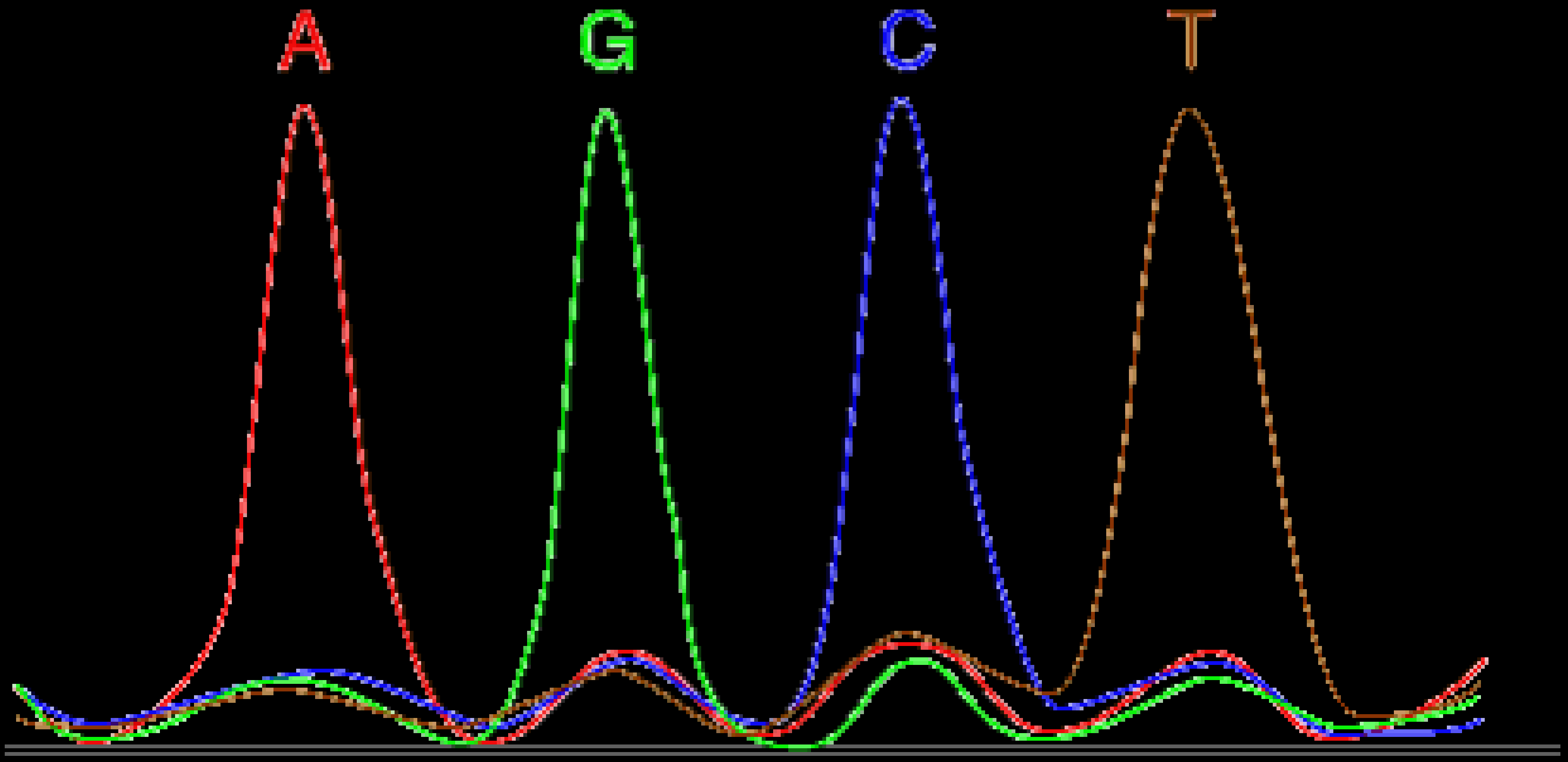
BENG 183

Trey Ideker

DNA Sequencing

The next generation



Sequencing topics to be covered in today's lecture

(1) Devils in the details:
DNA preparation prior to sequencing
Amplification: vectors or cycle sequencing
PAGE and Polymerases

(1) Next generation sequencing foundations:
EST sequencing, SAGE and MPSS

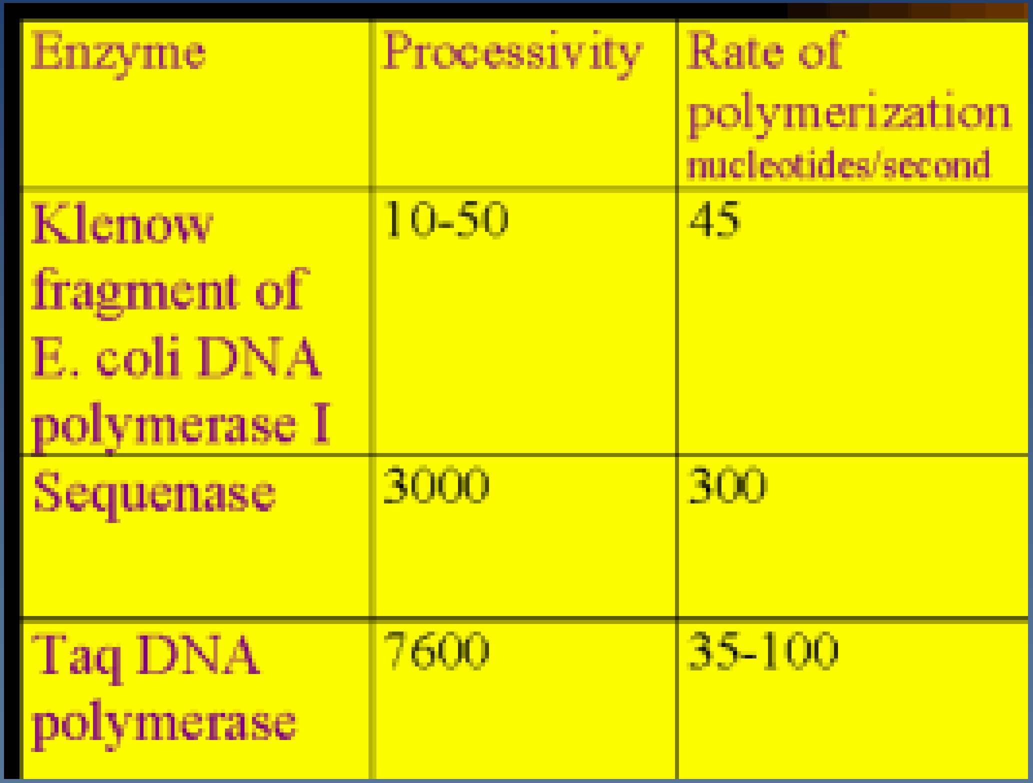
(2) Roche 454 pyrosequencing

(3) Illumina / Solexa sequencing

Polyacrylamide gels

- Gel is 7M urea and 4-8% acrylamide
- 1600 volts, heats gel to 65°C
- ~60 cm long
- Denaturing gel (two reasons why?)
- Resolves single DNA bp differences up to 1000 bp in length
  (why not longer???)

Polymerase Enzymes for DNA sequencing



Enzyme	Processivity	Rate of polymerization nucleotides/second
Klenow fragment of E. coli DNA polymerase I	10-50	45
Sequenase	3000	300
Taq DNA polymerase	7600	35-100

Processivity: average # of nucleotides synthesized before enzyme dissociates

DNA preparation

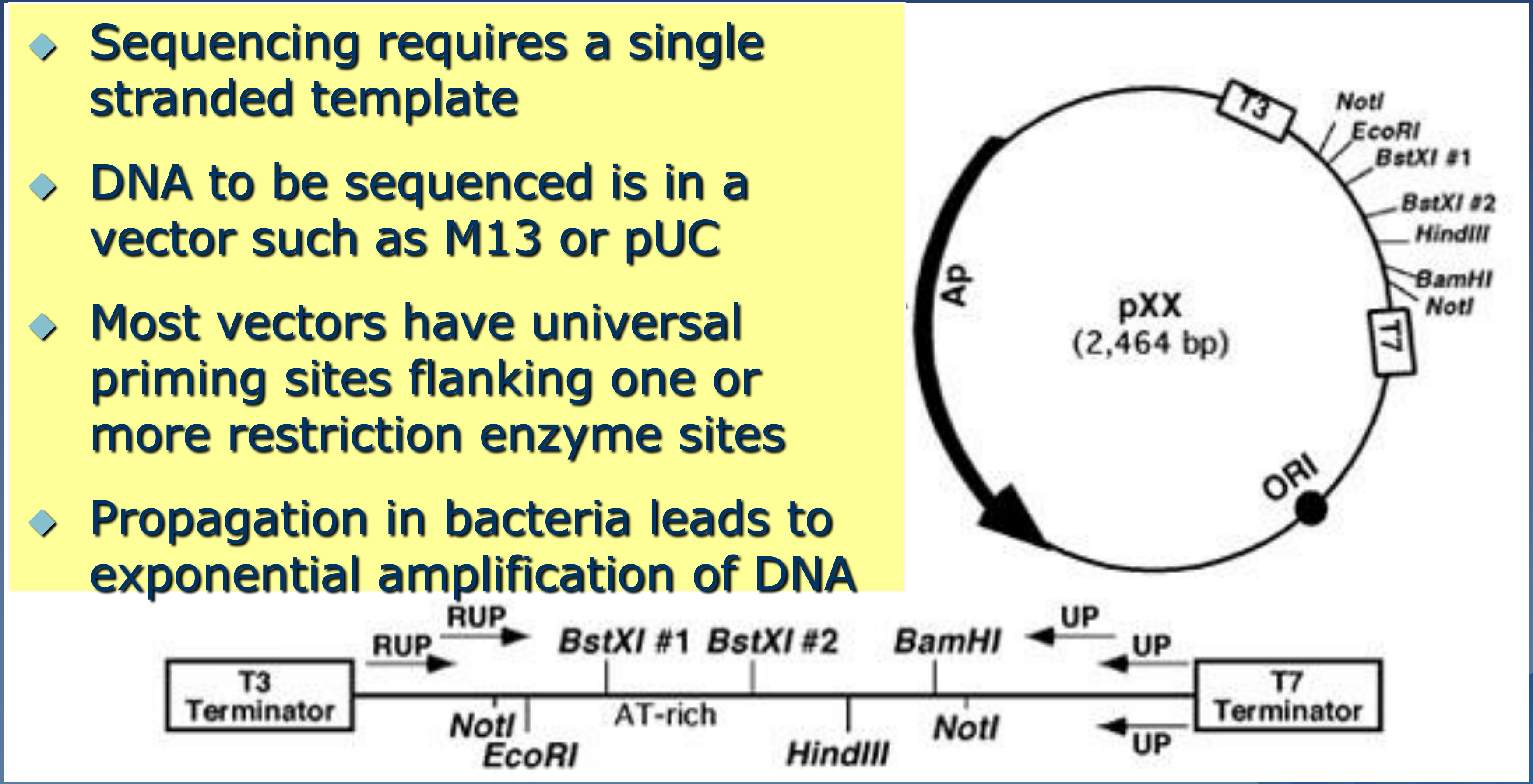
Many steps are required before DNA samples are loaded on gels:

- DNA isolation
- Fragmentation
- Amplification (bacterial vectors, PCR, *cycle sequencing*)
- Re-isolation of DNA (if vector amplified)

Most of these have also been automated

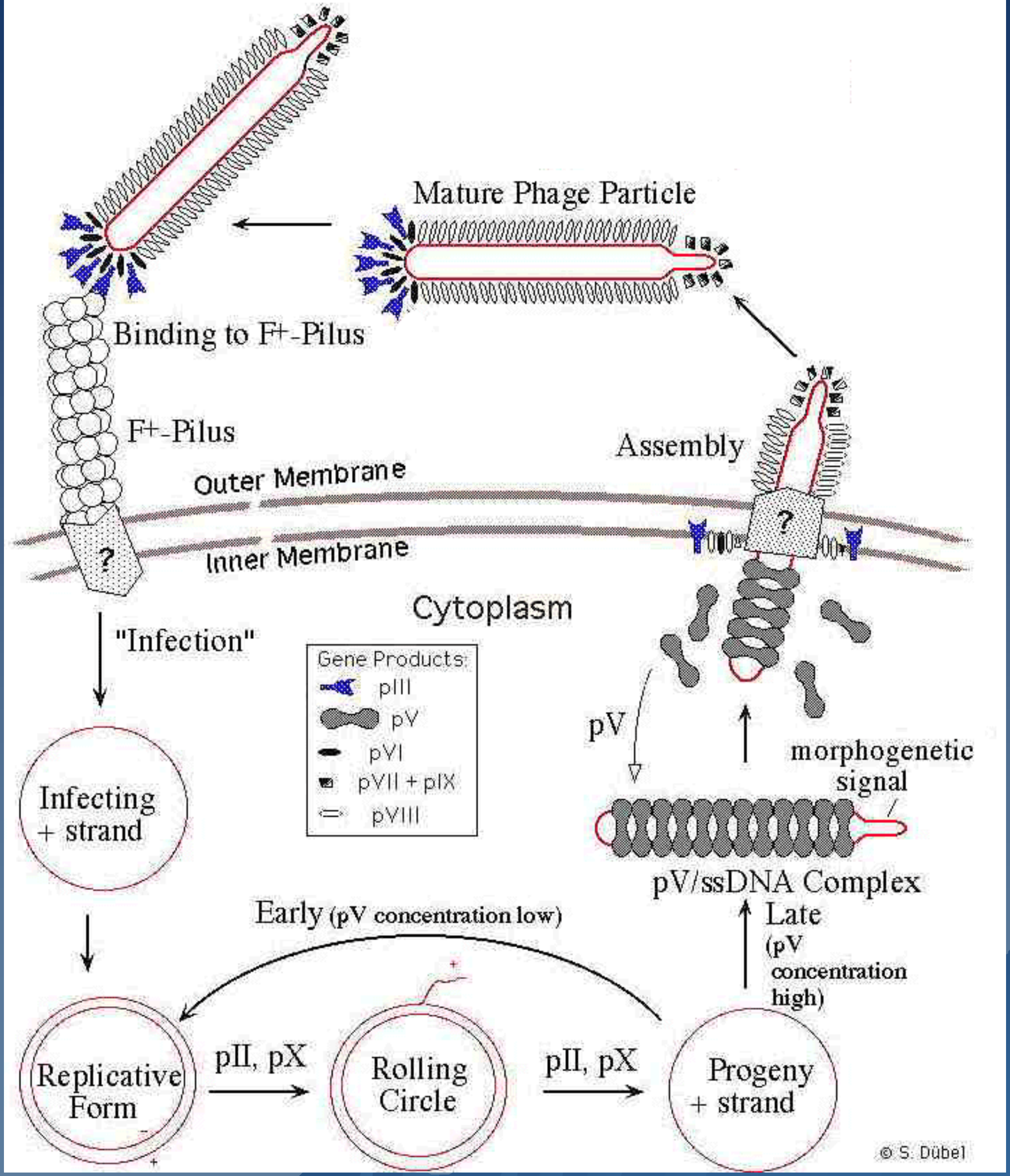
Vectors for sequencing

- Sequencing requires a single stranded template
- DNA to be sequenced is in a vector such as M13 or pUC
- Most vectors have universal priming sites flanking one or more restriction enzyme sites
- Propagation in bacteria leads to exponential amplification of DNA



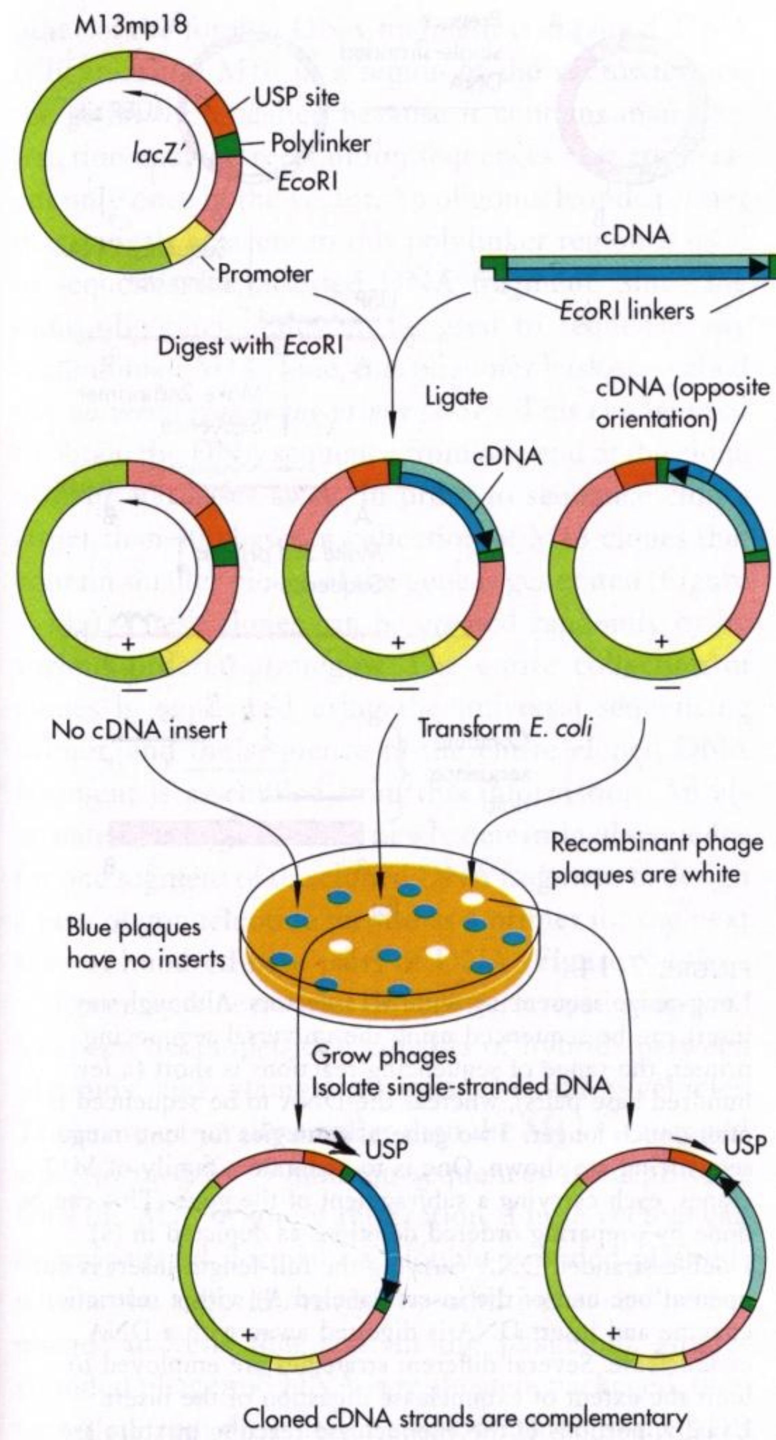
M13 vectors for DNA sequencing

- Circular DNA modified from the genome of the M13 bacteriophage.
- Transitions through both single and double stranded forms, making it ideal for sequencing applications
- Double stranded DNA is the replicative form and is used for cloning.
- DNA packaged in phage capsid is single stranded.



http://www.biochem.arizona.edu/classes/bioc471/pages/materials.html

M13 vectors for DNA sequencing


M13mp18
USP site
Polylinker
EcoRI
lacZ′
Promoter
cDNA
EcoRI linkers
Digest with EcoRI
Ligate
cDNA (opposite orientation)
cDNA
+
−
No cDNA insert
Transform E. coli
Recombinant phage plaques are white
Blue plaques have no inserts
Grow phages
Isolate single-stranded DNA
USP
USP
Cloned cDNA strands are complementary

From *Recombinant DNA 2nd edition, Watson et al.*

The X Prize Foundation

- In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."

- http://genomics.xprize.org/

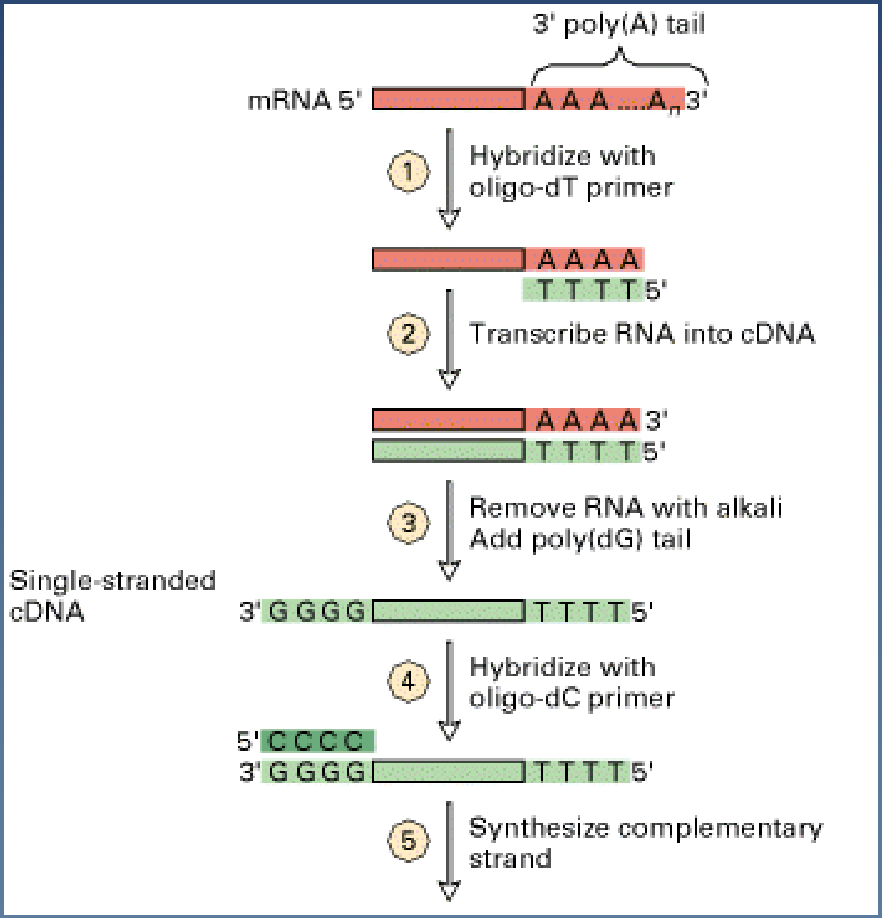
cDNA / EST sequencing projects

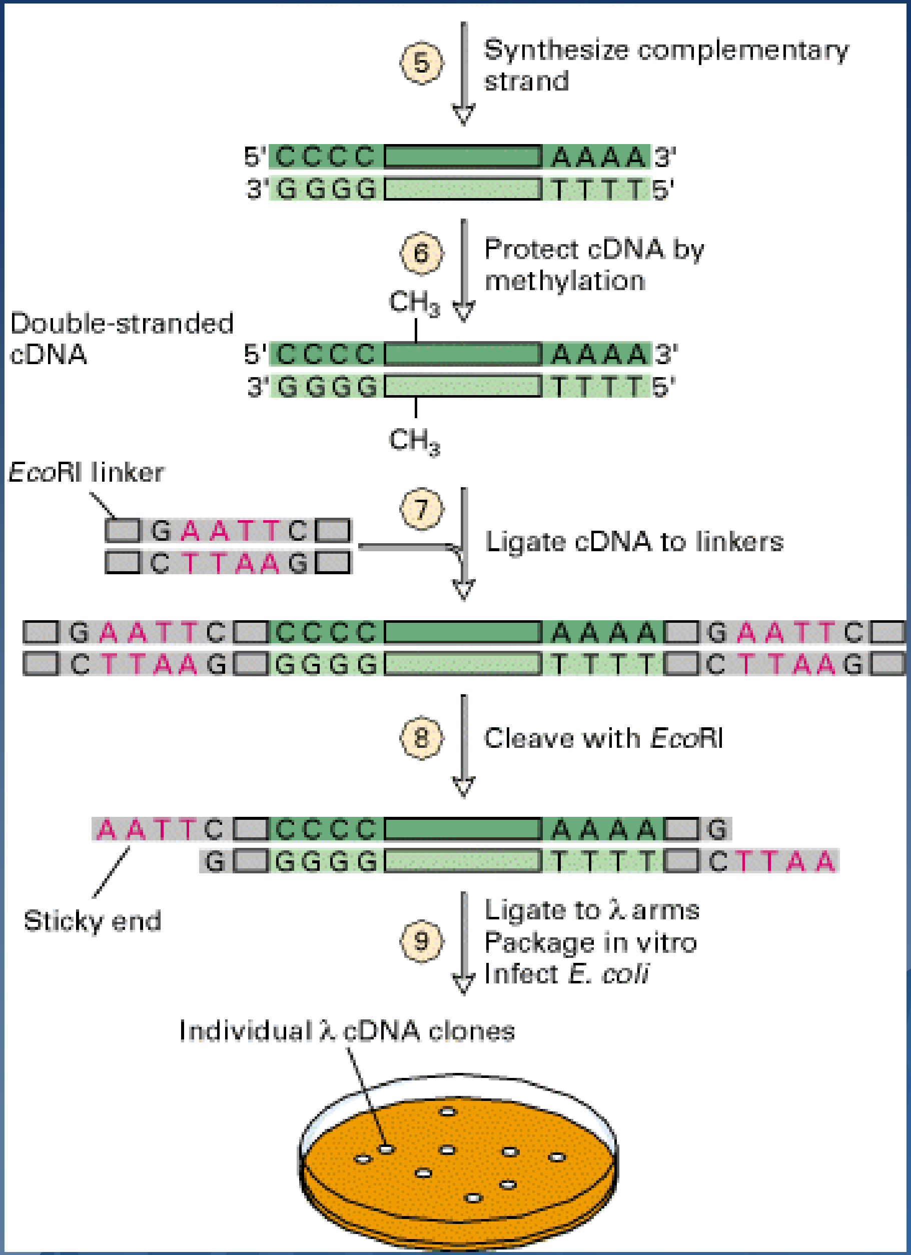
- cDNA = complementary or copy DNA
- EST = Expressed Sequence Tag

- Direct sequencing of cDNAs (yielding ESTs) through large-scale random sampling of sequences from a whole-cell RNA extract
- Statistical counting of distinct sequences provides an estimate of expression level
- Conversely, cDNA library can be normalized to capture rare messages
- Requires large scale sequencing to get statistical significance

cDNA / EST Sequencing:

Preparation of a cDNA library in phage λ vector


3' poly(A) tail
mRNA 5'
A A A ....A$_n$ 3'
1 Hybridize with oligo-dT primer
A A A A
T T T T 5'
2 Transcribe RNA into cDNA
A A A A 3'
T T T T 5'
3 Remove RNA with alkali Add poly(dG) tail
Single-stranded cDNA
3' G G G G
T T T T 5'
4 Hybridize with oligo-dC primer
5' C C C C
3' G G G G
T T T T 5'
5 Synthesize complementary strand


5 Synthesize complementary strand
5' C C C C
A A A A 3'
3' G G G G
T T T T 5'
6 Protect cDNA by methylation
Double-stranded cDNA
CH_3
5' C C C C
A A A A 3'
3' G G G G
T T T T 5'
CH_3
EcoRI linker
G A A T T C
C T T A A G
7 Ligate cDNA to linkers
G A A T T C C C C C A A A A G A A T T C
C T T A A G G G G G T T T T C T T A A G
8 Cleave with EcoRI
A A T T C C C C C A A A A G
G G G G G T T T T C T T A A
Sticky end
9 Ligate to λ arms Package in vitro Infect E. coli
Individual λ cDNA clones

SAGE Technology

Serial Analysis of Gene Expression

- Takes idea of sequence sampling to the extreme
- Generates short ESTs (9-14nt) which are joined into long concatamers and then sequenced
- 4^9 is 262,144, ~5-fold the number of human genes
- The count of each type of tag estimates RNA copy number
- >50X more efficient than cDNA sequencing because many RNAs are represented in a single sequencing run

Steps to SAGE

- Copy mRNA → ds cDNA using biotinylated (dT)
- Cleave with anchoring enzyme (AE) which cleaves within ~250bp of poly-A tail at 3’ end.
- Capture this segment on streptavidin beads
- Ligate to linkers containing a type IIs restriction site, which cleave DNA 14 bp away from this site.

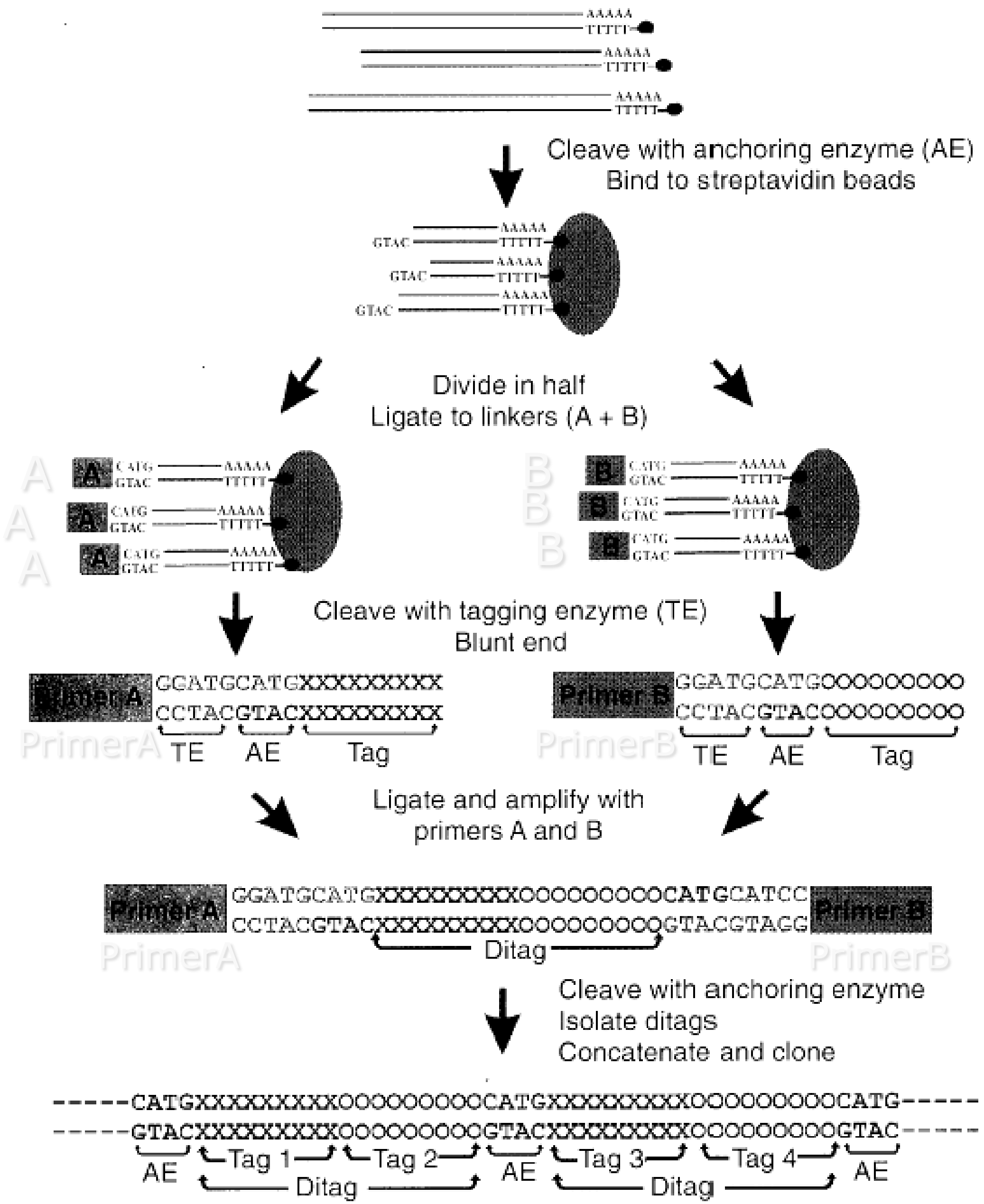

AAAAA
TTTTT
Cleave with anchoring enzyme (AE)
Bind to streptavidin beads
GTAC
Divide in half
Ligate to linkers (A + B)
A
CATG
GTAC
B
Cleave with tagging enzyme (TE)
Blunt end
Primer A
GGATGCATGXXXXXXXXX
CCTACGTACXXXXXXXXX
PrimerA
TE
AE
Tag
Primer B
GGATGCATGOOOOOOOOO
CCTACGTACOOOOOOOOO
PrimerB
TE
AE
Tag
Ligate and amplify with
primers A and B
Primer A
GGATGCATGXXXXXXXXXOOOOOOOOOCATGCATCC
CCTACGTACXXXXXXXXXOOOOOOOOOGTACGTAGG
Primer B
Ditag
PrimerA
PrimerB
Cleave with anchoring enzyme
Isolate ditags
Concatenate and clone
-----CATGXXXXXXXXXOOOOOOOOOCATGXXXXXXXXXOOOOOOOOOCATG-----
-----GTACXXXXXXXXXOOOOOOOOOGTACXXXXXXXXXOOOOOOOOOGTAC-----
AE
Tag 1
Tag 2
AE
Tag 3
Tag 4
AE
Ditag
Ditag

Fig. 1. Schematic of SAGE. The anchoring enzyme is Nla III and the tagging enzyme is Fok I. Sequences colored red and green represent primer-derived sequences, whereas blue represents transcript-derived sequences, with X and O indicating nucleotides of different tags. See text for further explanation.

Velculescu et al.
Science (1995)

WHY DI-TAGS?
Ditags are used to detect bias in the PCR amplification step.

The probability of any two tags being coupled in the same ditag is small.

Biased amplification can be detected as many ditags always having the same 2 tags present.

SAGE (continued)

Example of a concatemer:

CATGACCCACGAGCAGGGTACGATGATACATGGAAACCTATGCACCTTGGGTAGCACATG

TAG1 TAG2 TAG3 TAG4

Counting the tags:

Tag Sequence	Count
ATCTGAGTTC	1075
GCGCAGACTT	125
TCCCCGTACA	112
TAGGACGAGG	92
GCGATGGCGG	91
TAGCCCAGAT	83
GCCTTGTTTA	80

Tag Sequence	Count
GCGATATTGT	66
TACGTTTCCA	66
TCCCGTACAT	66
TCCCTATTAA	66
GGATCACAAT	55
AAGGTTCTGG	54
CAGAACCGCG	50
GGACCGCCCC	48

Massively Parallel Signature Sequencing (MPSS, Brenner)

- cDNA fragments are cloned onto microbeads
- Fragments are sequenced over multiple cycles of a ligation based sequencing method.
- This is carried out simultaneously on a million microbeads, each having a single DNA template
- Microbeads are arranged in a flow cell to form a closely packed planar array
- The bead array remains fixed while sequencing reagents are pumped through the flow cell

MPSS Overview

- Produces short seq. signatures
- Their relative abundance in a library gives a quant. estimate of expression of that gene.
- Millions of microbeads are used to capture cDNAs in solution
- The *initiating* adaptor adds Bbv1 site which cleaves DNA upstream
- The overhanging 4 bps are sequenced by hyb. to *decoders*
- The decoders add addl. Bbv1 sites
- Several rounds yield enough sequence to ID the RNA

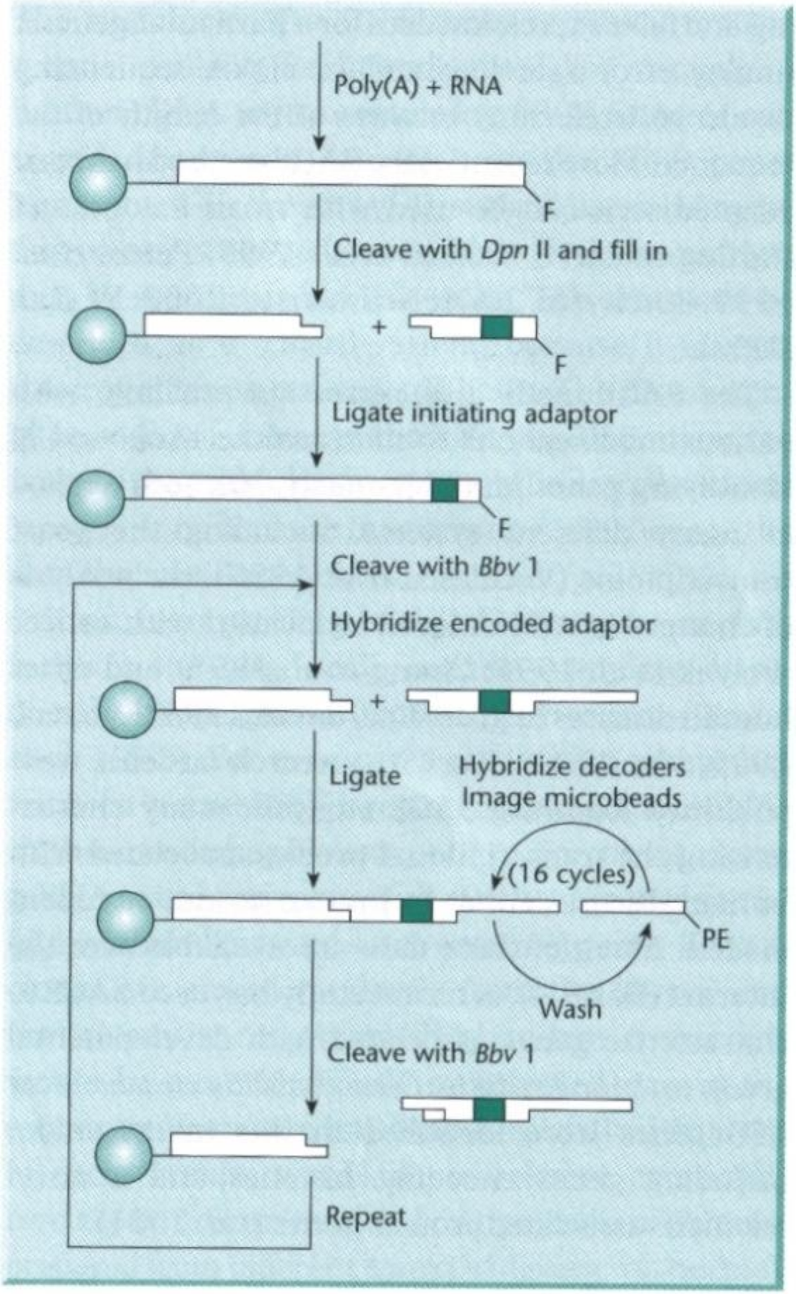
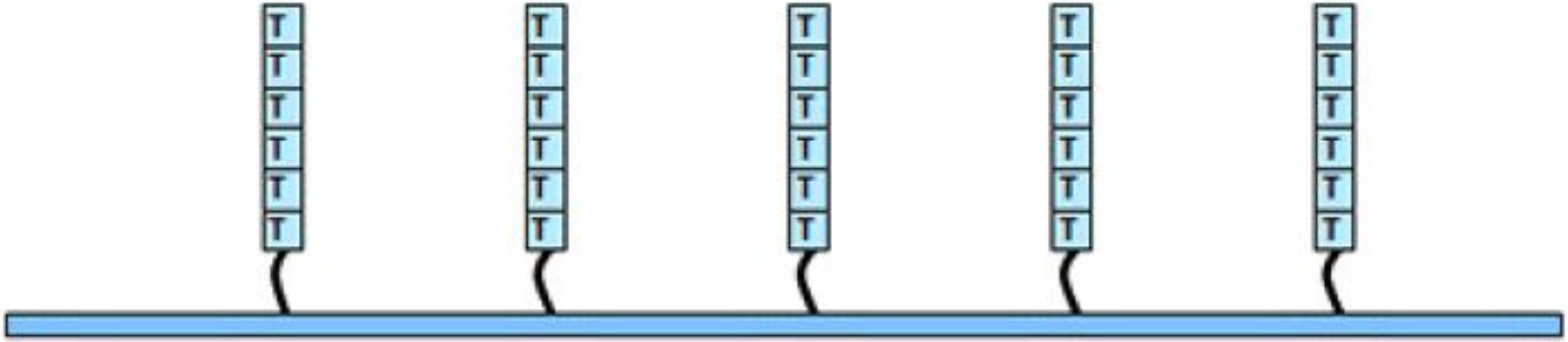


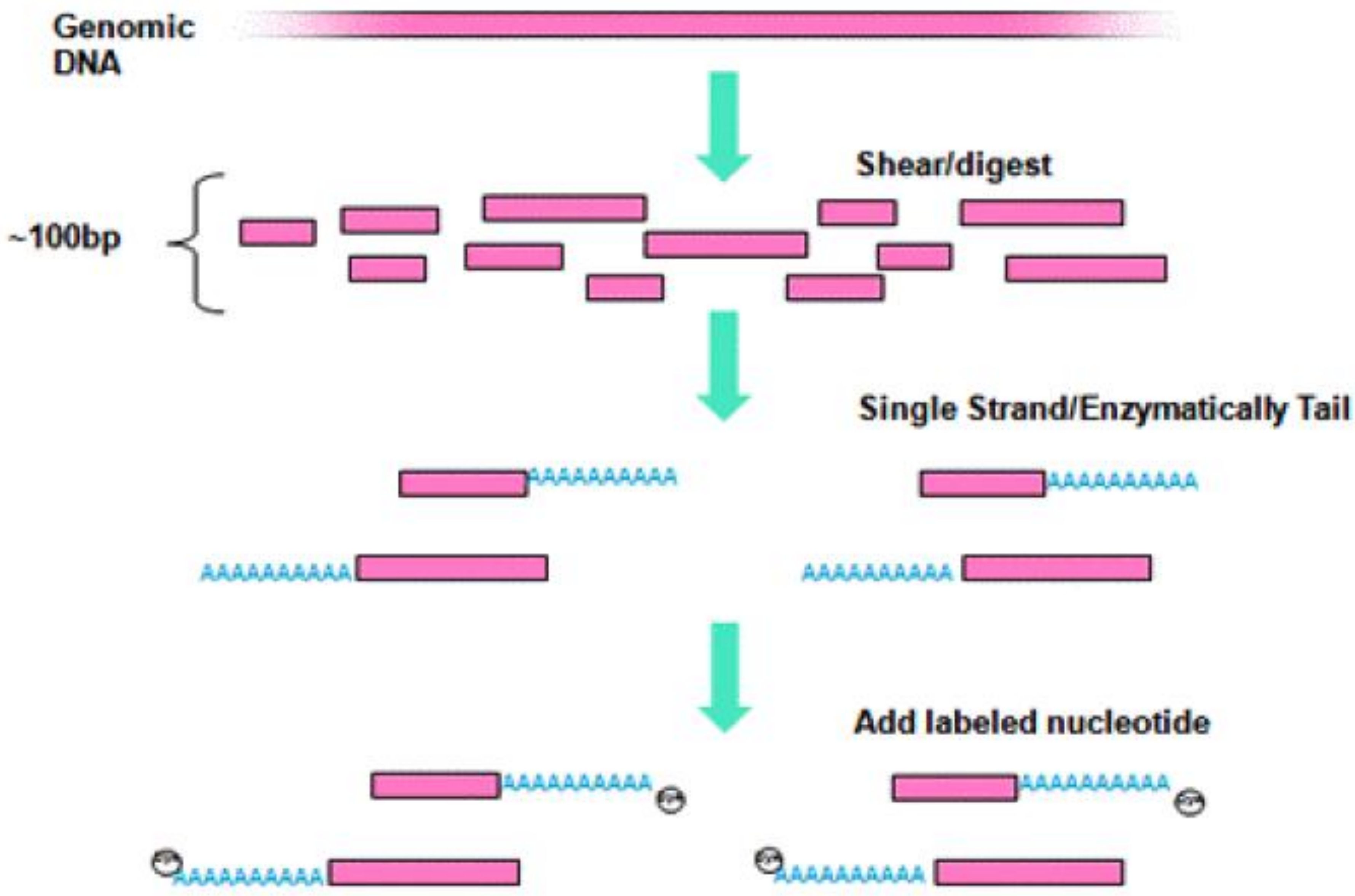
Fig. 9.7 Principle of massively parallel signature sequencing (MPSS) technique. PE = fluorescent label. (Adapted from Brenner 2000.)

Solexa Sequencing Overview

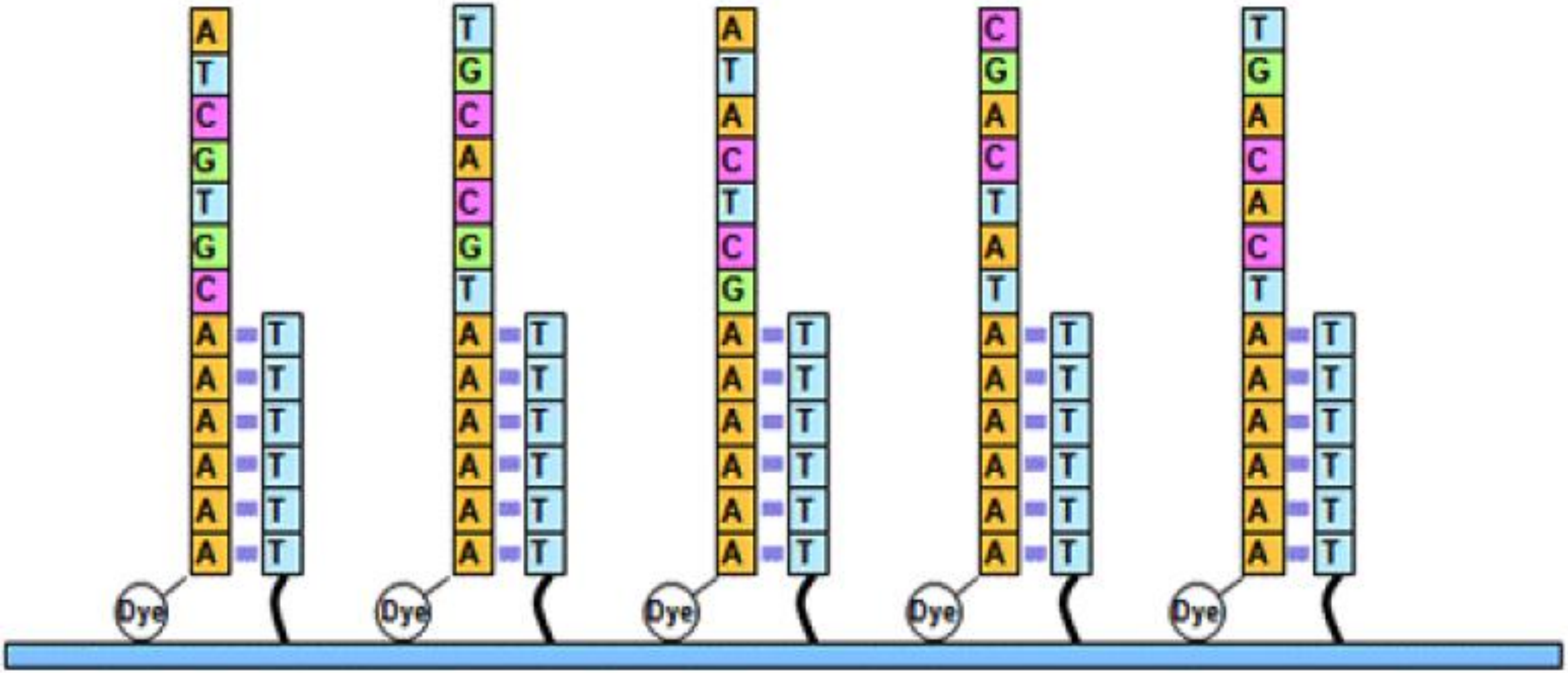
Step 1: Universal primers are immobilized on a glass surface inside a flow cell.



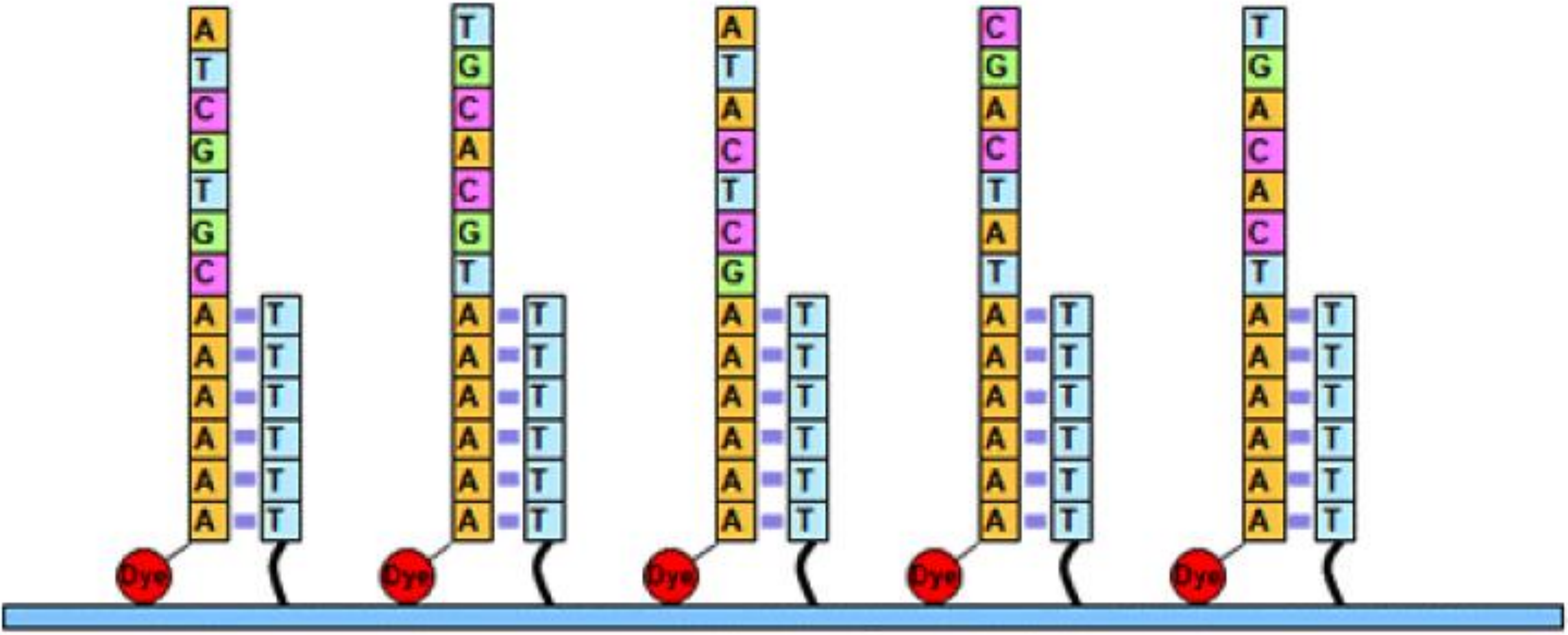
Step 2: Genomic DNA is converted into sequencing templates ready to load into the flow cell.



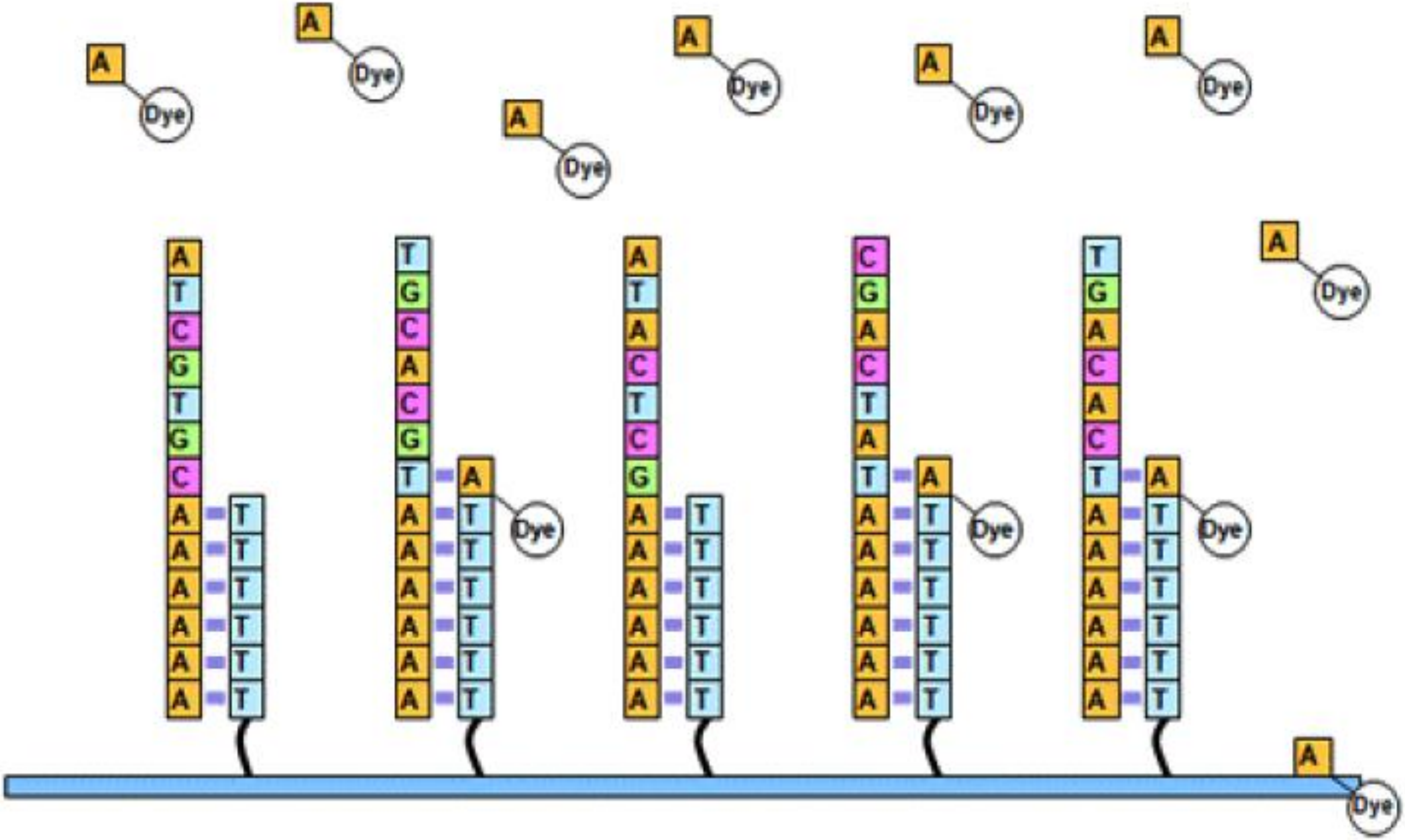
Step 3: Hybridize the DNA templates to the immobilized primers inside the flow cell.



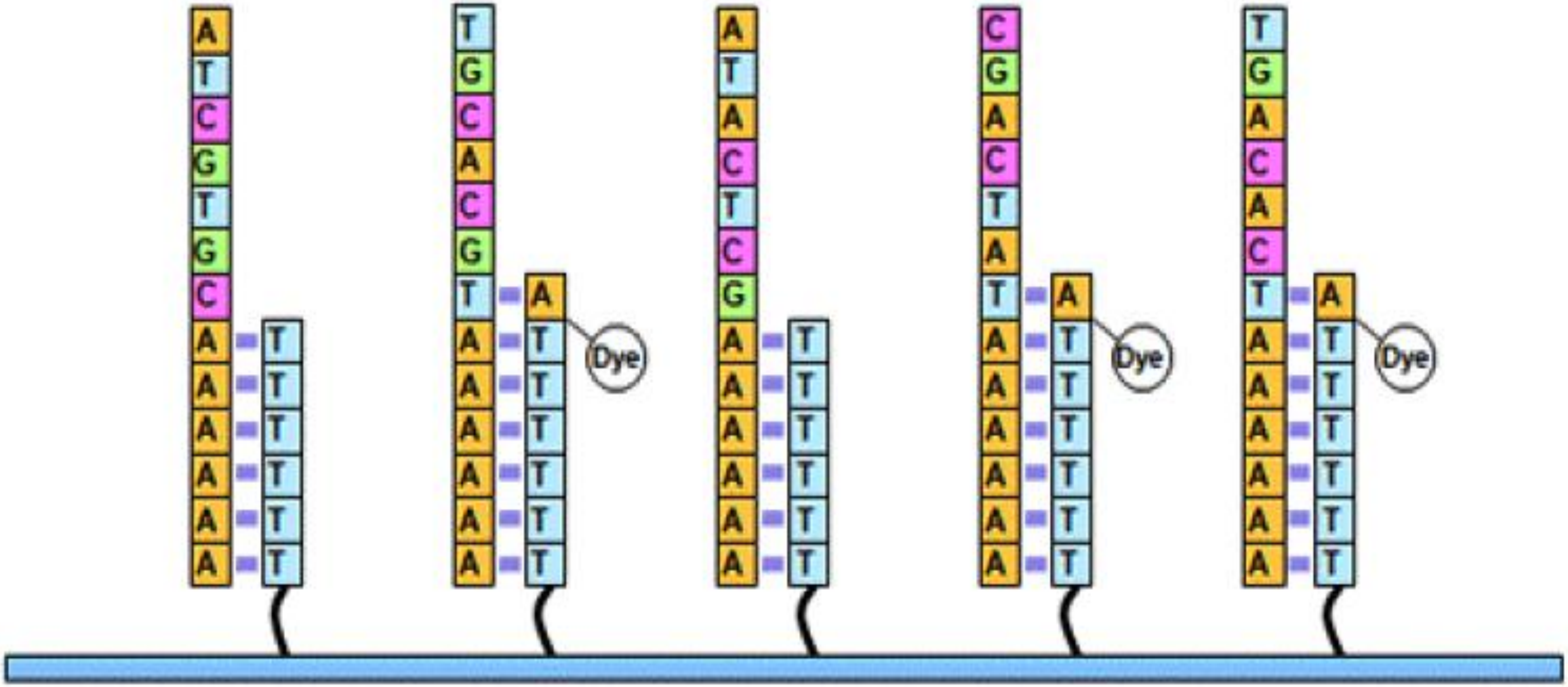
Step 4: Visualize the template:primer duplexes by illuminating the surface with a laser and imaging with an electronic camera connected to a microscope. Record the positions of all the duplexes on the surface. After imaging, the dye molecules are cleaved and washed away.


A T C G T G C A A A A A A
T T T T T T
Dye
T G C A C G T A A A A A A
T T T T T T
Dye
A T A C T C G A A A A A A
T T T T T T
Dye
C G A C T A T A A A A A A
T T T T T T
Dye
T G A C A C T A A A A A A
T T T T T T
Dye

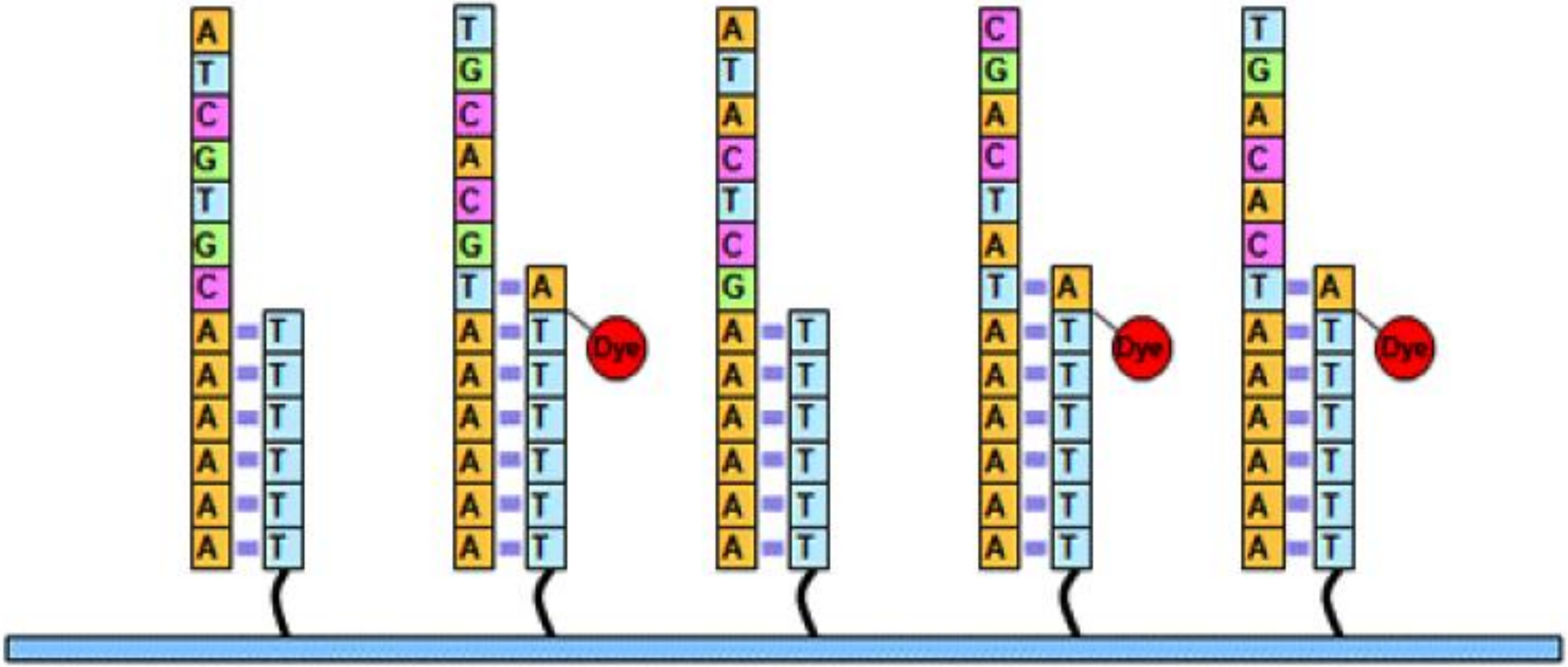
Step 5: Flow in DNA polymerase and one type of fluorescently labeled nucleotide (for example A). The polymerase will catalyze the addition of labeled nucleotide to the appropriate primers.



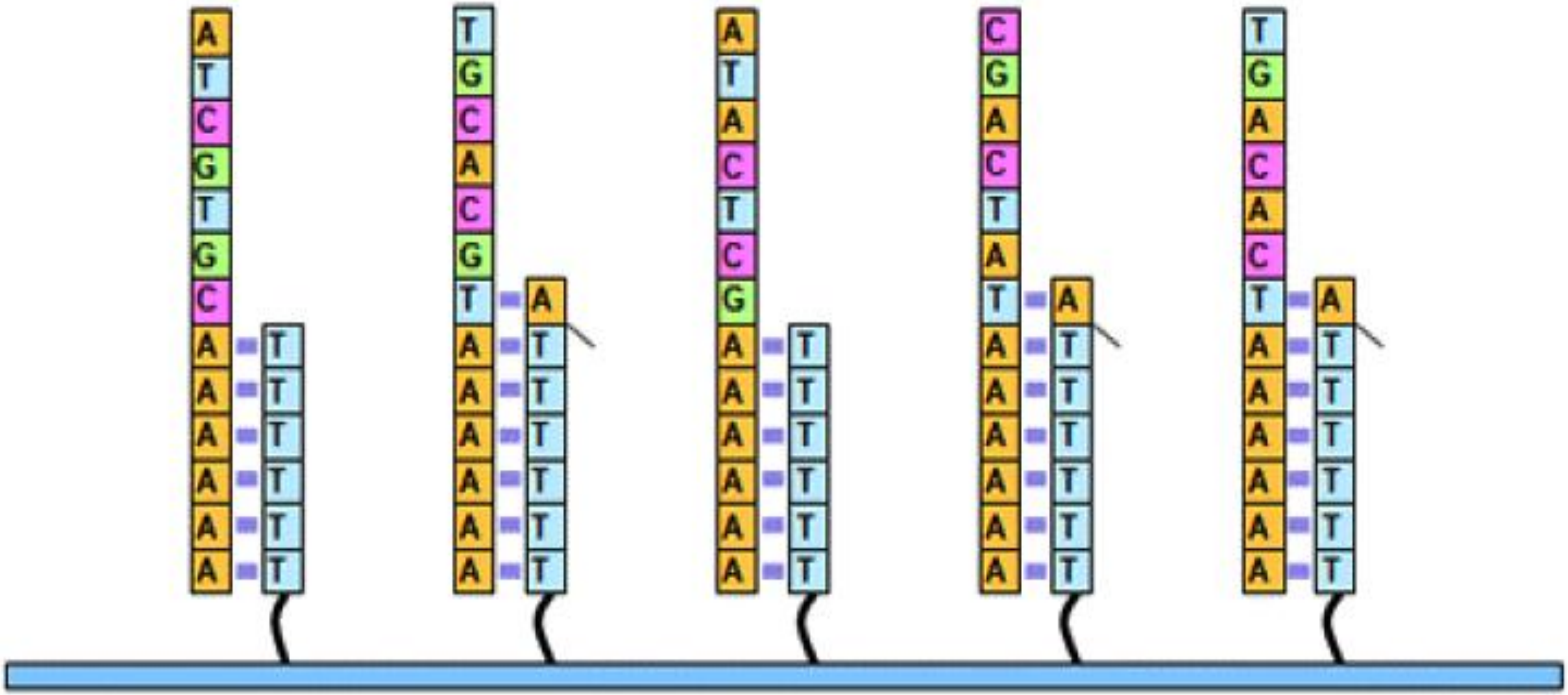
Step 6: Wash out the polymerase and unincorporated nucleotides.



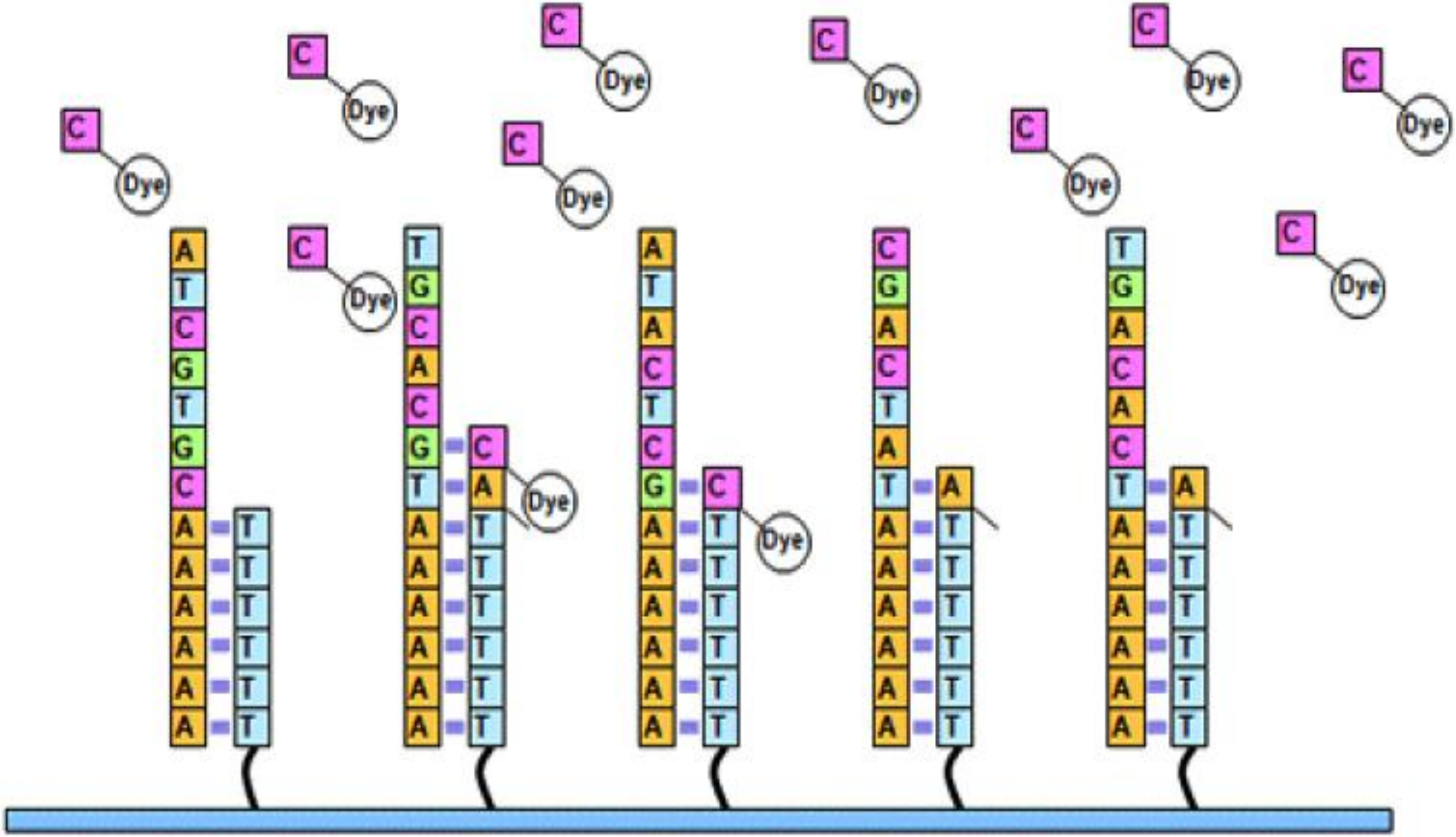
Step 7: Visualize the incorporated labeled nucleotides by illuminating the surface with a laser and imaging with the camera. Record the positions of the incorporated nucleotides.



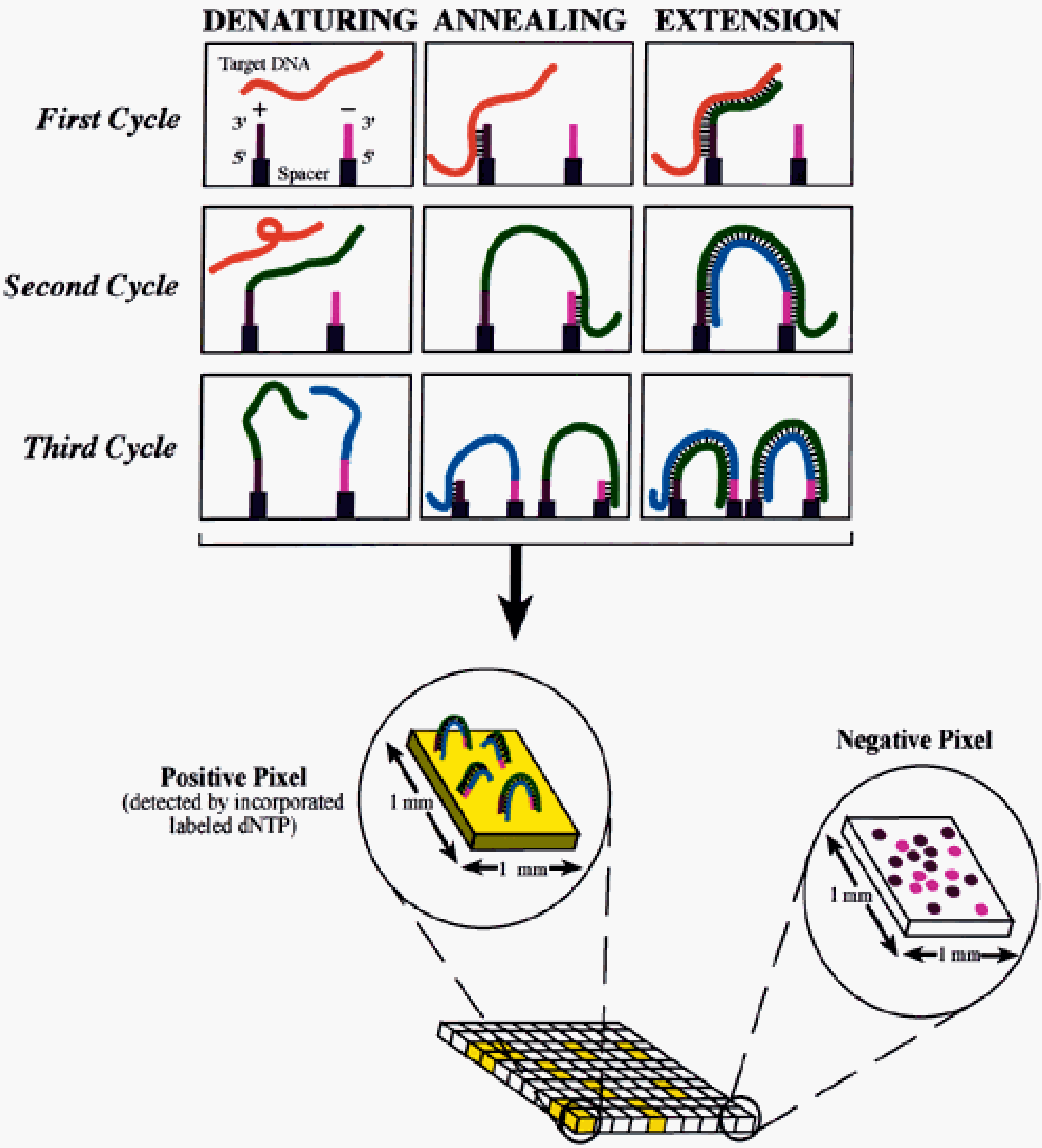
Step 8: Remove the fluorescent label on each nucleotide.



Step 9: Repeat the process from step 5 with the next nucleotide (stepping through A, C, G and T), until the desired read-length is achieved.


C Dye
C Dye
C Dye
C Dye
C Dye
C Dye
C Dye
C Dye
C Dye
C Dye
A T C G T G C A A A A A A
T T T T T T
T G C A C G T A A A A A A
C A T T T T T T
Dye
A T A C T C G A A A A A A
C T T T T T T
Dye
C G A C T A T A A A A A A
A T T T T T T
T G A C A C T A A A A A A
A T T T T T T

Bridge Amplification



Solexa Sequencing Video

- http://www.wellcome.ac.uk/Education-resources/Teaching-and-education/Animations/DNA/WTX056051.htm
- Related to bridge amplification: POLONIES http://arep.med.harvard.edu/Polonator/

Pyrosequencing

Note: No actual houses are burned down in pyrosequencing

Pyrosequencing (Roche 454)

- A **luciferase** is an enzyme which emits light in the presence of ATP.

Several organisms, such as the American firefly and the poisonous Jack-o-lantern mushroom, produce luciferases.

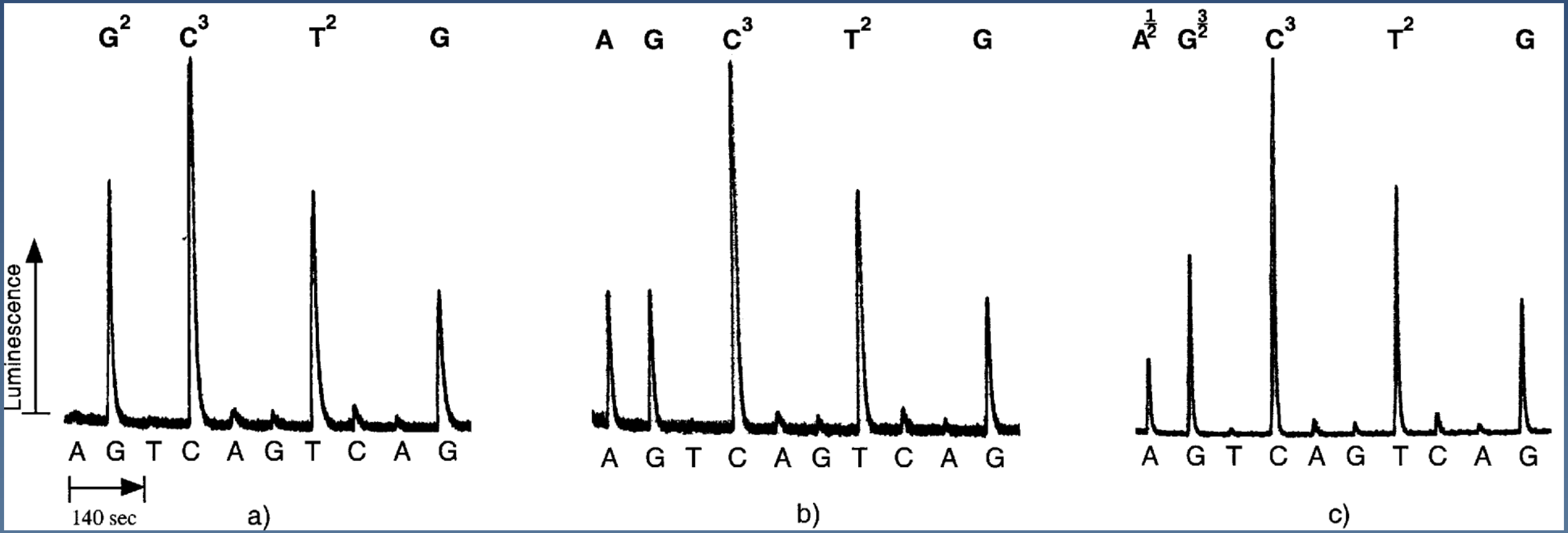
Detecting polymerase activity

- Recall: Pyrophosphate is also known as PPi, also known as “two phosphate groups stuck together”. During replication, each addition of a dNTP releases pyrophosphate
- In the reaction mixture, PPi allows adenosine phosphosulfate (APS) to be converted to ATP; this ATP allows luciferase to luciferate (emit light).
- Measures strand extension as it happens

Pyrosequencing cycle

- Add dATP. If light is emitted, your sequence starts with A. If not, the dATP is degraded (or elutes past immobilized primer).
- Add dGTP. If light is emitted, the next base must be a G.
- Then add T, then C. You now know at least one (maybe more) base of the sequence.
- Repeat!

Pyrosequencing output



Runs of bases produce higher peaks – for instance, the sequence for (a) is GGCCCTTG. Sample (c) comes from a heterozygous individual (hence the heights in multiples of ½)