Lecture2: High throughput sequencing and Read Mapping

▼ Briefly describe how Sanger sequencing for genome works.

▼ You take the same single stranded DNA and divide in into four test tubes. Each test tube contains all dNTPs + one of ddATP, ddCTP, ddTTP and ddGTP.

▼ When the DNA extends, if a dNTP is incorporated, the strand continues to grow, but if a ddNTP is incorporated, then the strand will terminate. So each test tube will have fragments of DNA of different lengths where the corresponding ddNTP has been incorporated

▼ The sample from each tube is then run in a single lane in a gel and the bands are read to deduce the DNA sequence.

▼ Briefly describe how Next generation sequencing works.

▼ Step 1: Modify the DNA: We fragment the DNA in the sample and add adpater fragments to both ends.

▼ Step 2: Clustering: The adapter groups are used to capture the fragments on a surface - a flow cell. The fragments are multiplied in place so that clusters of similar fragments form in each location where a particular fragment is attached.

▼ Step 3: Sequencing by synthesis: The surface is washed with fluorescent nucleotides such that only one nucleotide attaches. The surface is imaged usign 4-color imaging. The dye is cleaved so that nucleotides are free to extend again and the process is repeated by adding one fluorescent nucelotide, imaging, cleaving and repeating again.

▼ Why is NGS called massively parallel sequencing?

▼ Because NGS can sequence millions of DNA fragments at a time. The time taken to sequence is proportional to the length of the DNA fragment being

sequenced, not on the total number of fragments.

▼ Briefly describe how third generation single molecule sequencing works.

▼ Technology still under development, but existing technologies use polymerase and/or pores for sequencing DNA without amplification.

▼ It is not good for read count applications, but good for sequencing long DNA strands.

▼ What is the format of the raw data coming from NGS machines?

▼ FASTQ file

▼ Which type of data does a fastq file contains?

▼ Sequence ID

▼ Sequence

▼ Quality ID

▼ Quality score - uses ASCII of: sequence quality + 33

▼ What is Phred quality score and how is it calculated?

▼ Phred quality score is a metric used to indicate the measure of base read quality in next generation sequencing. The bigger the number, the better the quality

▼ Phred score of 20 denotes the likelihood of finding 1 incorrect base among 100 bases.

▼ Which tool can you use to check the quality of sequencing data?

▼ FASTQC

▼ What are the types of quality scores you can look at in FastQC report?

▼ Per base sequence quality score: If the average quality score across all reads at each base pair is good and the quality does not drop too much too quickly. Dropping around 40/50 bp is fine, dropping around 20bp is a problem.

▼ Per sequence quality distribution: For each sequence look at median quality score along the read. quality score on the x-axis. Most reads should

have a high quality score.

▼ Nucleotide content at each location: unbalanced ACGT is a problem. The graph should get smooth quickly.

▼ Per sequence GC content: Should follow a roughly normal distribution.

DNA sequence mapping algorithms

▼ What categories of DNA sequence alignment algorithms exist based on number of query and search sequences and which algorithms belong to these categories?

▼ Mapping 1 read to 1 read - Global alignment algorithm such as Needleman-Wunsch and Local alignment algorithm such as Smith-Waterman

▼ Mapping 1 read to many reads - BLAST which breaks a read into k-mers (all possible sequences of k-nucleotides) and indexes them, then matches k-mers of a query algorithm to stored indexes of all database sequences.

▼ Mapping many to many reads - such as Burrows-Wheeler Alignment (BWA) which uses burrows wheeler transformation to store the database sequences and LF mapping to find a match and retrieve the original sequence.

▼ Which popular bioinformatics tool uses suffix arrays for alignment algorithm?

▼ STAR - a tool used for RNA alignment uses suffix arrays.

▼ [WIP] What are suffix arrays?

▼ My understanding based on lecture is that suffix arrays are indexes built in computer's memory of a sequence and stored as a type of hash table (a combination of arrays and linked lists). The lecture explanation is not sufficient to understand what these really are or how the algorithm for searching suffix arrays works.

▼ Which popular bioinformatics tools use burrows-wheeler alignment algorithm?

▼ bwa and bowtie both use Burrows-Wheeler alignment algorithm for DNA sequence alignment

▼ What is the output of a DNA alignment algorithm such as Burrows Wheeler Alignment?

- ▼ A SAM or BAM (binary version of SAM) file

▼ What information does a SAM file contain?

- ▼ The data information contains the following:
  - ▼ original sequence and quality: Sequence ID, Sequence, Sequence quality,
  - ▼ Match information: whether match found (0) or not (4) or found to reverse strand (16), and mapping location if match is found - the chromosome number and location where match found.
  - ▼ Mismatch information: number of mismatches and where those mismatches occurred in the query sequence
- ▼ The header contains information about:
  - ▼ Reference genome used: which genome was used as reference, how many chromosomes it contained and what is the length of those chromosomes
  - ▼ sequencing platform information
  - ▼ Read group information - to collate information if same sample is run in multiple lanes. This information is important for normalization
  - ▼ Information about the alignment tool/algorithm used

▼ What percentage of sequences being unmapped to a reference human genome is acceptable in a typical run?

- ▼ 10-30% of sequences being unmapped is common to see (reference?)

▼ What tool can you use to visualize a SAM file?

- ▼ UCSC or Interactive Genome Viewer (IGV)

▼ What is a BED/BigBED file?

- ▼ A BED file is typically used to store genomic interval information, i.e., interesting information about certain locations in the genome. It can be used to store alignment information but is rarely used for that.

▼ Each row contains information about a location in the genome (chromosome number, start, end, strand) and some interesting information about that location such as a peak in CHIP-seq data.

▼ There is some information loss in BED compared to SAM files.

▼ BigBED is the binary version of a BED file.