Assembly in Practice: Part 3: Scaffolding

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

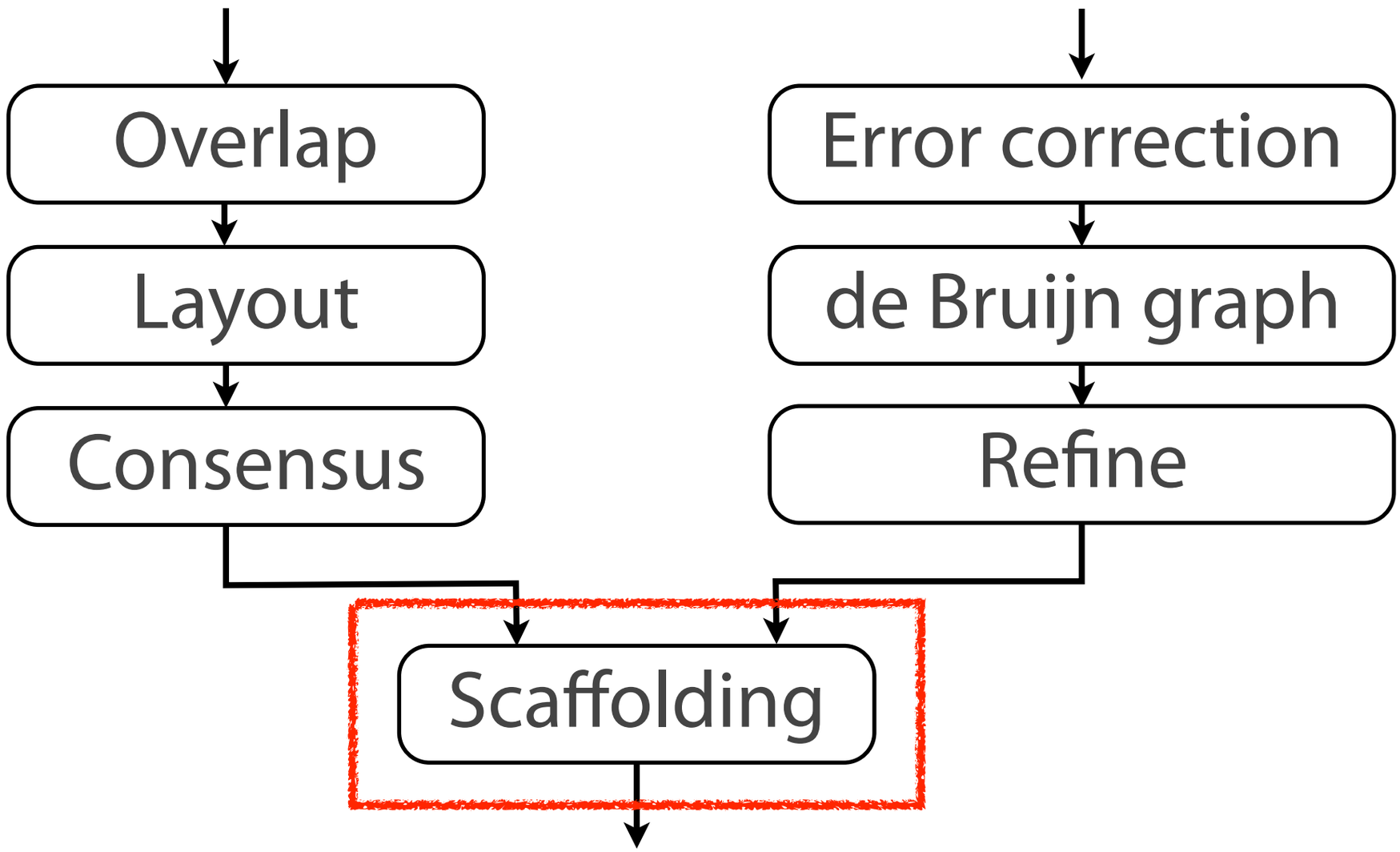
Department of Computer Science

Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

Assembly paradigms

1: Overlap-Layout-Consensus (OLC) assembly

2: de Bruijn graph (DBG) assembly



Scaffolding

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

Contig is a stretch of unambiguously assembled sequence

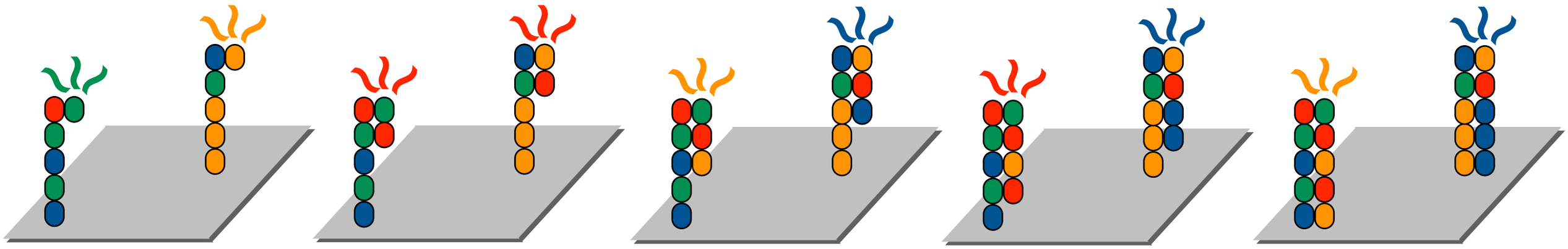
Scaffolding *orders* and *orients* contigs with respect to each other

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

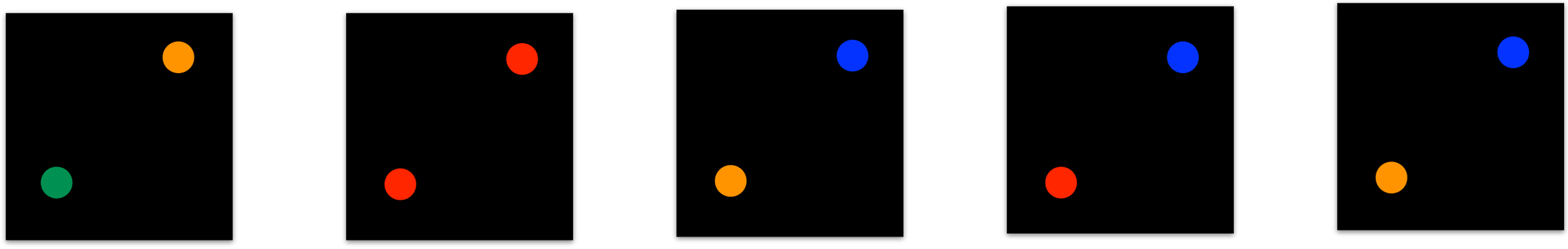
Scaffolding: paired-end sequencing

We discussed sequencing by synthesis

Process we discussed produces one contiguous read sequence



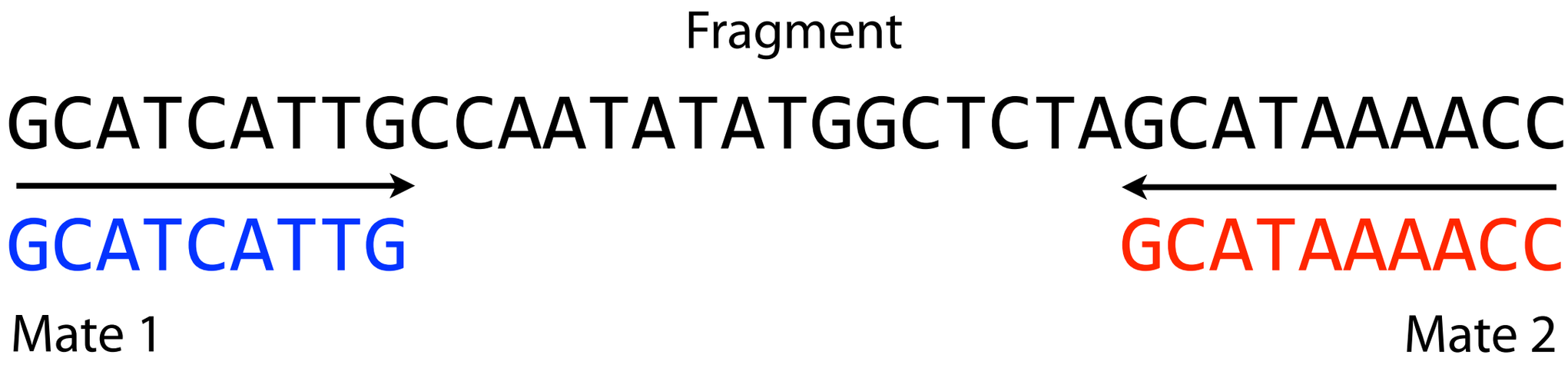
Cycle 1 Cycle 2 Cycle 3 Cycle 4 Cycle 5



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

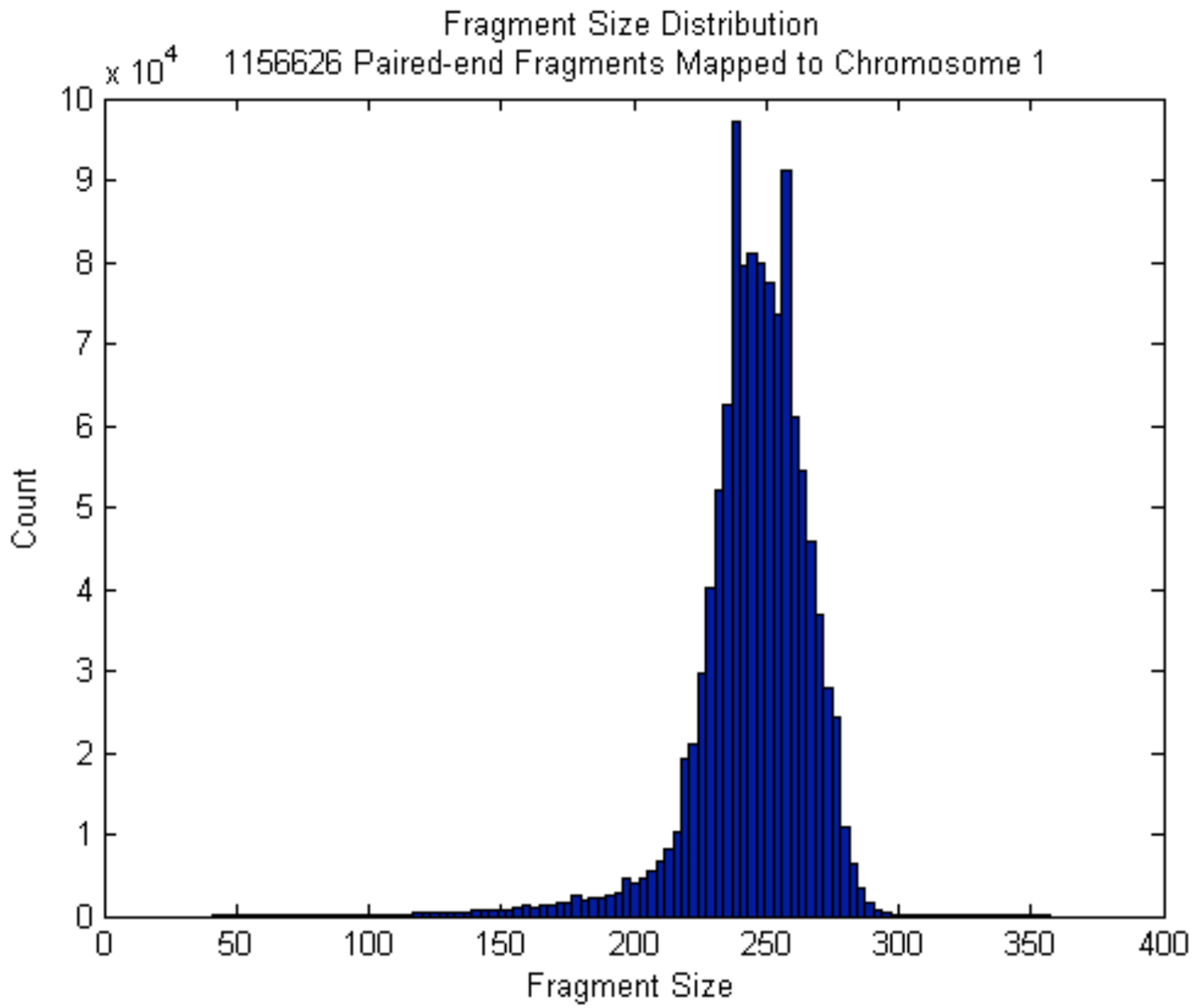


Depending on lengths, mates might overlap in the middle of the fragment

Scaffolding: paired-end sequencing

Example fragment length distribution

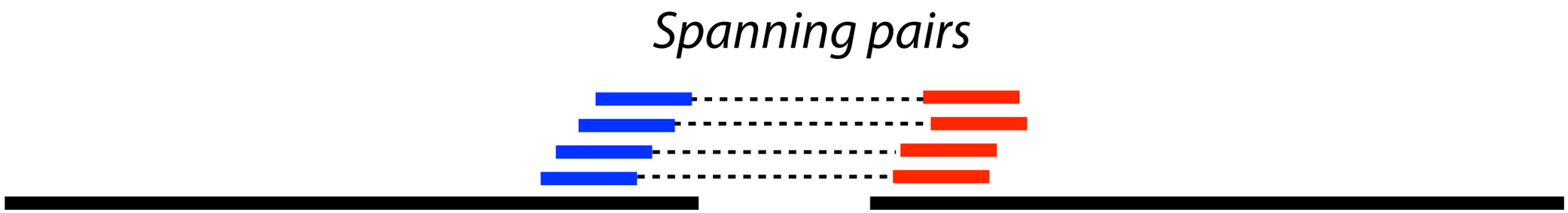
Fragments are not exactly the same length, but there's a clear peak around 250 nt, very few < 150 nt or > 300 nt



Scaffolding: paired-end sequencing

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

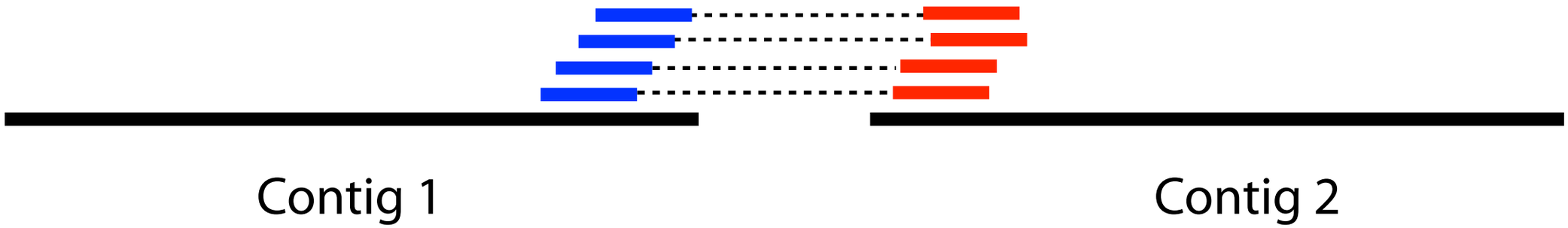
Assembly yields two contigs:



...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

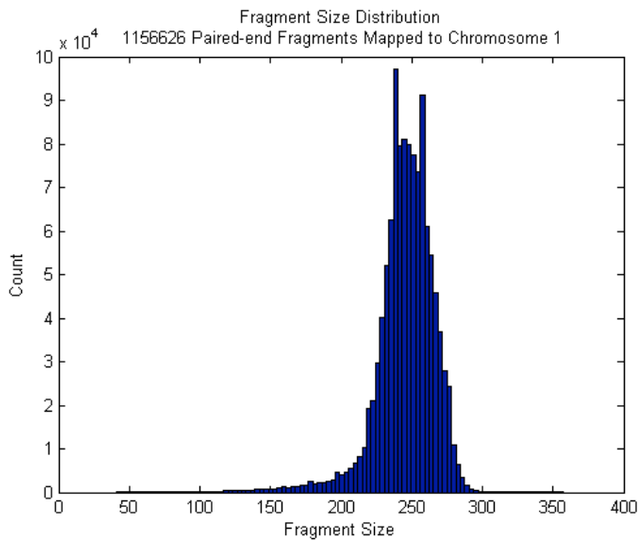


What does this tell us?

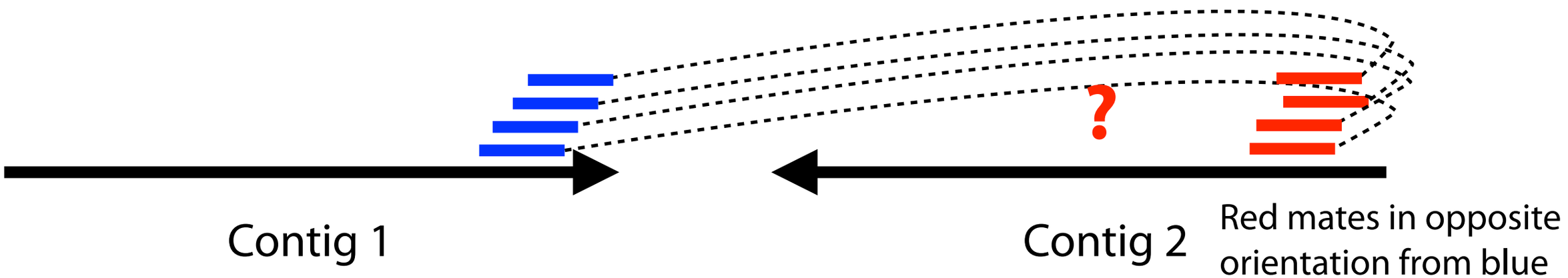
Contig 1 is close to contig 2 in the genome

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

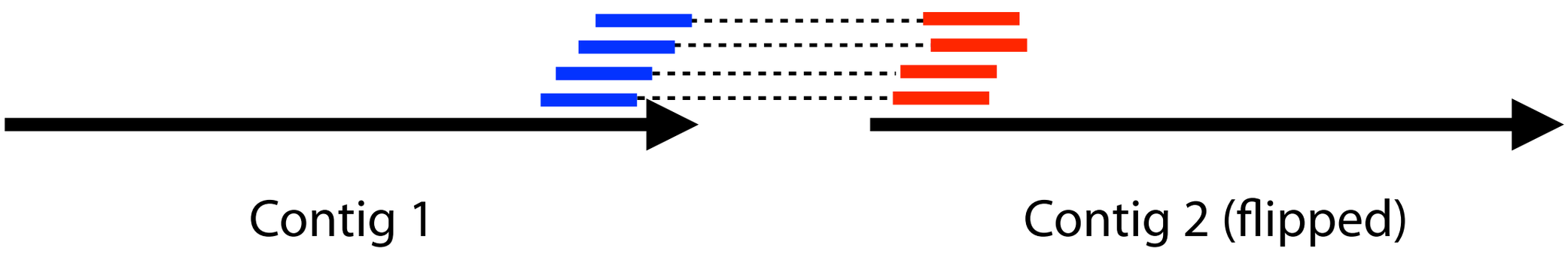
The more spanning pairs we have, the better our estimate



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

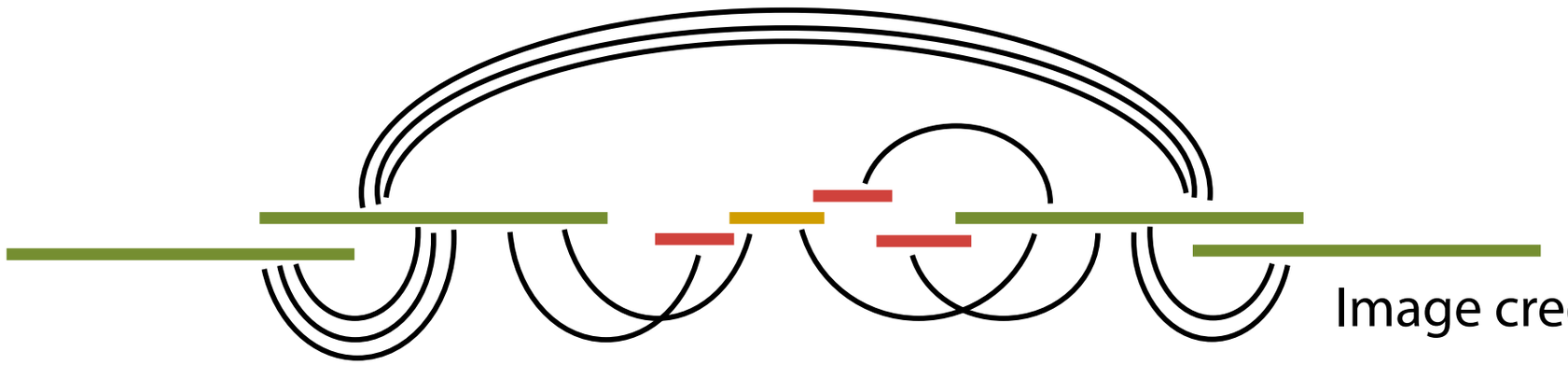


Image credit: Mike Schatz