Genome Assembly: 07.02.2019

Ben Langmead John Hopkins Dept of computer science -> youtube tutorial

Refrence genome + I have the reads of my species ->

Reads are DNA sequences -> overlapped using algorithms -> overlaps make string graps -> those help you produce a contig. These are aligned and put in order putting scaffold.

Graph: Comprised of nodes and edges (nodes are ppls and edges are connections between people).

De novo assembly (without reference genome) ->

Raw data -> FastQC report -> trim and filter data (remove adapter).

**Coverage:** Either coverage at a given position. **Average coverage** = total length of reads / total length of genome.

Assembly:

**Overlapping reads:** See if suffix of read 1 is similar to read 2 ->

Why are overlapping reads different? -> heterozygocity (can get paternal or maternal errors) or sequencing error.

**K-mers:** Build a k-mer index like an index of a book. Key terms in alphabetical order with page numbers -> Can look certain strings up

**Index:** Often called a hash table often sorted.

If there is an underrepresented error it might be a sequence error. -> maybe I can toss it.

**De Bruijn graph:** Main algorithm uses edges and nodes (nodes are k-mers not reads).

Reads: Paths in the graphs.

Assembly: Path through graph that visits evert edges.

Method is sensitive to sequencing errors.

Graph emited: You do something and it’s the output that comes from that.

Optimize k-mer length in some way or swap in different increments to see if we get different results -> even k-mers you end up with the reverse complement of reserve sequences and you can have palyndromes. -> **use ODD k-mers**

**Reads** -> K-mer index -> often tips are removed (wiki velvet assembler) -> node can generate a bubble. -> pop bubbles within oyur assembly.

Human genome is 50% repetitive.

**Repeat structure is important in relation with given read length.**

Restriction enzyme: often cleave the double stranded DNA, often just knit the double stranded DNA.

QC Scores: Need to sequences numeral molecules at a time.

Gene evolution of immune gene and olfactory genes in bats -> a few reference bat genomes ->

Do a literature review with an undergrad -> Or ask a M.Sc. student -> Ask Yale library how to do a literature review ->

**Rad seq:** relies on enzyme recognizing small sequences 8bp,they cleave seq at that point -> chop the genome and get eeything between 500-800bp. -> indexpensive way to sequence -> very common for pop lvl studies -> good to have reference genome.

**Exoms:** portion of genome that codes protein DNA.

**Tutorial on type along and use computational skills -> reinforce skills as we know it.**

Notes on syllabus on sequence assembly.