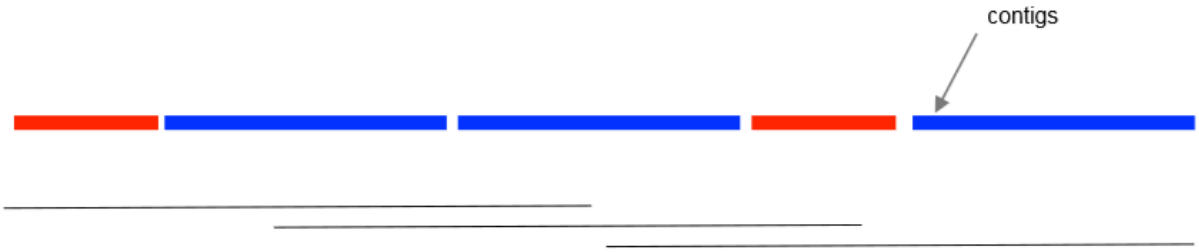


PacBio Sequencing

- Single molecule sequencing
  → one DNA polymerase molecule per well
- Synthesises normal DNA from modified nucleotides
  → fluorophores (one for each of the 4 bases)



- Produces up to 40, 000 base reads
  → low quality reads can be used to 'join' the Illumina contigs



Instrument	Method	Read Length	Yield	Quality	cost per base
Illumina	synthesis + fluorescence	250	++++	+++++	Low
~~SOLiD~~	ligation + fluorescence	75	++++	+++	Medium
Ion Torrent	non-term NTP + pH wells	300	++	+++	Medium
~~Roche 454~~	non-term NTP + luminescence	600	+	++++	High
PacBio	synthesis + ZMW	40,000	+++	+	High
Oxford Nanopore	https://nanoporetech.com/science-technology/how-it-works	40,000++	+	+	??

Closing the genome sequence

- Joining contigs
  → sort the contigs into a best guess order
    ➔ assume similar gene order to a related organism
    ➔ design oligonucleotide primers for PCR amplification of 'gaps' between contigs
- Predicts proteins encoded on the genome
  → predicts function by similarity to characterised proteins
  → insight into molecular tools available to the organism
- Comparison of genome sequences
  → particularly closely related isolates
  → phenotypic differences can be attributed to genotypic differences
    ➔ determines function of genomic elements

Single nucleotide polymorphisms (SNPs)

- SNPs are variation of one nucleotide at a particular position
  → any one of the four DNA bases may be substituted for any other
- Most SNPs have only two alleles
  → e.g. some people have a T at certain place in genome, everyone else has a G
  that place in the genome is a SNP with a T allele and a G allele

- Human genomes collectively contain more than 18 million SNPs
  → each individual has a unique SNP pattern/fingerprint
  → to fully classify an individual's haplotype, need to map all their SNPs or sequence their genome
- SNPs account for the majority of human sequence variation

- Mapping an individual's SNPs can provide physiological information
  → e.g. one of the genes associated with Alzheimer's is apolipoprotein E (ApoE)
    ➔ ApoE is involved in cholesterol transport
  → ApoE gene contains two SNPs that result in three alleles: E2, E3, E4 = three protein variants
    ➔ individual who inherits at least one E4 allele will have greater chance of getting Alzheimer's
    ➔ individual who inherits E2 allele will have less likely to develop Alzheimer's

Haplotype blocks

- Haplotype blocks are recombined DNA segments from parental chromosomes during meiosis
  → each has a unique SNP pattern
  → recombination between parental chromosomes during meiosis is non-random
    ➔ occurs at host spots = discrete segments of DNA are shuffled during meiosis
- Using SNP patterns to "barcode" different haplotype blocks
  → reduces complexity of individual's genotype
  → no need to physically sequence the entire genome or map every SNP to "fingerprint" someone

Disease management

- Determining SNP patterns (haplotype) of individuals tests for susceptibility for a specific disease
  e.g. heart disease, diabetes, Crohn's/Alzheimer's disease
  → polymorphisms may modify proteins and hence drug responses
    ➔ correlating treatment to specific SNP = specifically tailored treatment for individuals

- Nearly every gene and haplotype block is marked by a SNP(s)
  → absence of SNP on haplotype block can be used to flag presence of nearby defective genes
    ➔ helps assess the risk that someone will develop a particular disease
  → comparing patterns and frequencies of SNPs in patients and normal people
    ➔ identifies which SNPs are associated with which disease (genetic linkage analysis)

Personalised medicine

- The goal is to customise medical care for each individual
  → will depend on technological advances e.g. SNP arrays and ultra-fast DNA sequencing
- Predicting susceptibility to complex disease
  → e.g. type II diabetes, obesity
- Diagnosing complex disease or syndromes
  → e.g. schizophrenia, autism

- Classifying disease
  → e.g. tumour typing
- Managing disease by individualised treatment via pharmacogenetics
- Cure disease by gene therapy

- Pharmacogenetics = study of relationship between genetic variation and response to medications
  → individuals will react/respond differently to drugs
    ➔ may require differing doses
    ➔ may be more/less susceptible to side-effects

- Producing susceptibility/risk profiles for broad range of diseases and treatments for individual
  → reference map of SNPs
    ➔ sequencing >100 individual human genomes to have a 95% confidence that all SNPs occurring at 1% or greater are mapped
    ➔ developed rapid (and cheap) screening methods to map at least 10,000 SNPs in a patient

SNP microarrays

- HapMap describes the common patterns of human sequence variation
  → characterised 600,000 SNPs i.e. 1 SNP per 5 kb of genome sequence
  → allowed for the development of SNP microarrays

- SNP arrays are used to detect polymorphisms within a population
  → type of microarray = set of DNA sequences representing entire set of genes of an organism, arranged in a grid pattern for genetic testing
  → gene chip contains thousands of spots, each containing single-stranded 25 base reference DNA (oligonucleotides)
  → each reference DNA is complementary to an SNP allele
- Genomic DNA to be tested is fragmented
  → amplified as a single strand
  → labelled
  → put on the chip
  → binding (hybridisation) conditions are strictly controlled
    ➔ favours perfect matching between probe DNA and chip DNA

Clinical example

- Blood clotting factors are γ-carboxylated on Glu residues in a process that requires vitamin K
  → Vitamin K epoxide reductase (VKORC1) is the key enzyme

- Non-carboxylated factors are inefficient for blood clotting so inhibition of VKORC1 prevents blood clotting disease (thromboembolic disease)
  → Coumarins (warfarin) used to inhibit VKORC1 and prevent blood clotting
    ➔ too much leads to serious side effects such as bleeding
  → VKROC1 SNP haplotypes have been screened for correlation to warfarin sensitivity
    ➔ e.g. different races show different dose-response
    high African > Caucasian > Chinese *low*
  → a SNP in the VKORC1 promoter region is associated with a low warfarin dose requirement

- Clinically, a safe warfarin dose can be predicted by determining a patient's VKORC1 SNP haplotype

- SNP analysis is powerful but does not give all genetic information on an individual
  → e.g. variations not represented in arrays, like INDELs and mobile elements
- Sequencing the entire human genome will give a complete genetic information on an individual