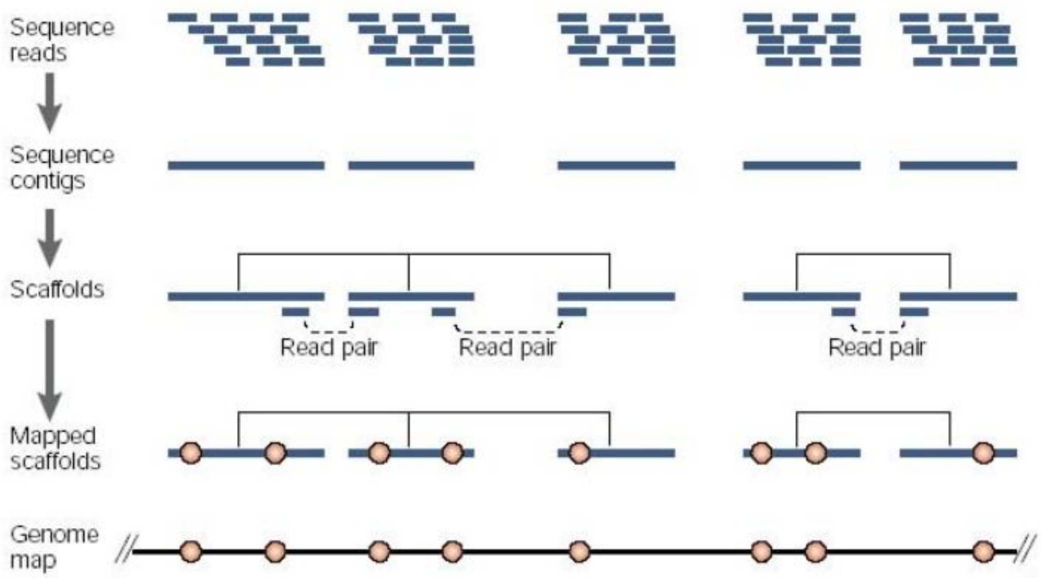
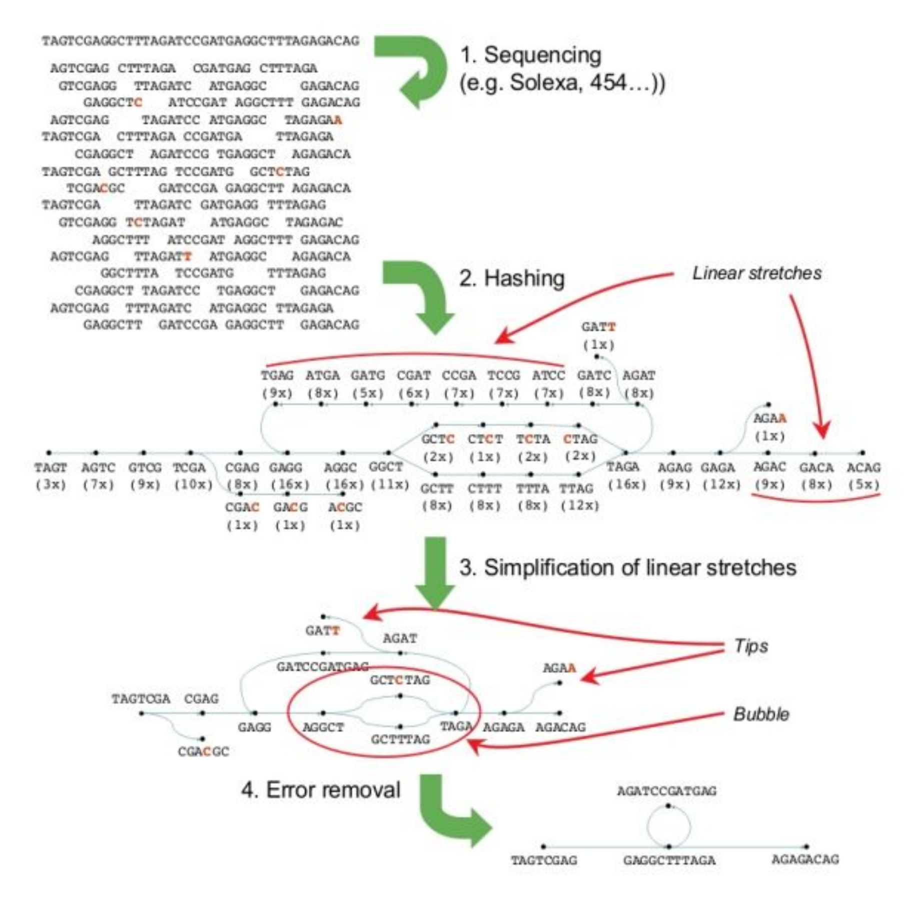
Scaffolding

de novo whole-genome shotgun assembly



Contig construction



Scaffolding

Both OLC and DBG are concerned with constructing the longest, most accurate *contigs* possible

Scaffolding orders and orients *contigs* with respect to each other

For this we can use data from various sources, especially paired ends.

Contig: is a stretch of unambiguously assembled sequence.

Scaffold: may contain gaps.

Paired-end sequencing

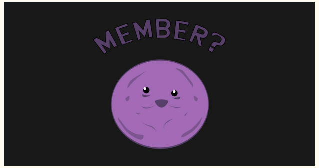
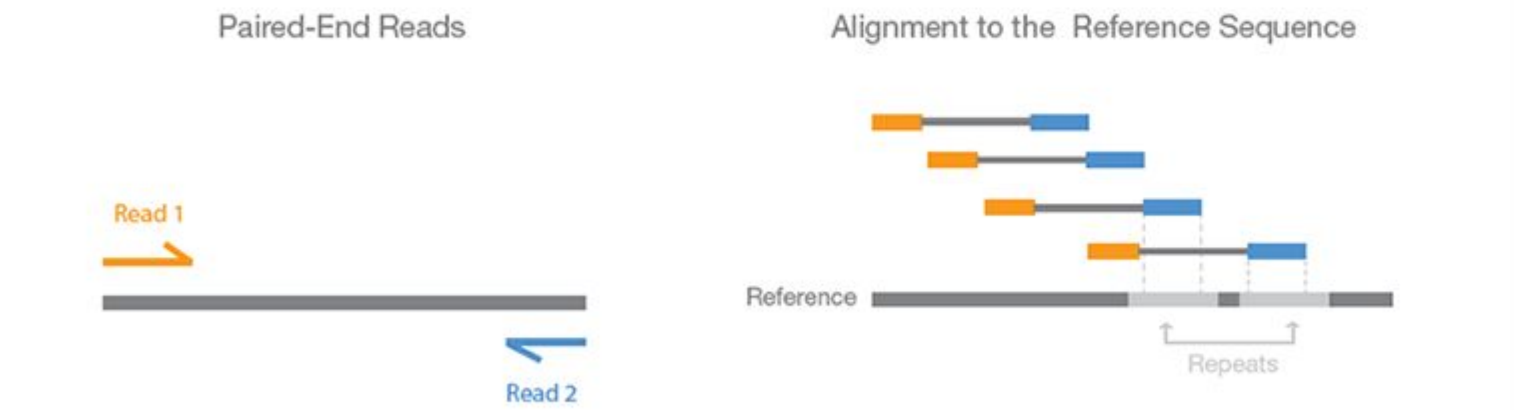


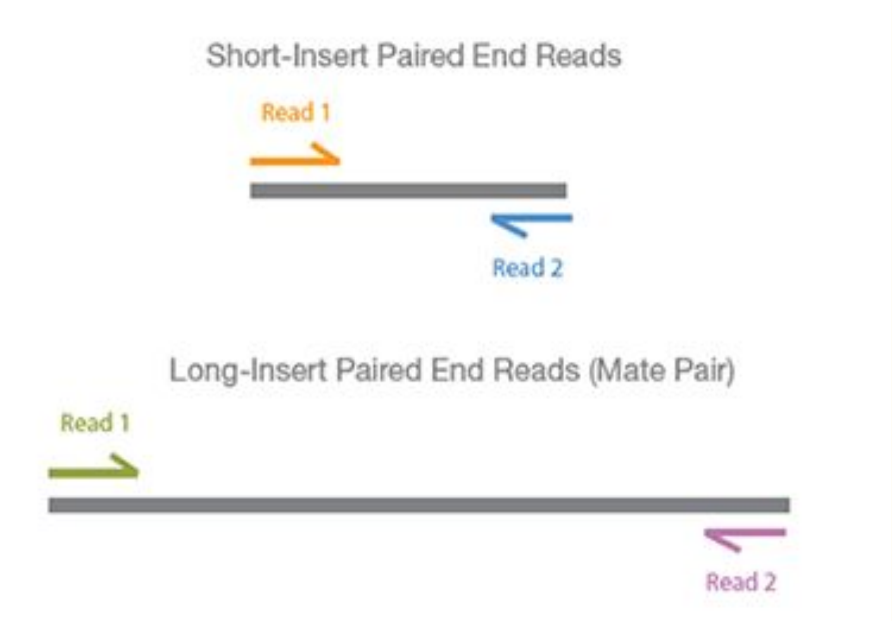
Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Vocabulary

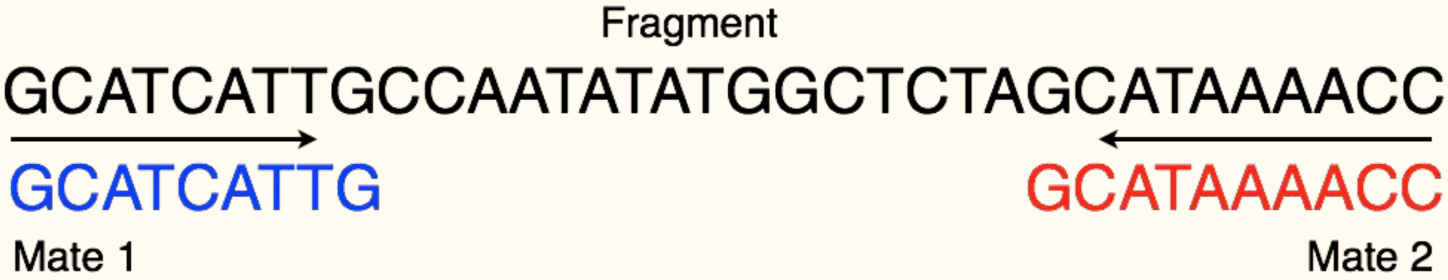
- **Paired end reads:** read1, insert $<$ 500 bp, read2
- **Mate pair reads:** read1, insert $>$ 1 kbp, read2
- **k-mer:** any sequence of length k
- **Contig:** gap-less assembled sequence
- **Scaffold:** sequence which may contain gaps


Short-Insert Paired End Reads
Read 1
Read 2
Long-Insert Paired End Reads (Mate Pair)
Read 1
Read 2

Scaffolding: paired-end sequencing

Alternative protocol produces a *pair* of reads taken from either end of a longer fragment

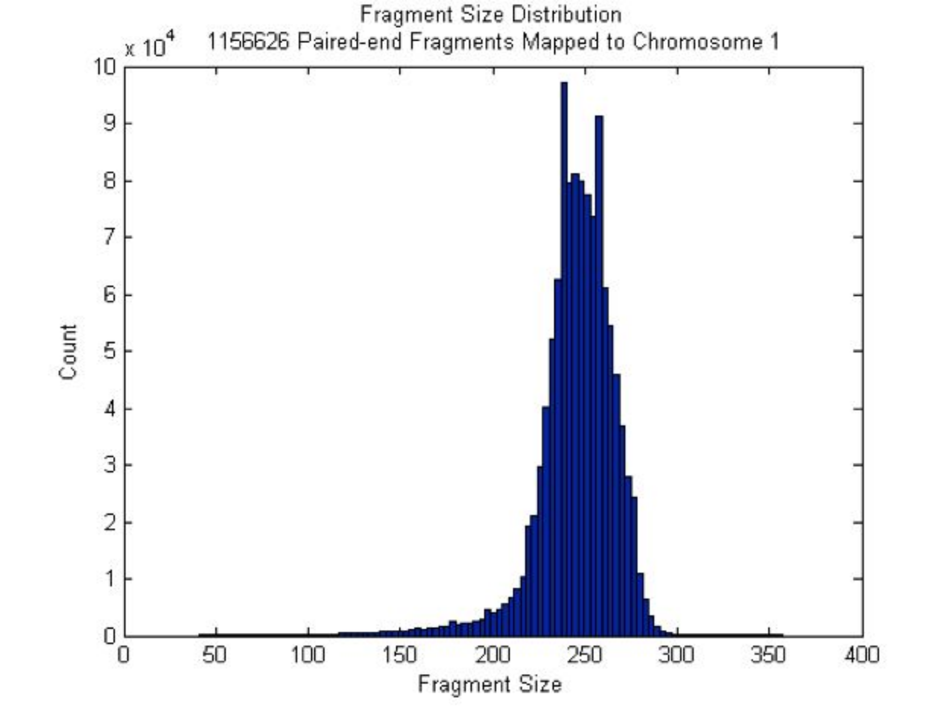
Paired reads are also called *mates* to distinguish them from the *unpaired* reads we've been discussing



Depending on lengths, mates might overlap in the middle of the fragment

Scaffolding: paired-end sequencing

Example fragment length distribution

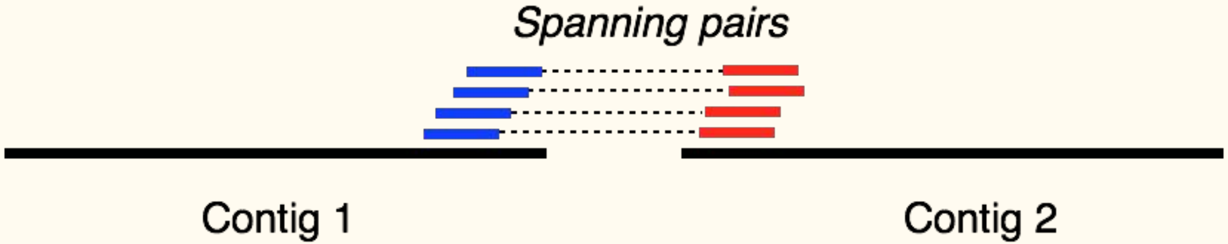


Fragments are not exactly the same length, but there's a clear peak around 250 nt, very few < 150 nt or > 300 nt

Scaffolding: paired-end sequencing

Say we have a collection of pairs and we assemble them as usual.

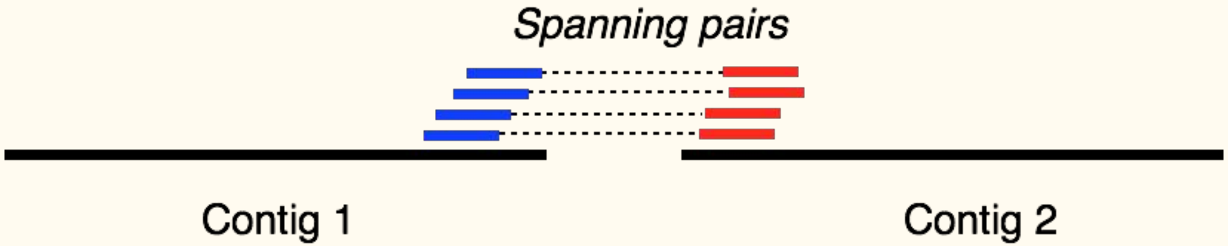
Assembly yields two contigs:



...and we discover that some of the mates at one edge of contig 1 are paired with mates in contig 2

Call these *spanning pairs*

Scaffolding: paired-end sequencing

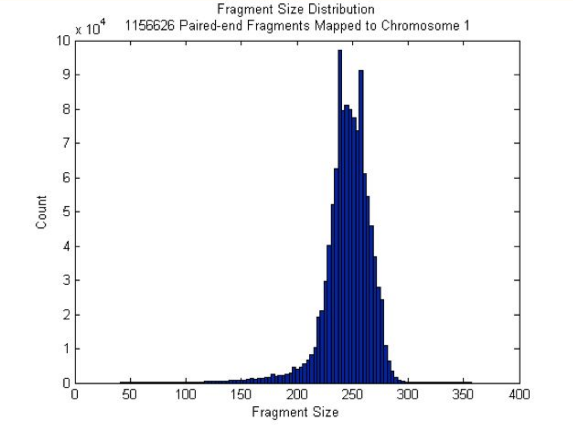


What does this tell us?

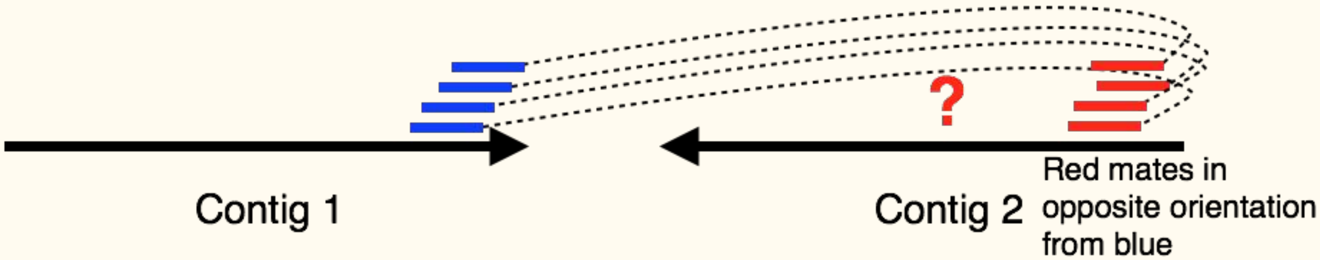
Contig 1 is close to contig 2 in the genome.

In fact, we can *estimate distance between contigs* using what we know about fragment length distribution.

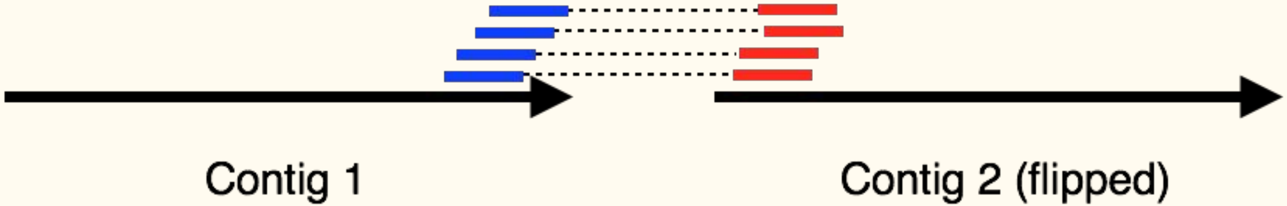
The more spanning pairs we have, the better our estimate.



Scaffolding: paired-end sequencing



What does the picture look like if contigs 1 and 2 are close, but we assembled contig 2 “backwards” (i.e. reverse complemented)



Pairs also tell us about contigs’ relative *orientation*

Scaffolding

Scaffolding output: collection of *scaffolds*, where a scaffold is a collection of contigs related to each other with high confidence using pairs.

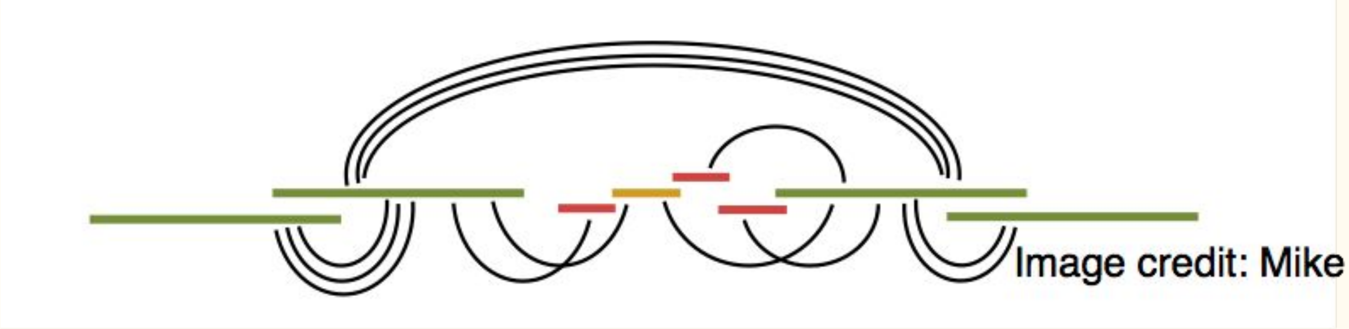
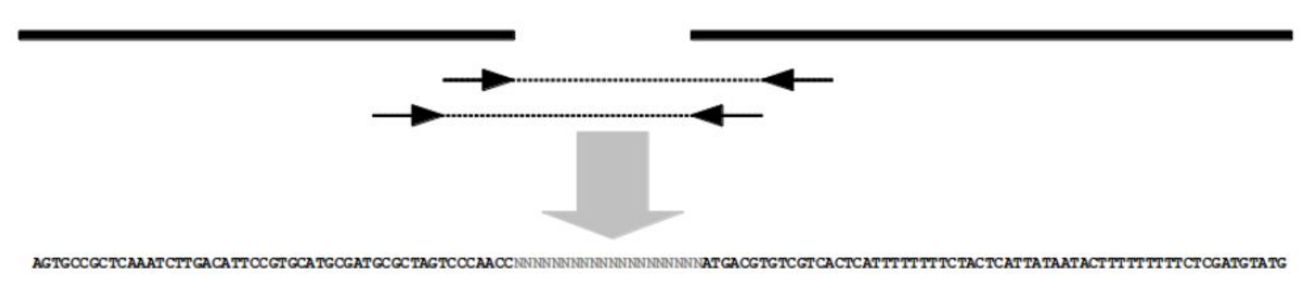


Image credit: Mike Schatz

Scaffolds construction in practice (lab)

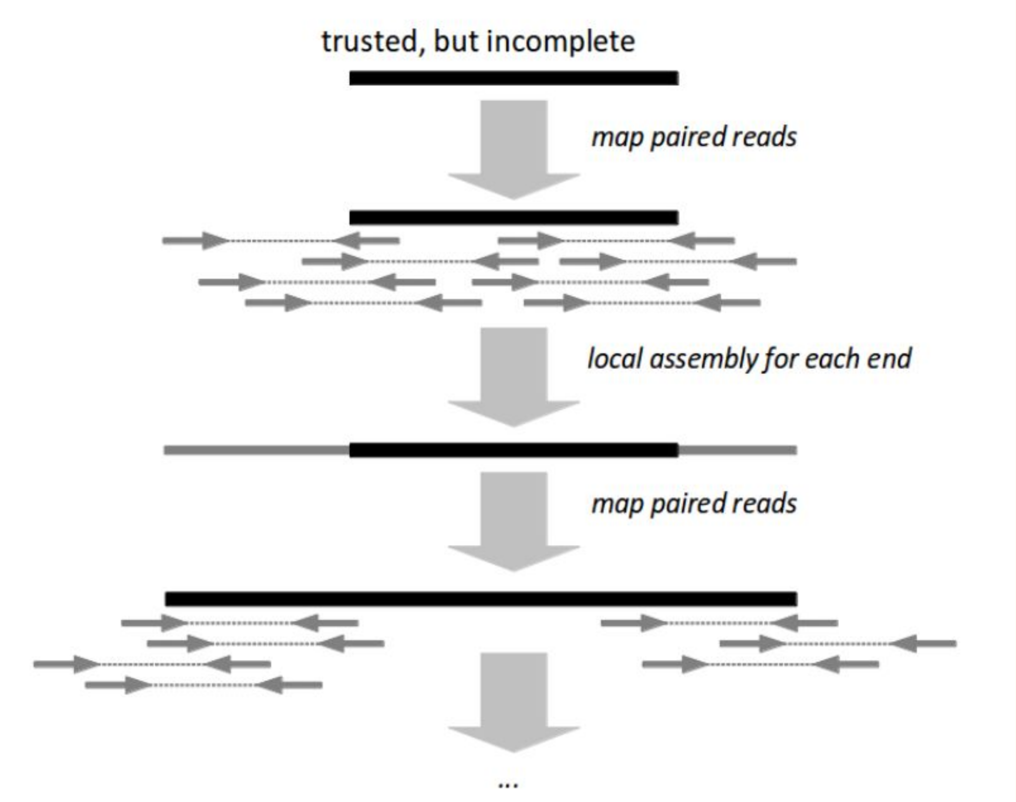
- Scaffolding using pairing information
- This is why libraries containing **multiple insert sizes** are used:
  - Gnerre 2011:
    - 45x overlapping PE reads (insert size: 180 bp, reads: >100 bp)
    - 45x short jump MP reads (insert size: 3 kb)
    - 5x (optional) long jump MP reads (insert size: 6 kb)
    - 1x (optional) fosmid jump MP reads (insert size: 40 kb)
  - Ribeiro 2012:
    - 50x overlapping PE reads (insert size: 180 bp, reads: >100 bp)
    - 50x PacBio reads (reads: 1-3 kb)


AGTGCCGCTCAAATCTTGACATTCCGTGCATGCGATGCGCTAGTCCCAACCNNNNNNNNNNNNNNNNNNNNATGACGTGTCGTCACTCATTTTTTTTCTACTCATTATAATACTTTTTTTTTCTCGATGTATG

Scaffolds construction

- Scaffolding using pairing information
- Techniques:
    - Jumping libraries
    - Linked reads (10x Genomics, Dovetail Genomics - Chicago libraries)
    - Long reads (PacBio, Oxford nanopore)
    - Structural maps (BioNano)

Gap closing / contig extension


trusted, but incomplete
map paired reads
local assembly for each end
map paired reads
...