

CSC 334: TOPICS IN COMPUTATIONAL BIOLOGY

“Algorithms for Genomic Data”

Fall 2015

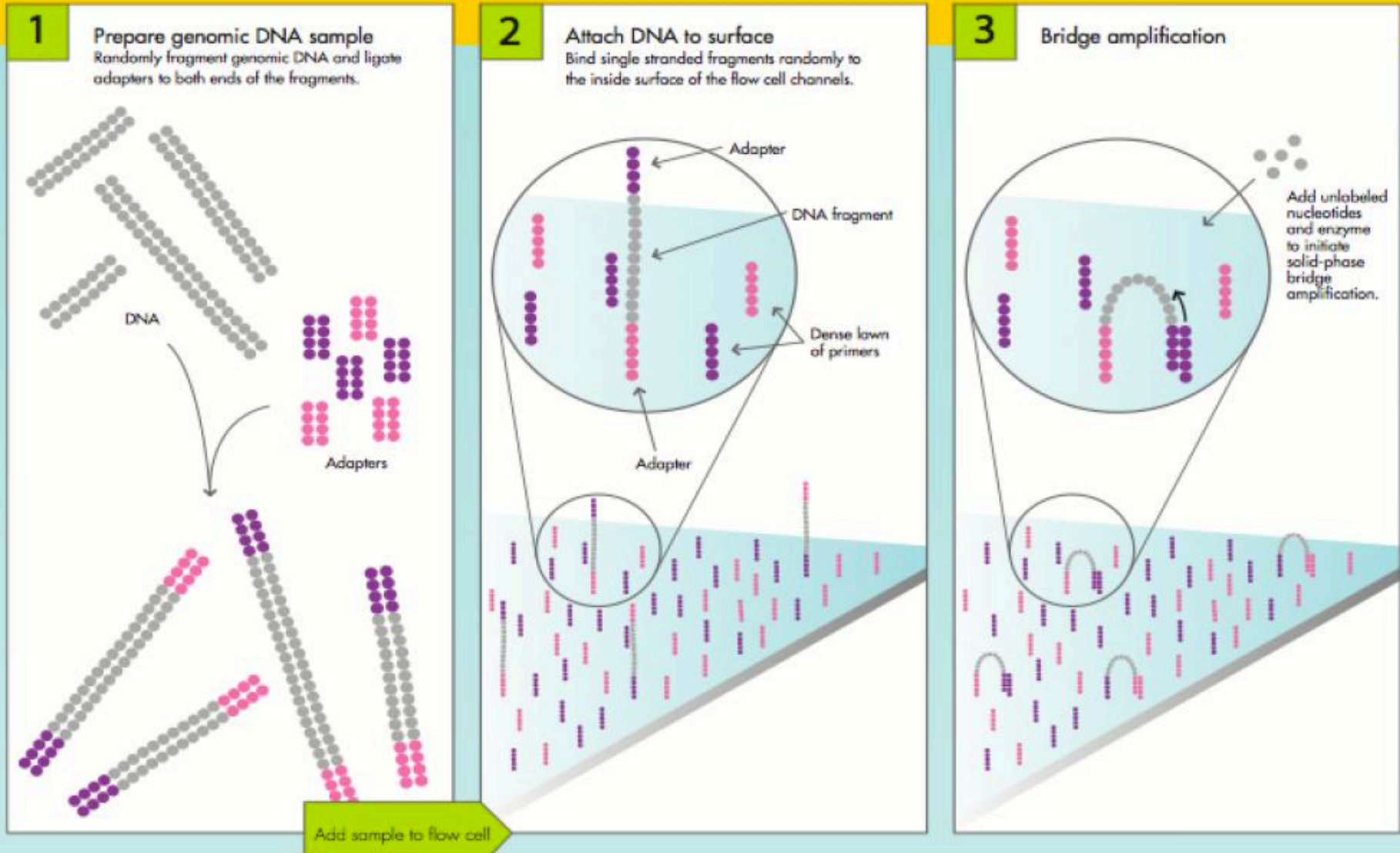
Smith College

Instructor: Prof. Sara Sheehan

Outline: 9/11

- Next-generation sequencing
- String alignment

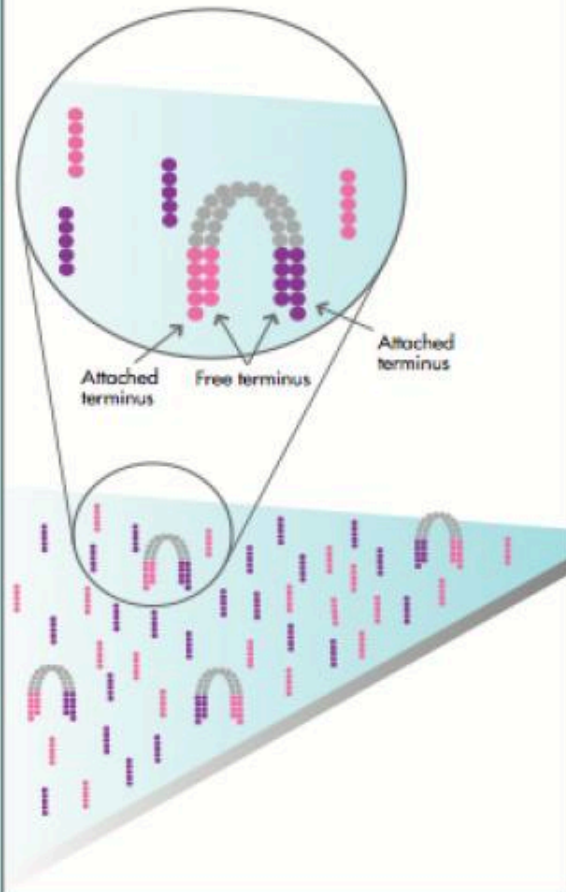
Next-generation sequencing 1-3



Next-generation sequencing 4-6

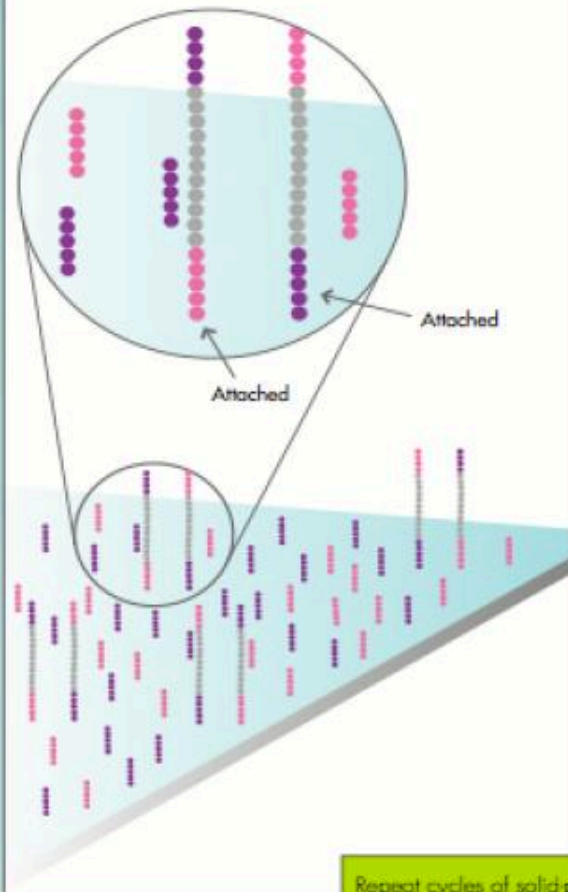
4

Fragments become double stranded



5

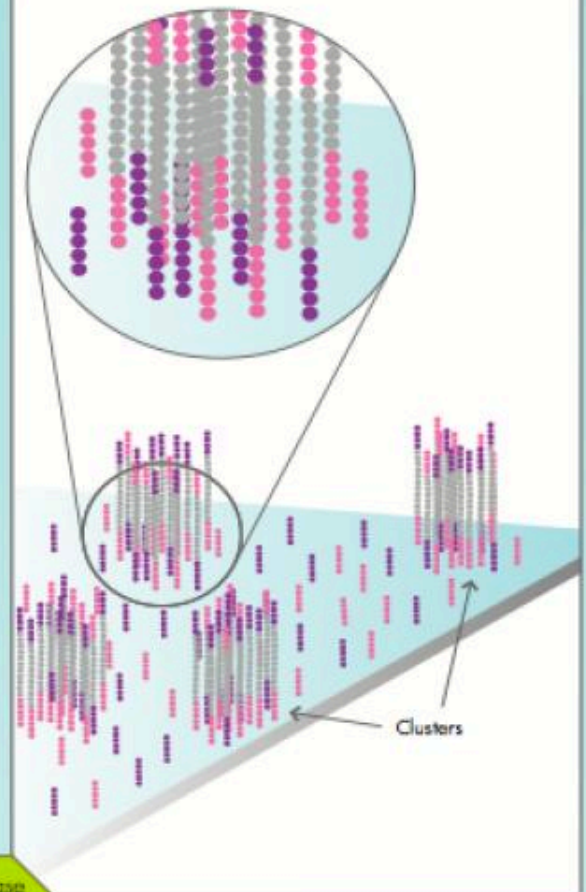
Denature the double stranded molecules



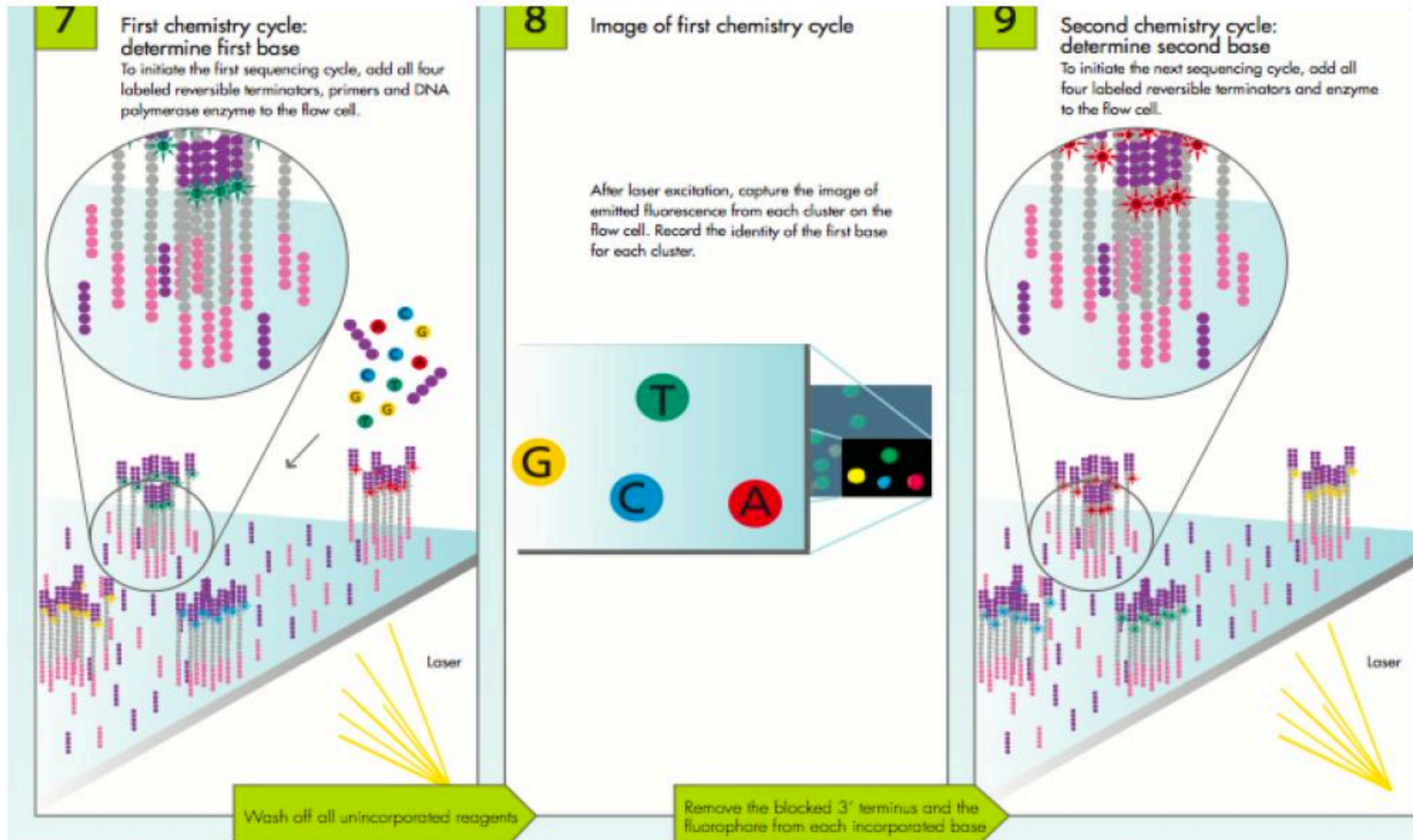
Repeat cycles of solid-phase bridge amplification

6

Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.



Next-generation sequencing 7-9



Next-generation sequencing 10-12

10

Image of second chemistry cycle is captured by the instrument

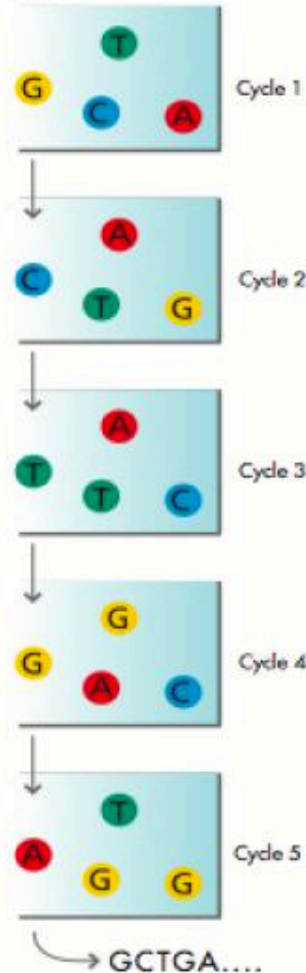
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.



11

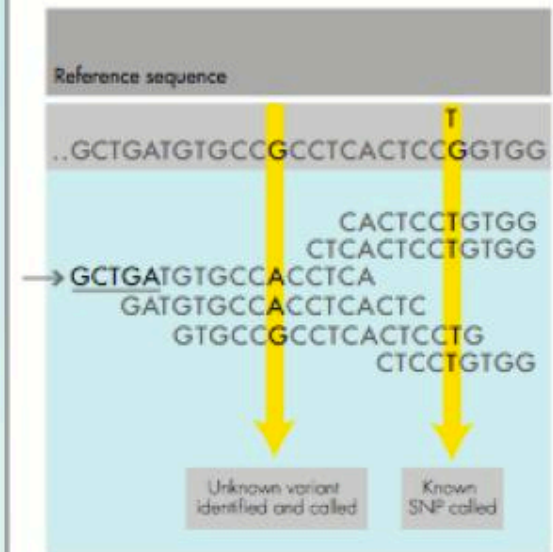
Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.



12

Align the new data to a reference and identify sequence differences



String Alignment

Example

- ACGGCTAGTTACG

- TCGTAGTATACCGA

- How should we line them up to get the best overlap?

Example alignments: human, chimp, macaque

	Start
Human	GTTTGTAGG - - - ATGTTCAACAAATGCTCCTTTCATTCTCTATTACAGACC TGCCGCA
Chimpanzee	GTTTGTAGG - - - ATGTTCAATAAATGCTGCTTTCATCTCTCTATTACAGACC TGCCGCA
Macaque	GTTTGTAGG - - - ATGCTCAATAAATGCTCCTTTCATTCTCTCATTACAAACT TGCCGCA
Human	GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee	GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque	GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Human	GATCTGGAGACTAA - CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Chimpanzee	GATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Macaque	TATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Human	CAGAATACGATTTAGCAAATTACTTCTTAAGATATTATTTTACATTTCTATATTCTCCTA
Chimpanzee	CAGAATACGATTTAGCAAATTACTTCTTAAGATACTATTTTACATTTCTATATTCTCCTA
Macaque	CAGAATATGATTTAGCAAATTACTTCTTAAGATATTATTTTGCAC TTCTATATTCTCCTA
Human	CCCTGAGTTGATGTGTGAGCAATATGTCACCTTTCATAAAGCCAGGTATACA - - - TTATG
Chimpanzee	CCCTGAGTTGATGTGTGAGCCGATATGTCACCTTTCATAAAGCCAGGTATACA - - - TTATG
Macaque	CCCTGAGTTGATGTGTGAGCAATATGTCACCTTCCACAAAGCCAGGTATATATACATTACG
	H I I Y S T F L S K
Human	GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCA AAAATTTTAAATTTT
Chimpanzee	GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAGAATTTTAAATTTT
Macaque	GACAGGTAAGTAAAAA - CATATTATTTATTCTAGGTTTTTGTCCAAGAGTTTAAATTTT
Human	AACGTGTTGCGCGTGTGTTGGTAA - - - TGTAAAACAAACTCAGTACA
Chimpanzee	AACGTGTTGCGCGTGTGTTGGTAA - - - TGTAAAACAAACTCAGTACA
Macaque	AACGTGTTGTGCAATGTGTTGGTAA - - - CGTAAAACAAATTCAGTACG