Read Mapping

4 Main Stages

1 Biological Question

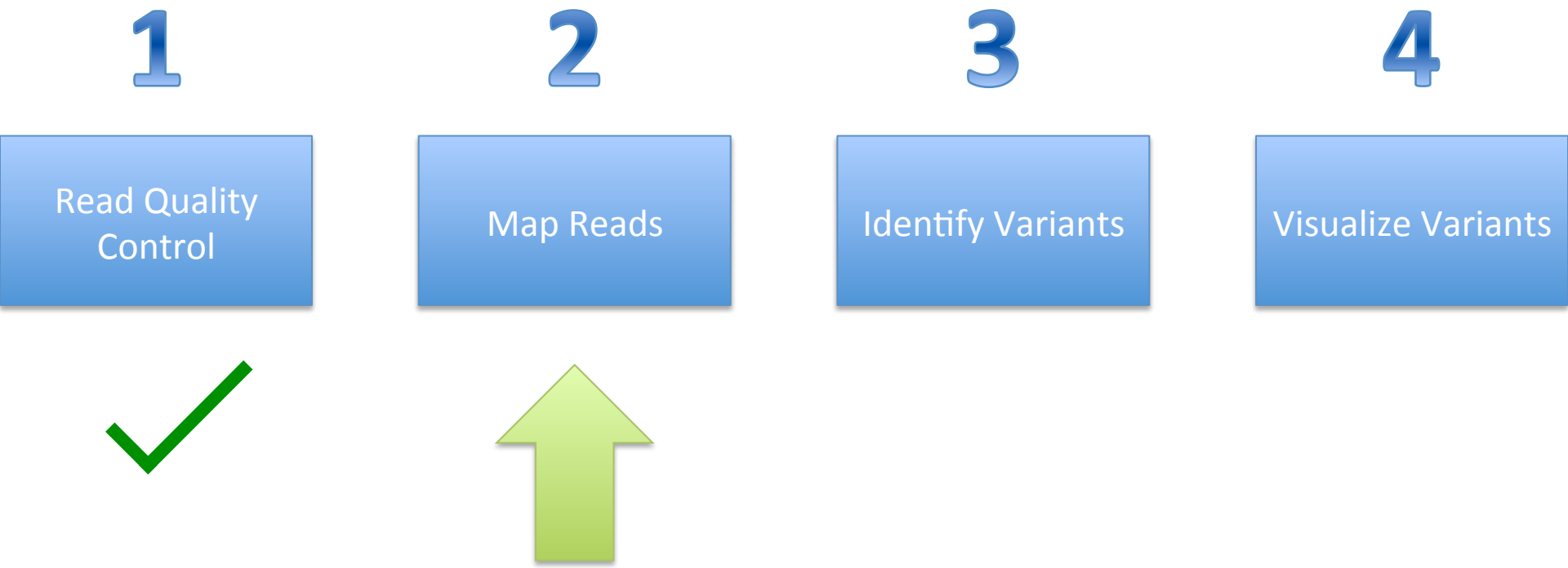
2 Design & Conduct Experiment

3 Prepare NGS Library & Sequence

4 Sequencing Analysis

Class time

4 Typical Stages of Variant Analysis



Why Map Reads?

- Fastq files are listed in an arbitrary order based on where they were located on the flow cell during sequencing. This is not informative in anyway, and checking each read separately for specific information would not be possible.
- Imagine throwing 1,000s or 1,000,000s of copies of Hamlet into a paper shreder, and trying to see how many of the copies said ‘to be not to be’ rather than ‘to be **or** not to be’. If the text was in order an impossible task simply becomes daunting.

What do we map reads to?

- Reference genomes.
- Other reads.

Not all Reference genomes are created equal

1. Accuracy
    1. Your organism
    2. Public reference
    3. Incomplete (contigs)
    4. “similar” organsims
2. Annotations:
    1. Without annotations: know where/what mutation is
    2. With annotations: insight into effect of mutation

Where to find reference genomes

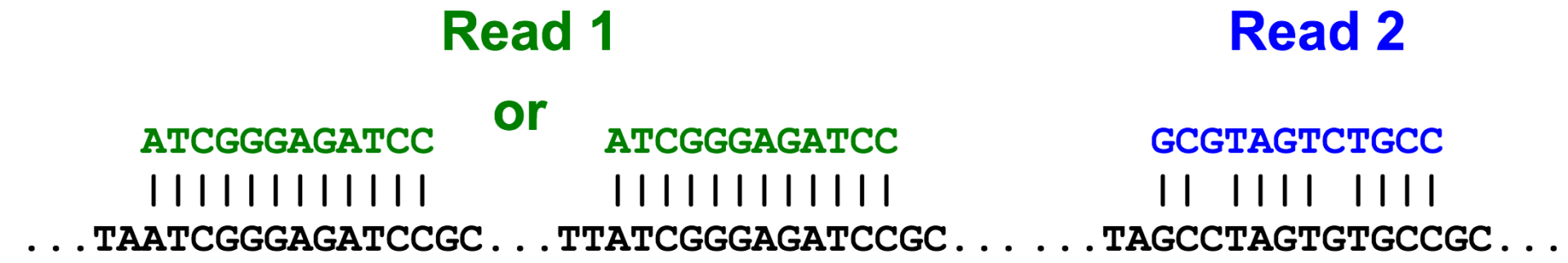
1. Public databases such as NCBI.
    1. This should be your goal whenever possible.
2. Created yourself starting from:
    1. Publicly available reference
        1. The closer the reference is the easier it will be.
        2. breseq can improve references using gdtools apply command.
    2. Next-generation sequencing itself and assembled.
        1. We will optionally cover this towards the end of class.
        2. Exceptionally difficult (though as Oxford Nanopore comes online this will hopefully become easier to even trivial).

How does read mapping work

- Build an index to facilitate “speedy” read mapping.
  - Allows a read to be placed on the reference at a location (**a “seed”**) where it knows at least a piece of the read matches perfectly (or with only a few mismatches).
  - By jumping right to these spots in the reference, rather than trying to fully align the read to every place in the genome, it saves a ton of time.
  - Indexes can be reused for mapping any set of reads.
- Align reads using the index
  - Full reads are aligned around “seed” locations on a base by base basis.

Mapping score vs. Alignment score

- **Mapping score** is the probability that a read is mapped to the correct location in the genome.



- **Alignment score** is how well a read maps to the given

```
ATTTGCGATCGGATGAAGACG..AA
||||||||||||||||XX|||XXX||
ATTTGCGATCGGATGTTGACTTTAA
```

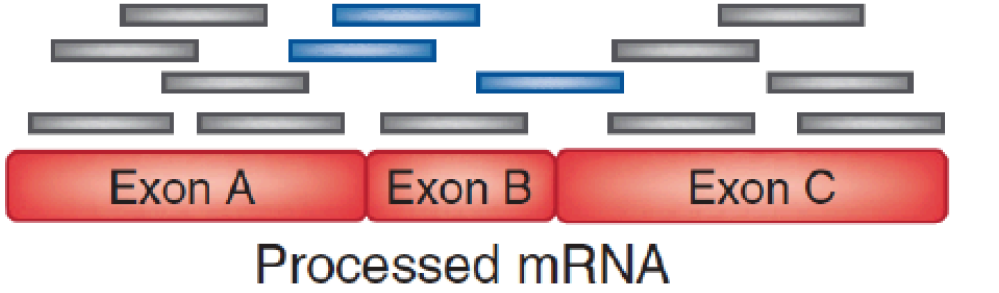
TYPES OF MAPPING

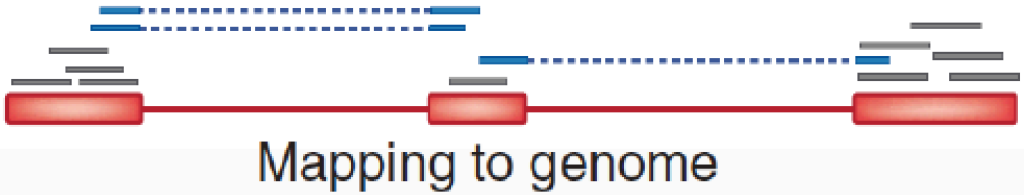
Paired End Mapping

- There is an expected insert size distribution based on the DNA fragment library.
- Mapping one read anchors the paired read to a specific location, even if the second read alone maps multiple places equally.
- Only one read in a pair might be mappable. (Known as singletons/orphans).
- Both reads can map with an unexpected insert size or orientation (known as discordant pairs).

Split Read Alignment

- Useful for predicting structural variants (or splice variants in RNA-seq, as pictured).
- Not many mappers do this directly, usually happens in a post-processing pipeline step.





Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).

Onto the tutorials!

- Complete the bowtie2 read mapping tutorial to learn how to actually map reads back to a reference file.
- Be sure to start an idev session!

- We still have additional presentations today on: SNV and SV calling, and a bonus presentation on SAM/BAM format. Some of this likely before break.