Problems in Genome Assembly

C. Ryan Campbell

Duke University

c.ryan.campbell@duke.edu

24 Oct 2017

Overview

1. Genome Assembly
   - The Old Way
   - The New Way
2. What is the best approach?
3. Assessing Genome Quality
   - Activity
4. The New(est) Way
   - BioNano
   - 10X
   - Hi-C
5. What problems remain?

Today's Goals

- Why is assembly important?
- What are the differences in long and short read platforms?
- How do these differences impact assembly?

Genome Assembly

- Mapping (or assembling) all the bases in the genome of an organism
- List all the bases for each chromosome
- Annotate the genes and exonic/intronic regions
- Why do we want to do this?
- Is this assembly for an individual or a species?

Assembling the Human Genome

- Was once limited by the throughput of Sanger sequencing
- Genomic DNA was separated by chromosome
- Cloned into bacteria
- Sequenced little by little

Shotgun Sequencing

- Eventually we moved on to “shotgun” sequencing
- Randomly sequence as much of the genome as possible
- Assemble into a complete genome
- What are the possible issues with this method?

Next Generation Sequencing (NGS)

- Still technically a “shotgun” method
- NGS produces a lot more data than Sanger
- Think of it as many shotguns
- What are the types of NGS we’ve talked about?

Illumina - Short Reads

- “Short” reads - 100-250 bp
- Sequence as a paired-end read
- Can cover up to 500bp of an 800bp fragment
- Accurate and cheap

PacBio - Long Reads

- “Long” read - greater than 1,000 bp
- Sequence fragment in a loop (shorter fragments get sequenced at higher coverage)
- High end now up to 20,000 bp
- Less accurate, more expensive per base

What is the best approach?

- Given a finite budget should you:
  - Use a lot of “short” reads?
  - Use fewer “long” reads?
  - Use even less of a mix of both?

How do we know what is “best”

- How is genome assembly quality measured?
- What makes a “good” genome?

How do we know what is “best”

How do we know what is “best”

- Quantitatively (N50 scores)
  - How much of the expected genome have you assembled?
  - How long are the pieces of the genome you have assembled?
- Qualitatively (BUSCO)
  - Are the genomic features we expect to be there present?
  - Look for commonly conserved genes
  - Benchmarking Universal Single-Copy Orthologs

Common Terms

- Contig
  - A run of contiguous bases
  - No gaps allowed
- Scaffold
  - A set of contigs scaffolded together with gaps of N's
  - Must know how far apart the contigs are

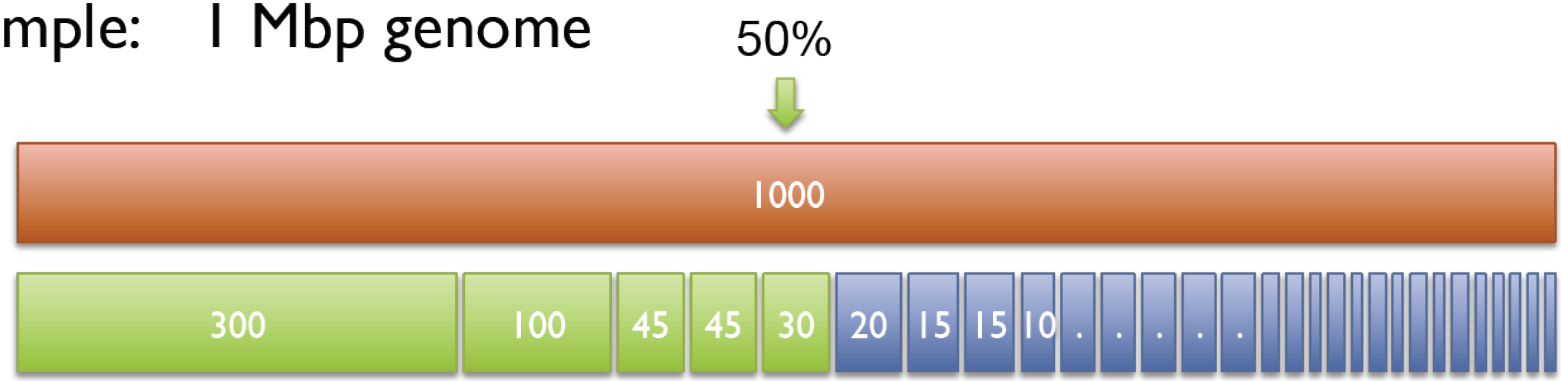
Scaffold N50

- Measure of genome assembly quality
- Arrange the scaffolds from longest to shortest
- N50 = Length of the scaffold at 50% of the total genome size
- What is a perfect N50?

N50 size

Def: 50% of the genome is in contigs larger than N50

Example: 1 Mbp genome



N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:

N50 values are only meaningful to compare when base genome

The Problem

- Common Assembly Problems:
- Size - 3bil bp = 5,000 books
- Chromosomal Arrangement - not just “base 1-3,000,000,000”
- Repeats - tricky regions that are common in mammals

Your Directions

- Assemble the sequence data your group has been given
- Should make a coherent paragraph, taken from a speech
- The “data” are randomly generated reads:
- Long - 75 characters, error prone
- Short - 16 characters, accurate

Your “Output”

1. Estimate paragraph length
2. Estimate the number of repeats
3. Estimate the proportion of errors
4. Estimate your contig N50

- IMPORTANT: “|” character is the end of the read
- 15-20 mins

The Answer

```
Even though large tracts of Europe and many old and
famous States have fallen or may fall into the grip of
the Gestapo and all the odious apparatus of Nazi rule,
we shall not flag or fail.  We shall go on to the end.
We shall fight in France, we shall fight on the seas
and oceans, we shall fight with growing confidence and
growing strength in the air, we shall defend our
island, whatever the cost may be.  We shall fight on
the beaches, we shall fight on the landing grounds, we
shall fight in the fields and in the streets, we shall
fight in the hills; we shall never surrender.
```

The Answer

Even though large tracts of Europe and many old and famous States have fallen or may fall into the grip of the Gestapo and all the odious apparatus of Nazi rule, we shall not flag or fail. We shall go on to the end. We shall fight in France, we shall fight on the seas and oceans, we shall fight with growing confidence and growing strength in the air, we shall defend our island, whatever the cost may be. We shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills; we shall never surrender.

470 chars, 581 w/spaces 11 repeats

The New(est) Ways

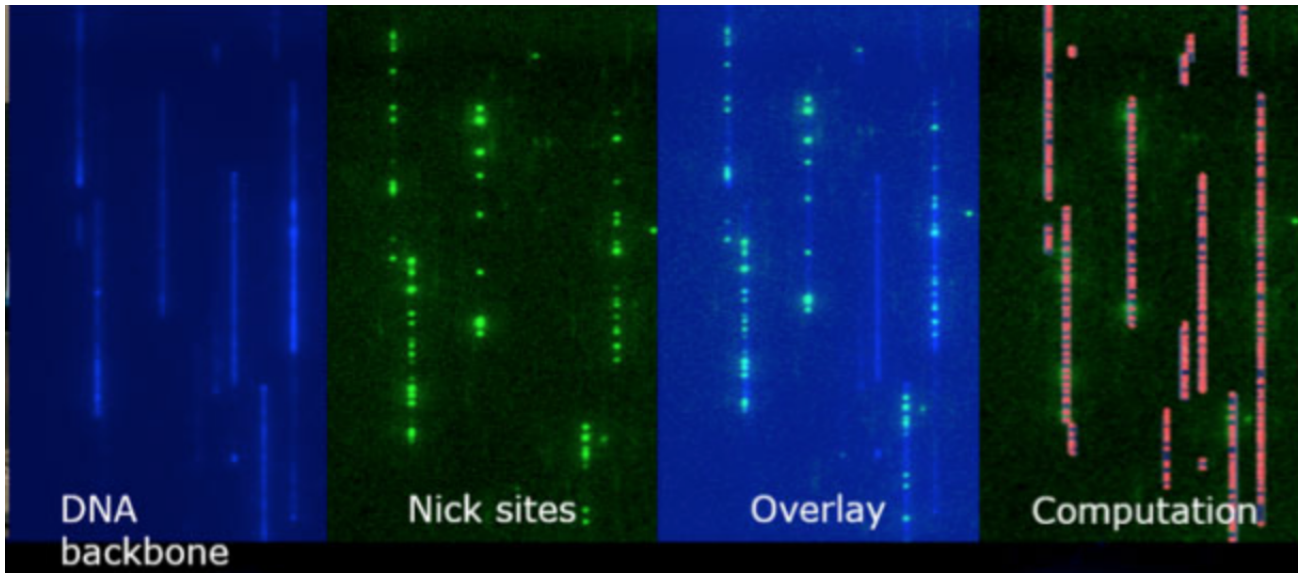
- Bio-Nano
- 10X Genomics
- Hi-C

BioNano Genomics

- Technology to “super”-scaffold long stretches
- Prepare high quality DNA, keeping as much of the chomosome intact as possible
- Flourescence-tag sites with a known 8bp sequence
- Use this as a backbone to build longer scaffolds

BioNano

- Collect long DNA
- Label known 8bp sites
- Image

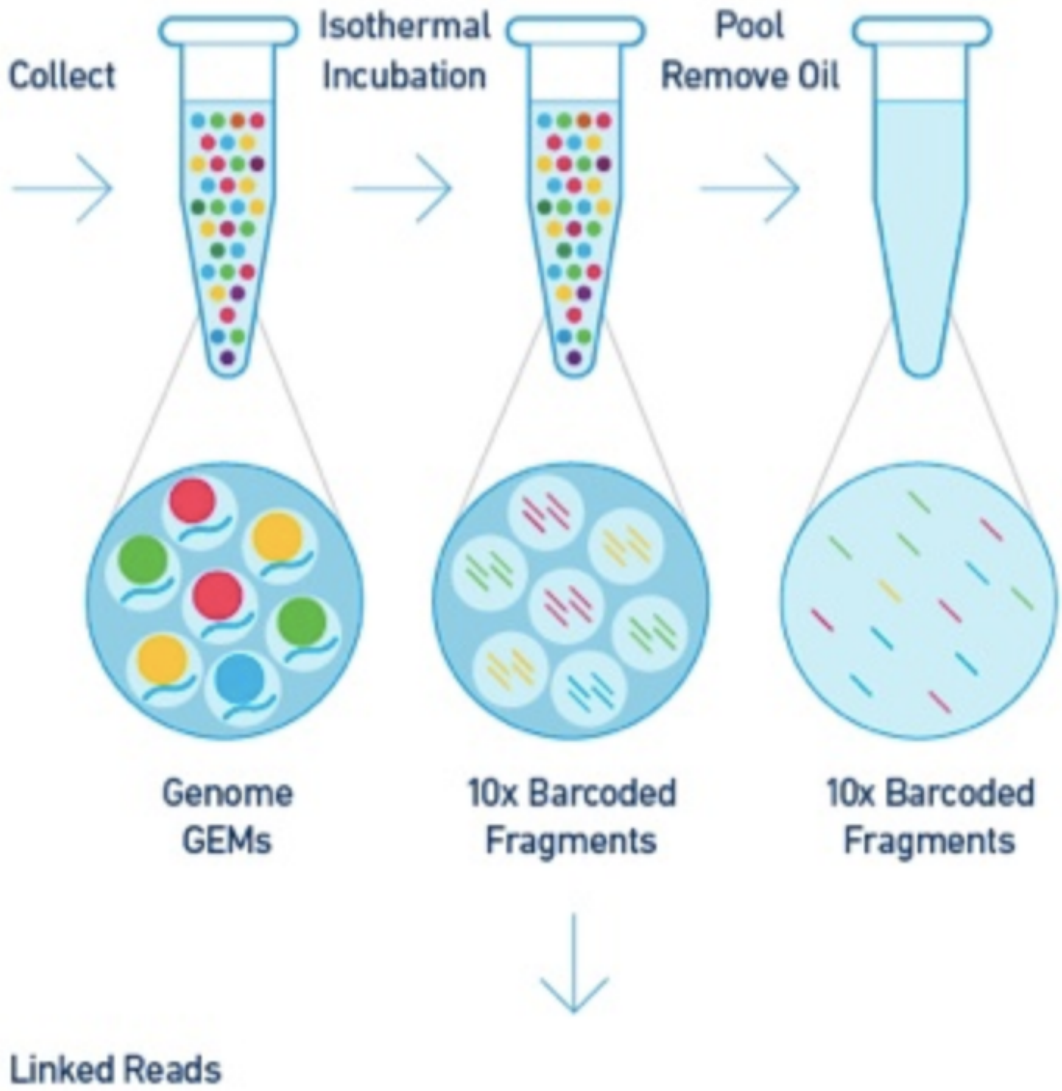


10X Genomics

- Built on Illumina technology
- Within oil droplets - barcode large DNA (100kb) and then fragment
- Sequence as usual
- Use the barcodes to successfully rebuild the 100kb pieces
- Simulates long read PacBio tech with Illumina

10X Genomics

- Label 100kb frags
- Barcode within beads
- Seq and reassemble

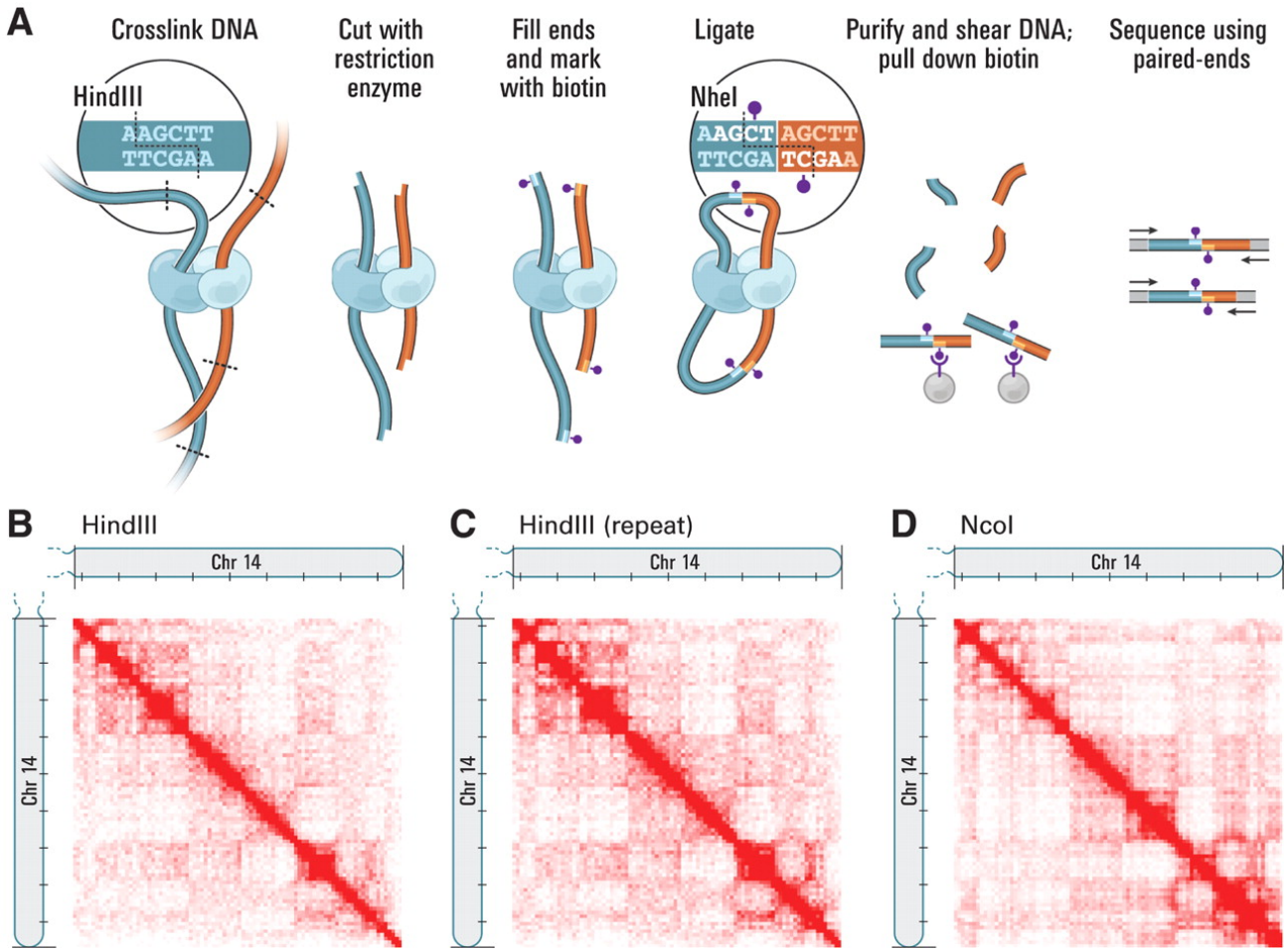


Hi-C

- Within a nucleus link DNA as it is found naturally
- Connects neighboring DNA sites, even across chromosomes
- Fragment and sequence, maintaining these linkages
- Captures a snapshot of chromosomal interactions

Hi-C

- Link nuclear DNA
- Fragment
- Sequence across linkage



Hi-C

- Two main uses:
- 1) What parts of the genome are interacting with others?
  - Open chromatin regions
  - Interactions across chromosomes
- 2) Which fragments are on the same chromosome?
  - Much of the data is within chromosome
  - Can use to improve assembly

Remaining Problems

- Centromeres - very long and oddly repetitive
- Cost - still expensive, especially to get structural variation
- Chasing down humans - there are always techniques and features that are available to human biologists that other model organisms (and non-model organisms) are pursuing

The End