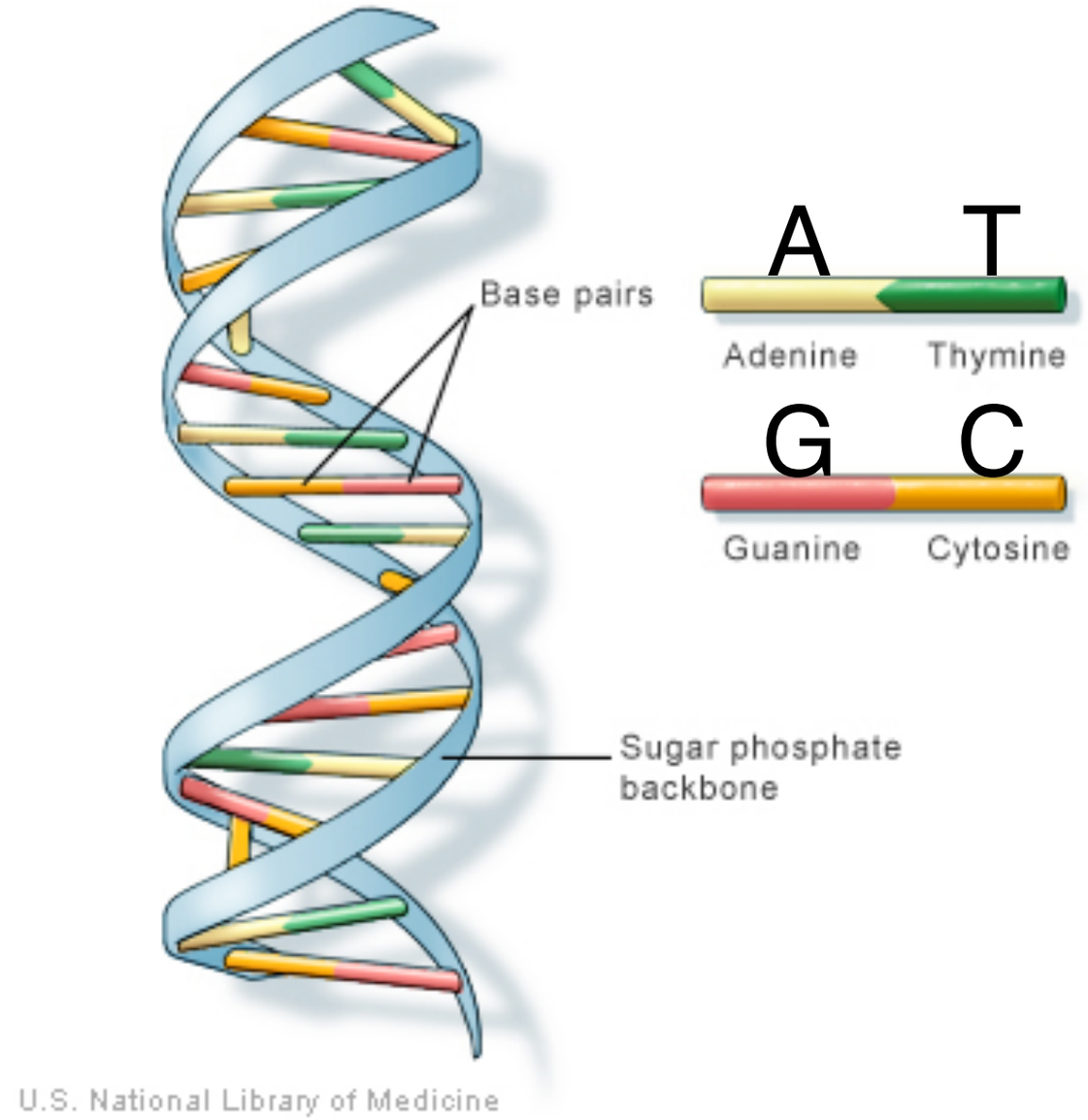
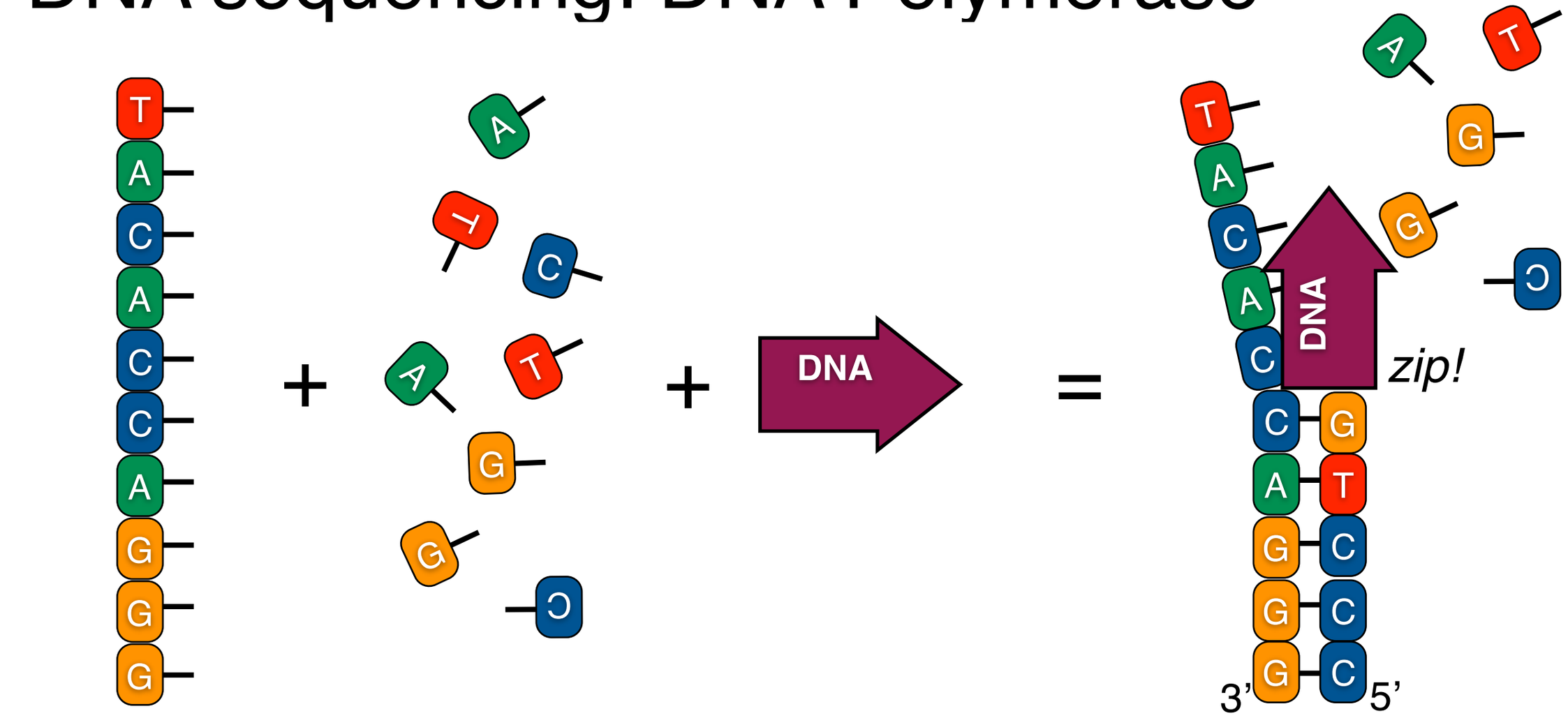
DNA sequencing: double helix



Picture: http://ghr.nlm.nih.gov/handbook/basics/dna

DNA sequencing: DNA Polymerase



Single-stranded DNA *template*

Free nucleotides

DNA polymerase

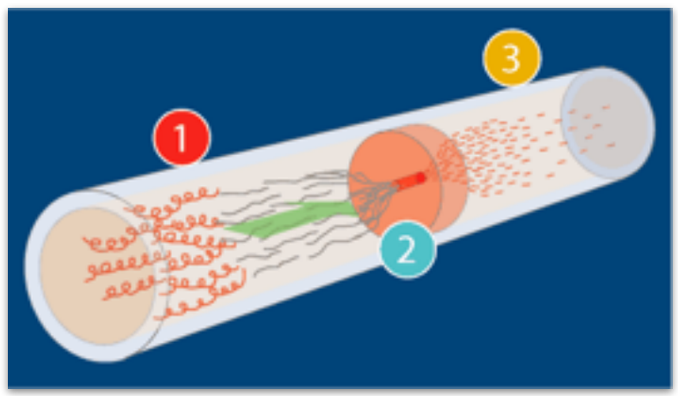
Strand synthesis

DNA polymerase moves along the template in one direction, integrating complementary nucleotides as it goes

Sequencing by synthesis

1. Take DNA sample, which includes many copies of the genome, and chop it into single-stranded fragments (“templates”)

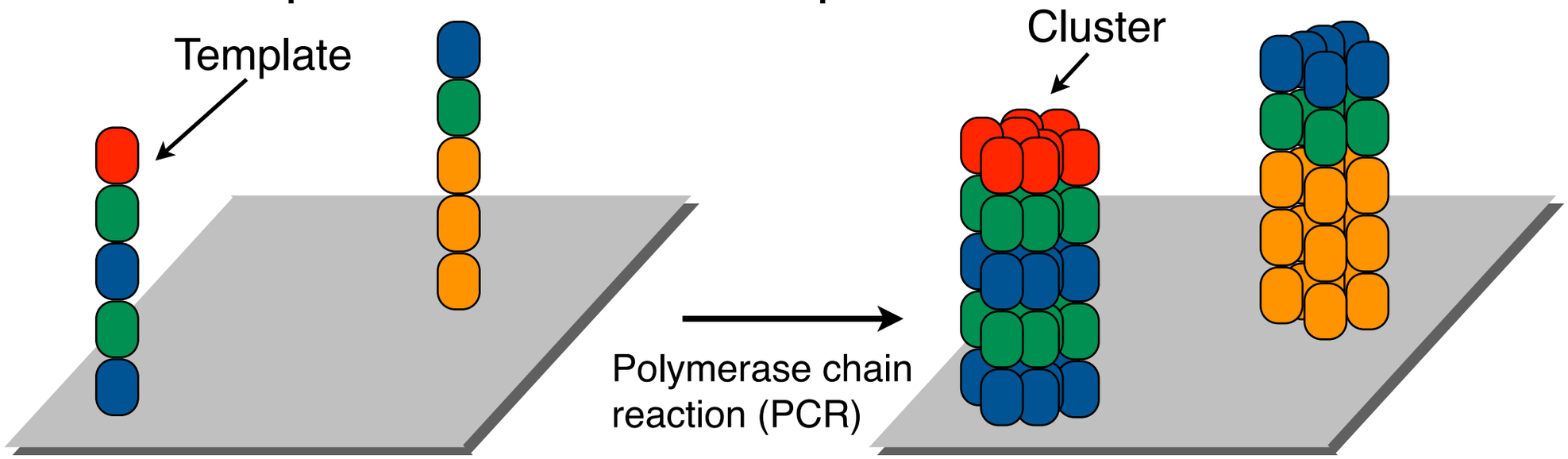
    E.g. with ultrasound waves, water-jet shearing (pictured), divalent cations



Picture: http://www.jgi.doe.gov/sequencing/education/how/how_1.html

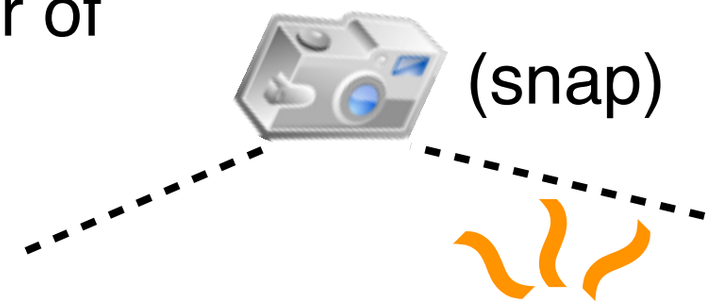
2. Attach templates to a surface

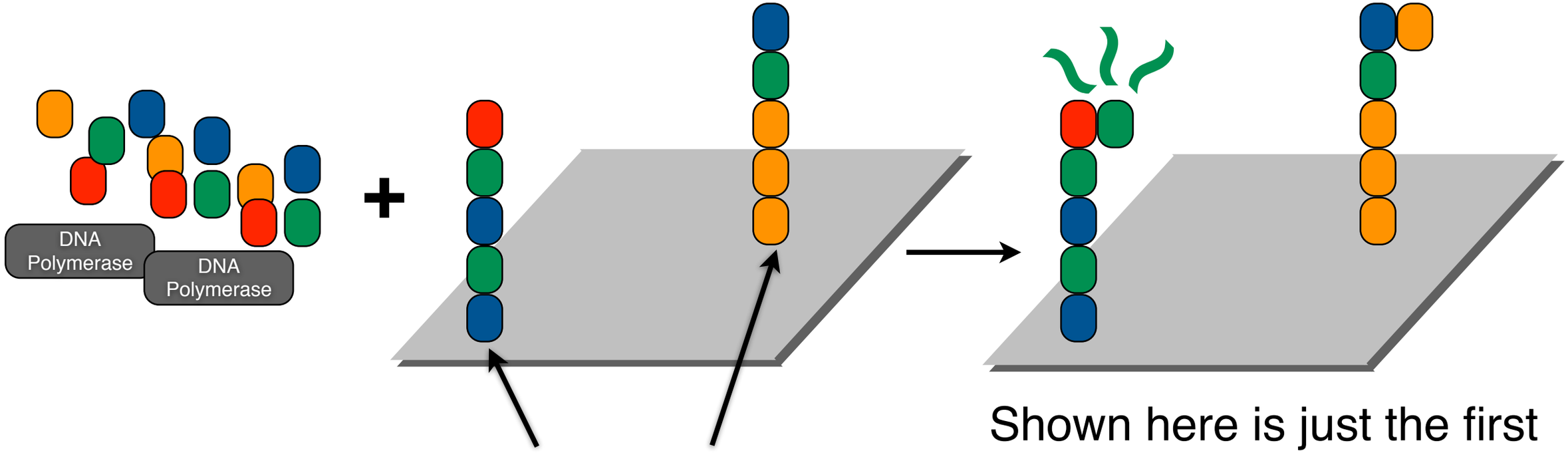
3. Make copies so that each template becomes a “cluster” of clones



Sequencing by synthesis

4. Repeatedly inject mixture of *color-labeled* nucleotides (A, C, G and T) and DNA polymerase. When a complementary nucleotide is added to a cluster, the corresponding color of light is emitted. Capture images of this as it happens.

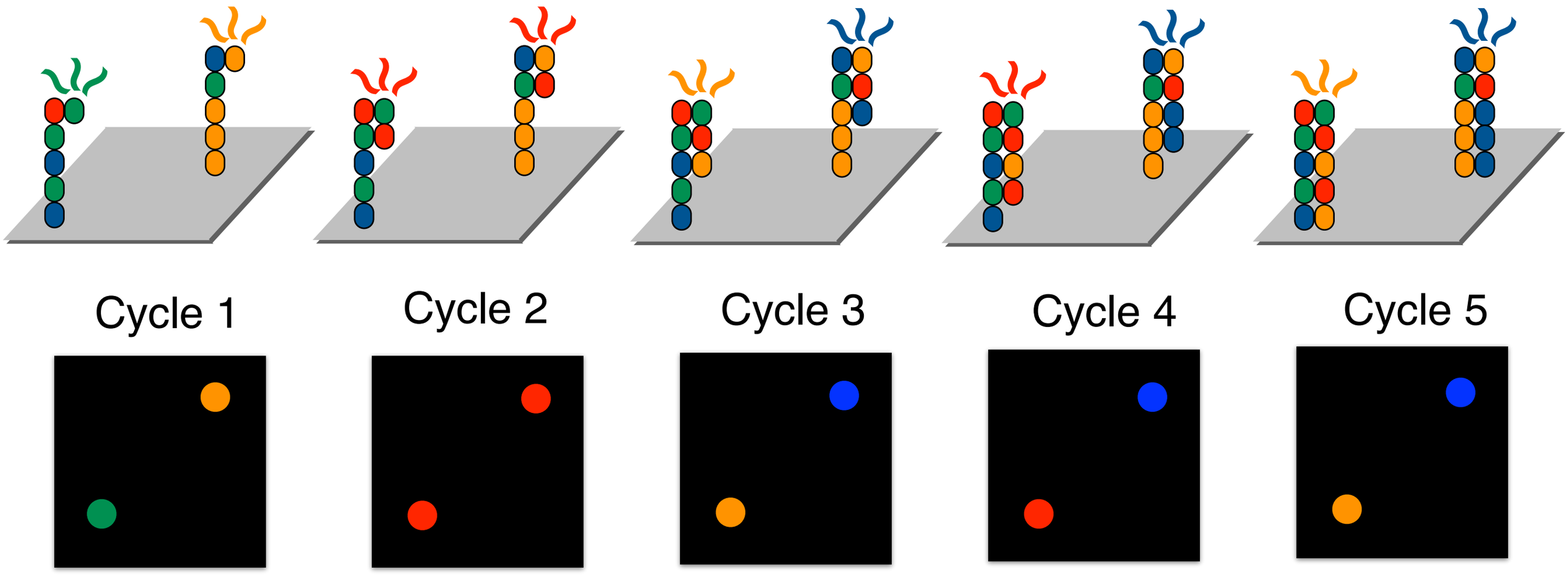




Shown here is just the first *sequencing cycle*

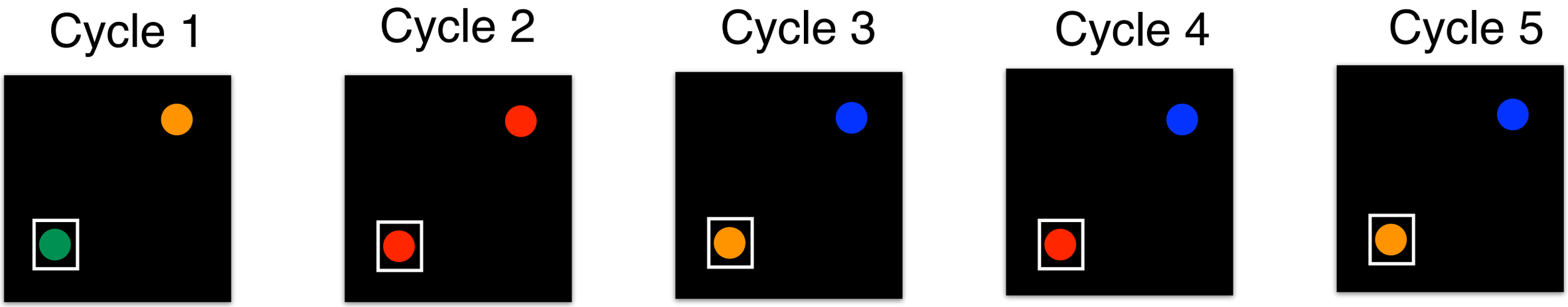
Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides

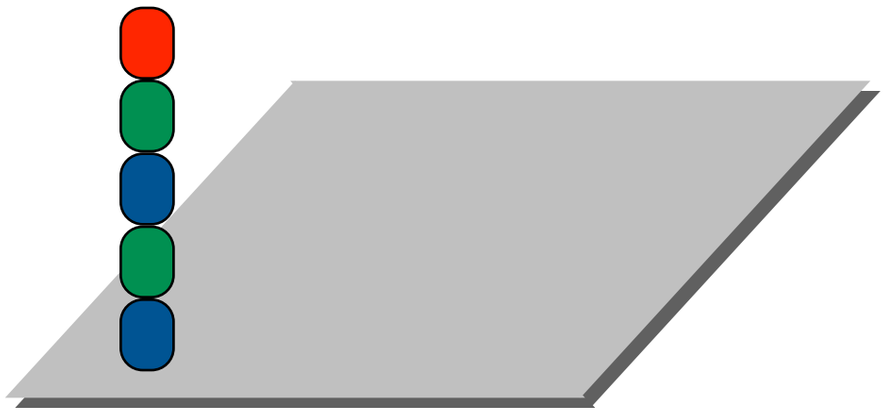


Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides



“Base caller” software looks at this cluster across all images and “calls” the complementary nucleotides: TACAC, corresponding to the template sequence



TACAC is a “sequence read,” or “read.” Actual reads are usually 100 or more nucleotides long.

Sequencing by synthesis

A modern sequencing-by-synthesis instrument such as the HiSeq sequences *billions* of clusters simultanously

A single “run” takes about 10 days to generate about 600 billion nucleotides of data

Cost of the reagents is $5-10K per run; multiplexing (sequencing many samples per run) further reduces cost per genome