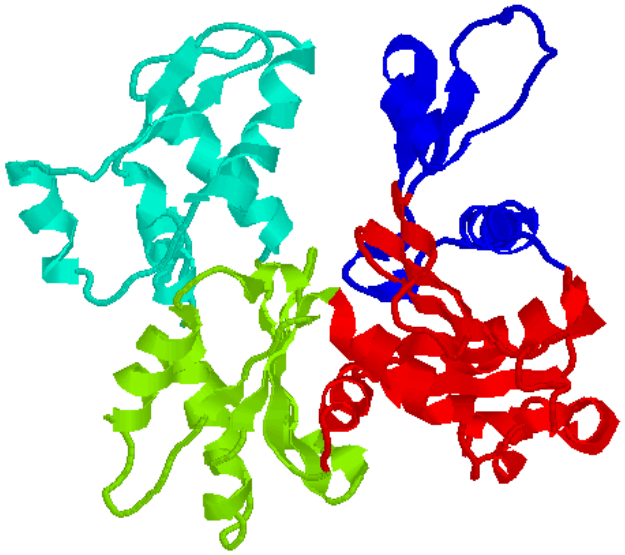
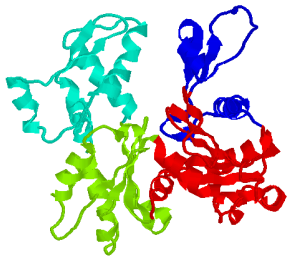
Genomic sequencing and its data analysis



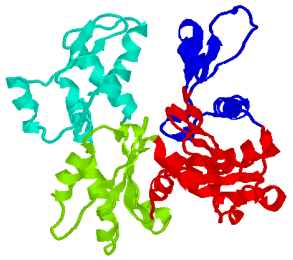
Dong Xu

Digital Biology Laboratory
Computer Science Department
Christopher S. Life Sciences Center
University of Missouri, Columbia
E-mail: xudong@missouri.edu
http://digbio.missouri.edu



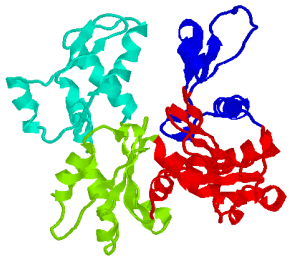
Lecture Outline

- Introduction to sequencing
- Next-generation sequencers
- Role of bioinformatics in sequencing
- Theory of sequence assembly
- Celera assembler
- Assembly of short reads



What is DNA Sequencing?

- A DNA sequence is the order of the bases on one strand.
- By convention, we order the DNA sequence from 5’ to 3’, from left to right.
- Often, only one strand of the DNA sequence is written, but usually both strands have been sequenced as a check.

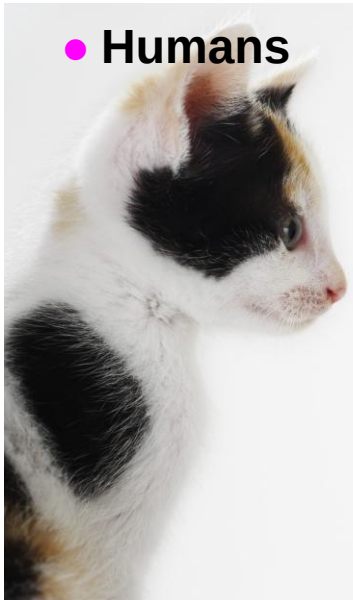


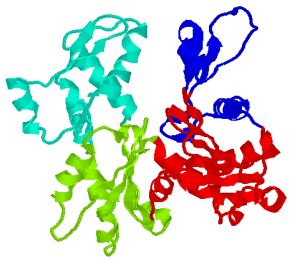
Sequencing

- Bacteria
- Fungi, yeast
- Insects: mosquito, fruit fly, moth, honey bee

- Plants: Arabidopsis, rice, corn, grapevine, …
- Animals: mouse, hedgehog, armadillo, cat, dog, horse, cow, elephant, platypus, …

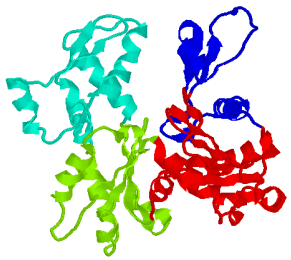
- Humans





Importance of Sequencing

- Basic blueprint for life
- Foundation of genomic studies
- Vision: personalized medicine
  - Genetic disorders
  - Diagnostics
  - Therapies
- $1000 genome



Lecture Outline

- Introduction to sequencing
- Next-generation sequencers
- Role of bioinformatics in sequencing
- Theory of sequence assembly
- Celera assembler
- Assembly of short reads

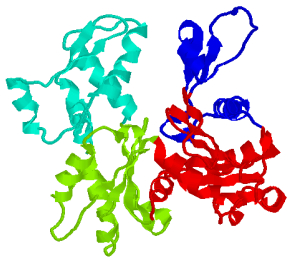
New Sequencers

Applied Biosystems
ABI 3730XL
1 Mb / day

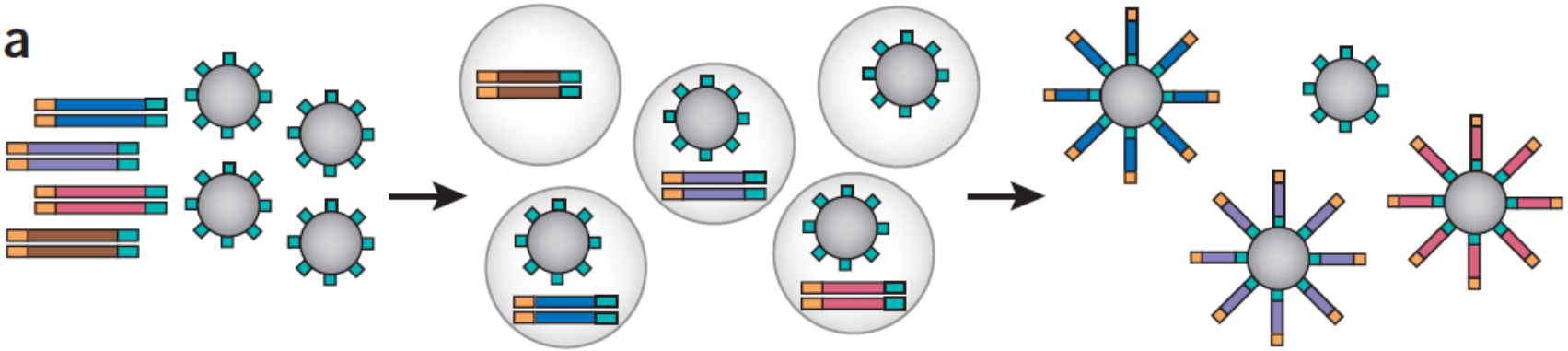
Roche / 454
Genome Sequencer FLX
100 Mb / run

Applied Biosystems
SOLiD
3000 Mb / run

Illumina / Solexa
Genetic Analyzer
2000 Mb / run

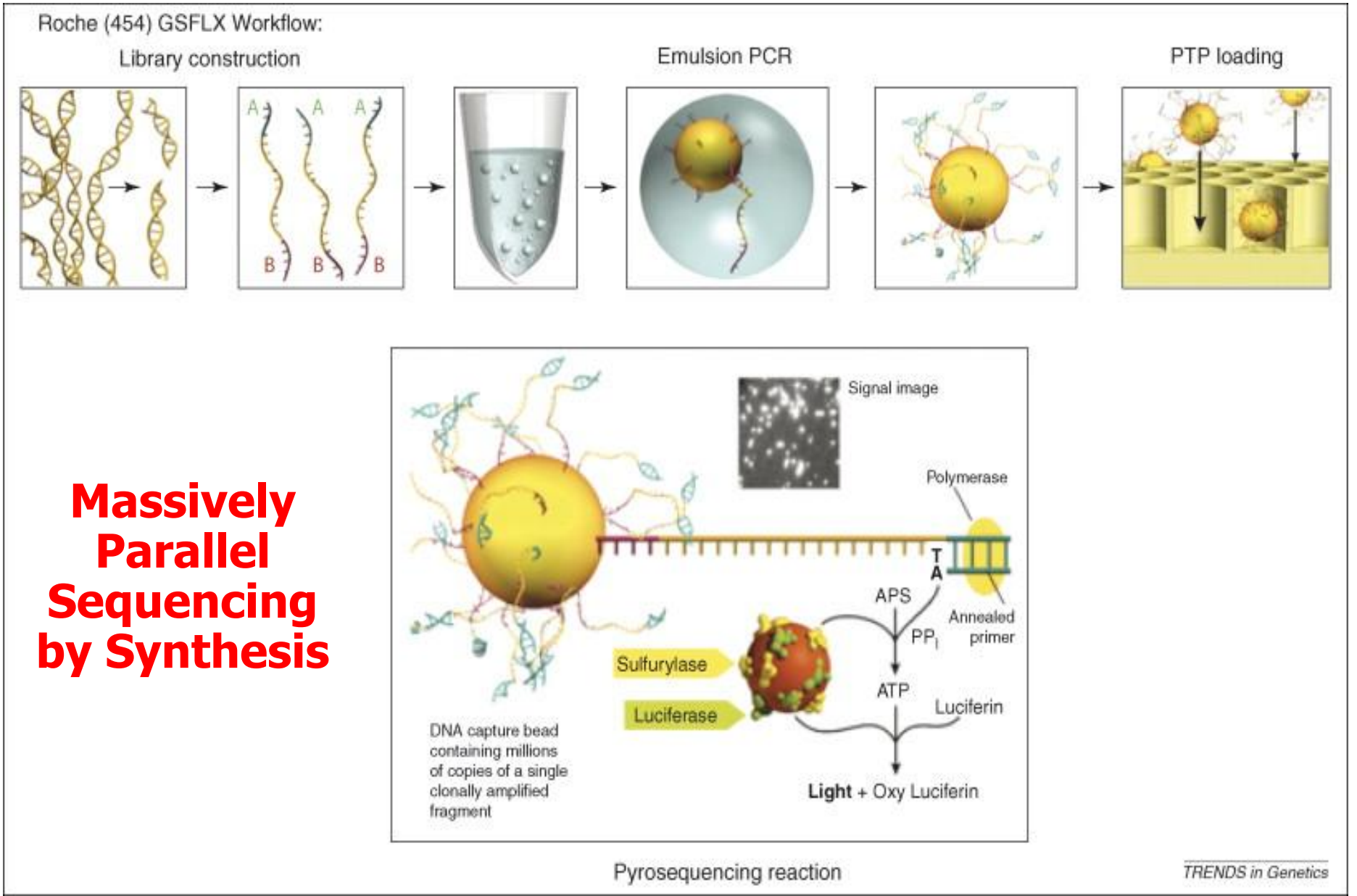


Emulsion PCR

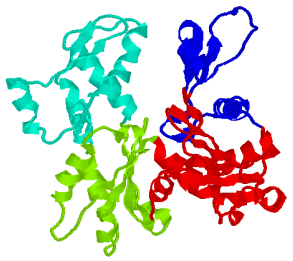


- Fragments, with adaptors, are PCR amplified within a water drop in oil.
- One primer is attached to the surface of a bead.
- Used by 454, Polonator and SOLiD.

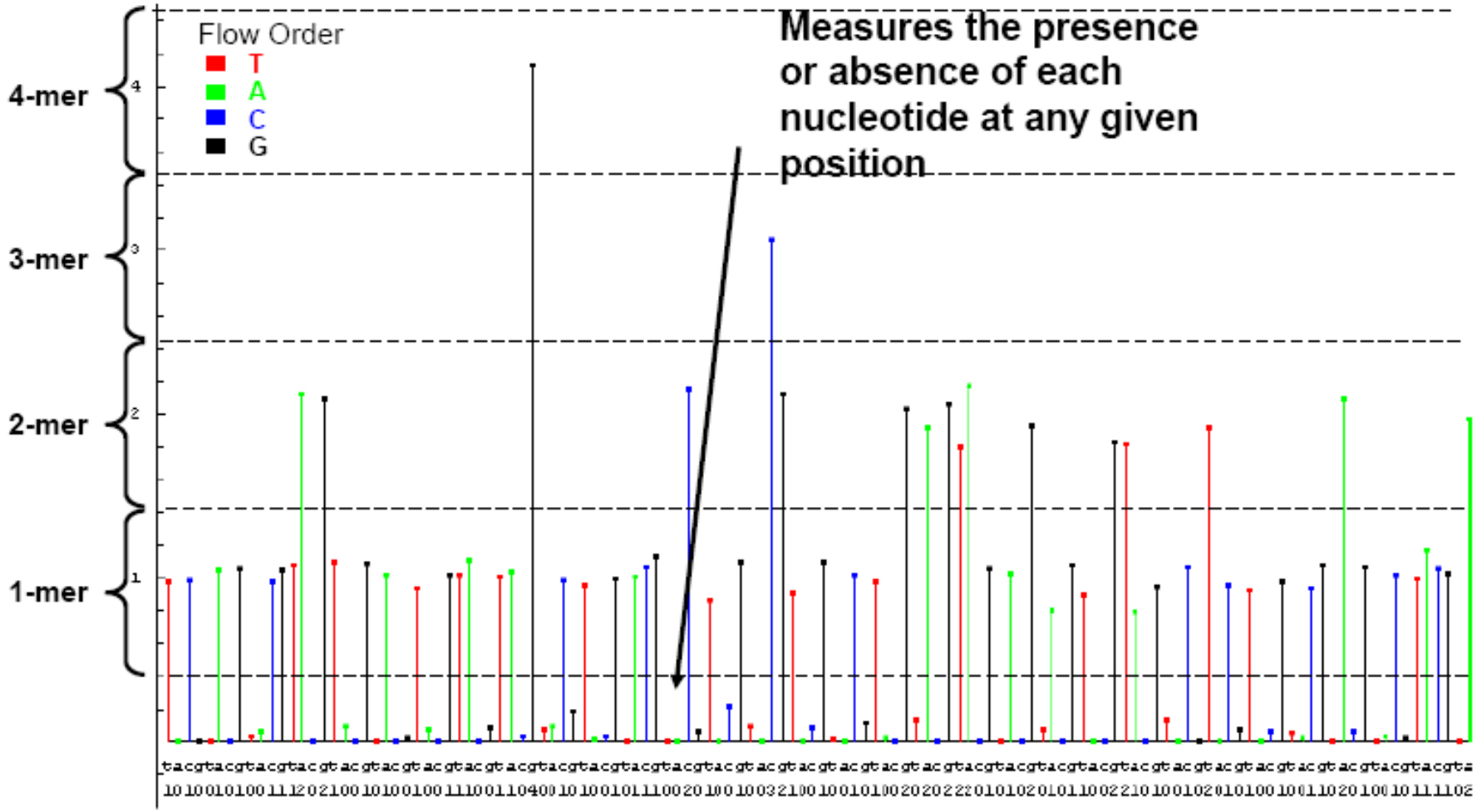
Roche (454) Workflow



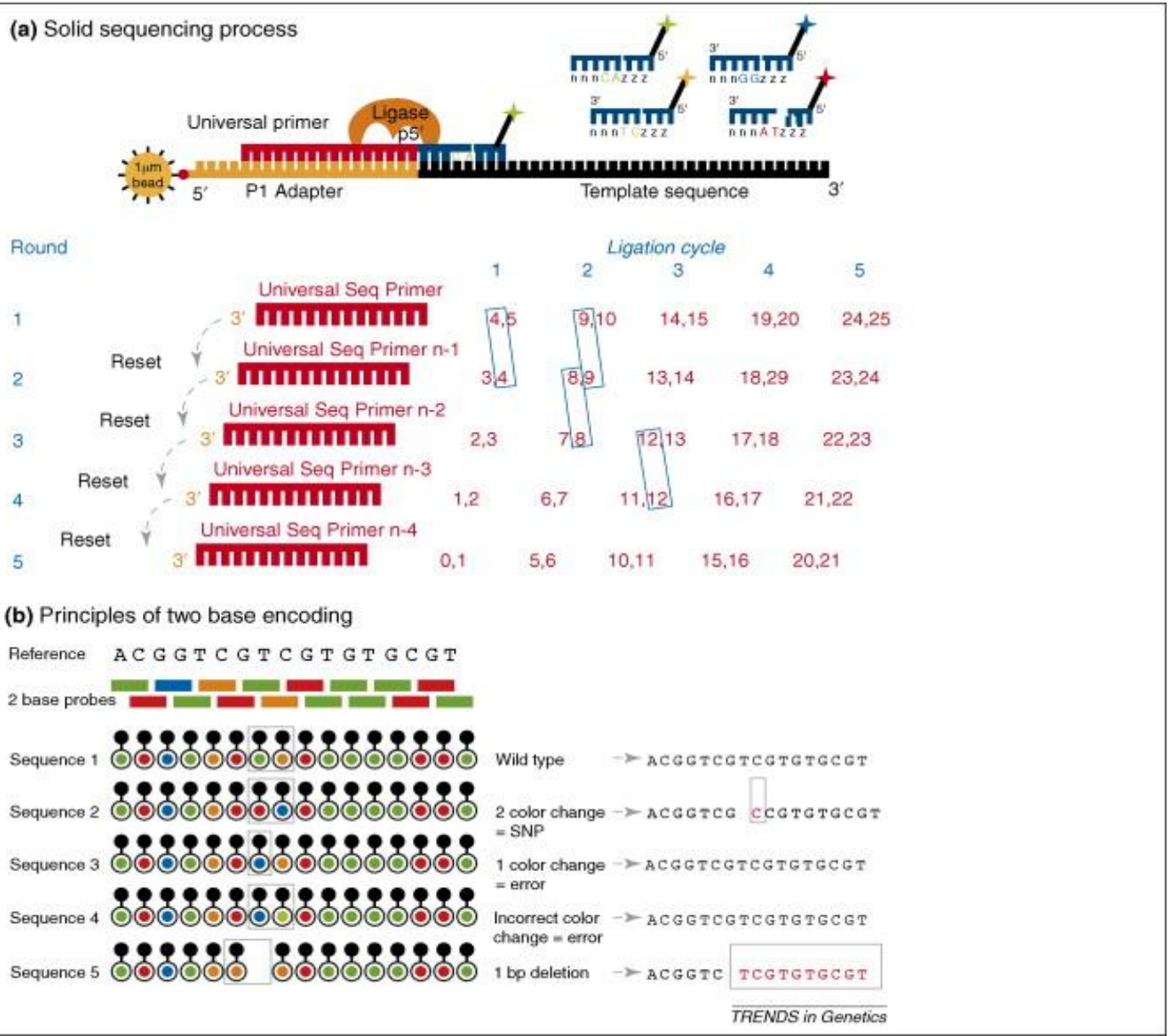
Massively Parallel Sequencing by Synthesis



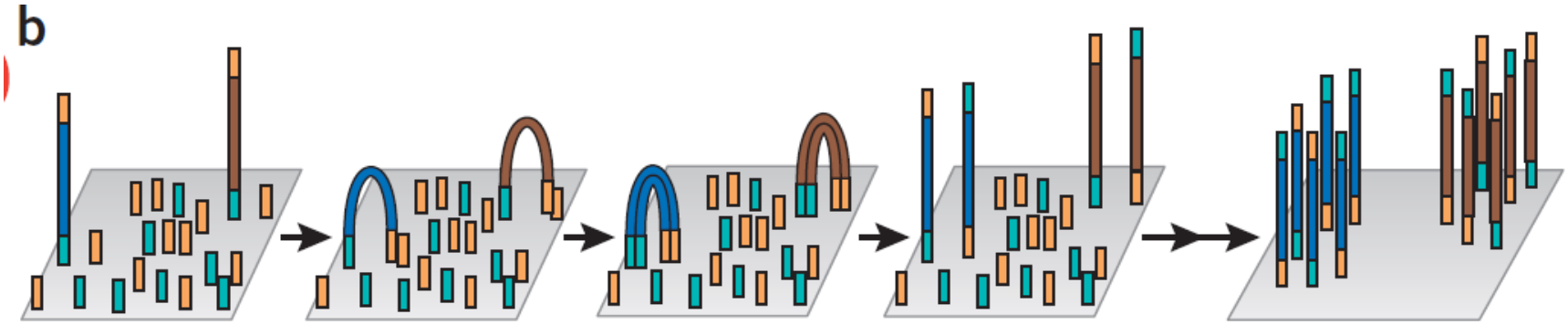
Flowgrams and Base Calling



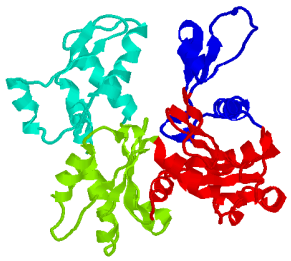
ABI SOLiD Workflow



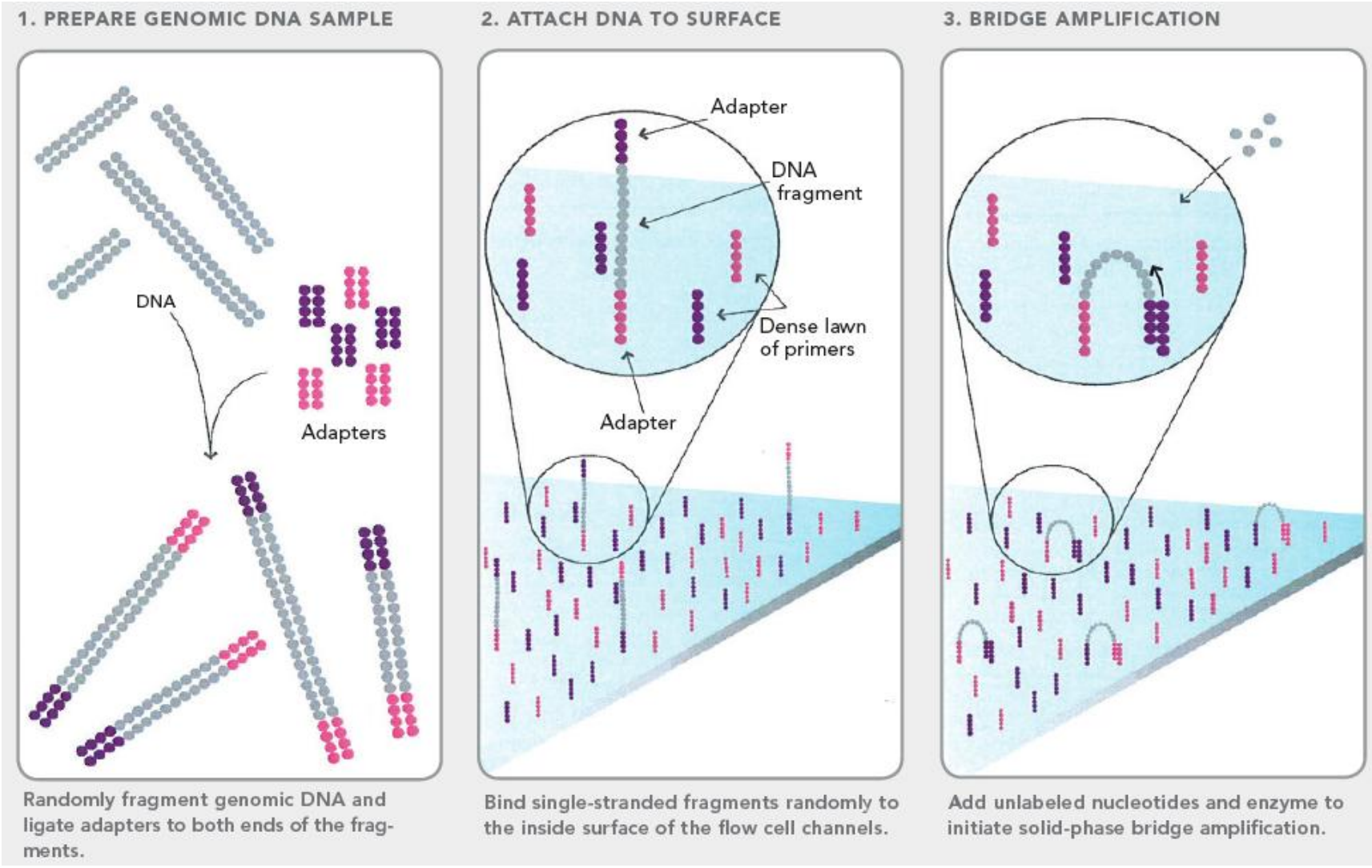
Bridge PCR

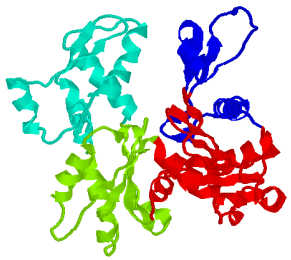


- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- Used by Solexa.

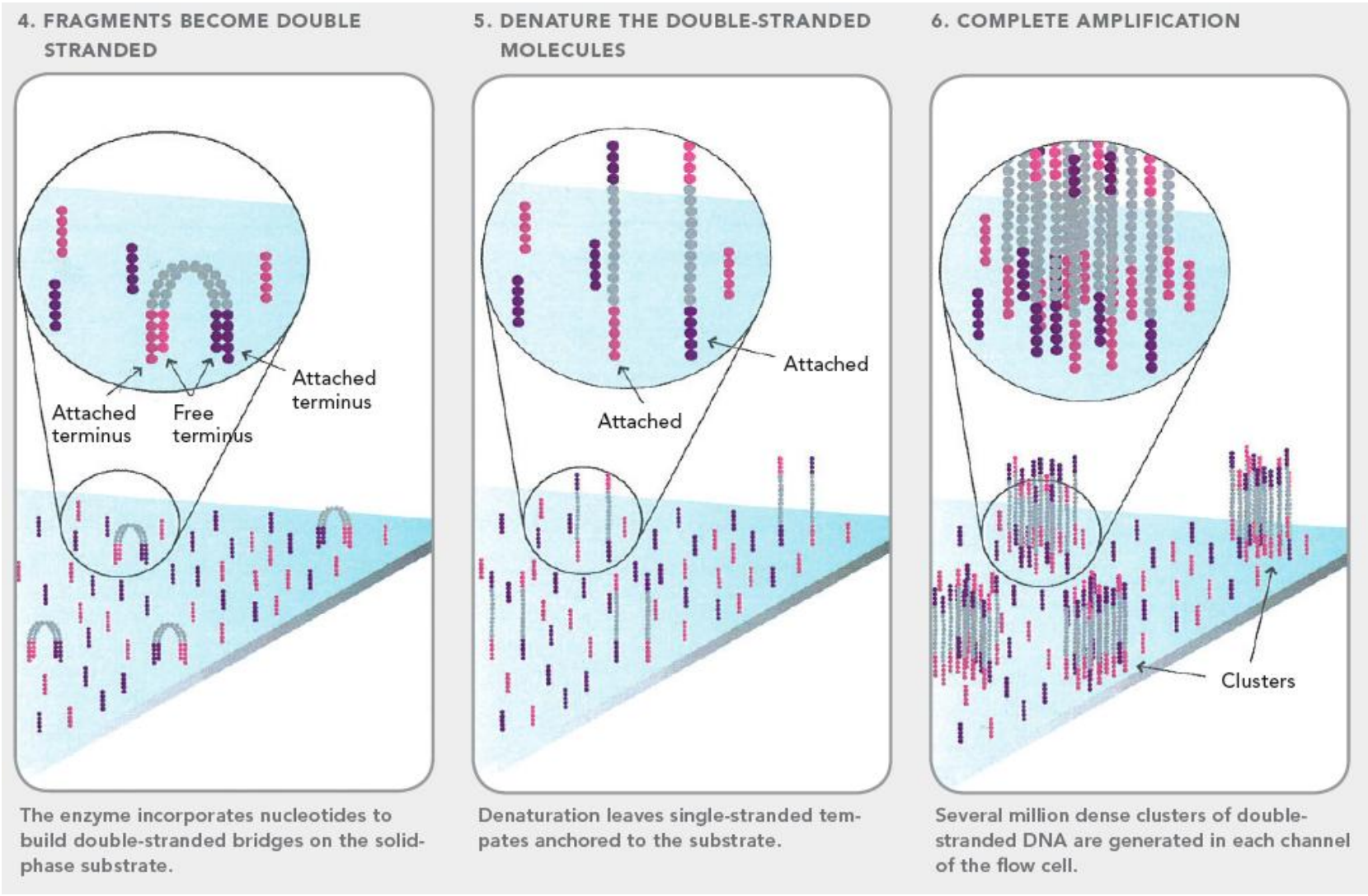


Illumina (Solexa) Workflow

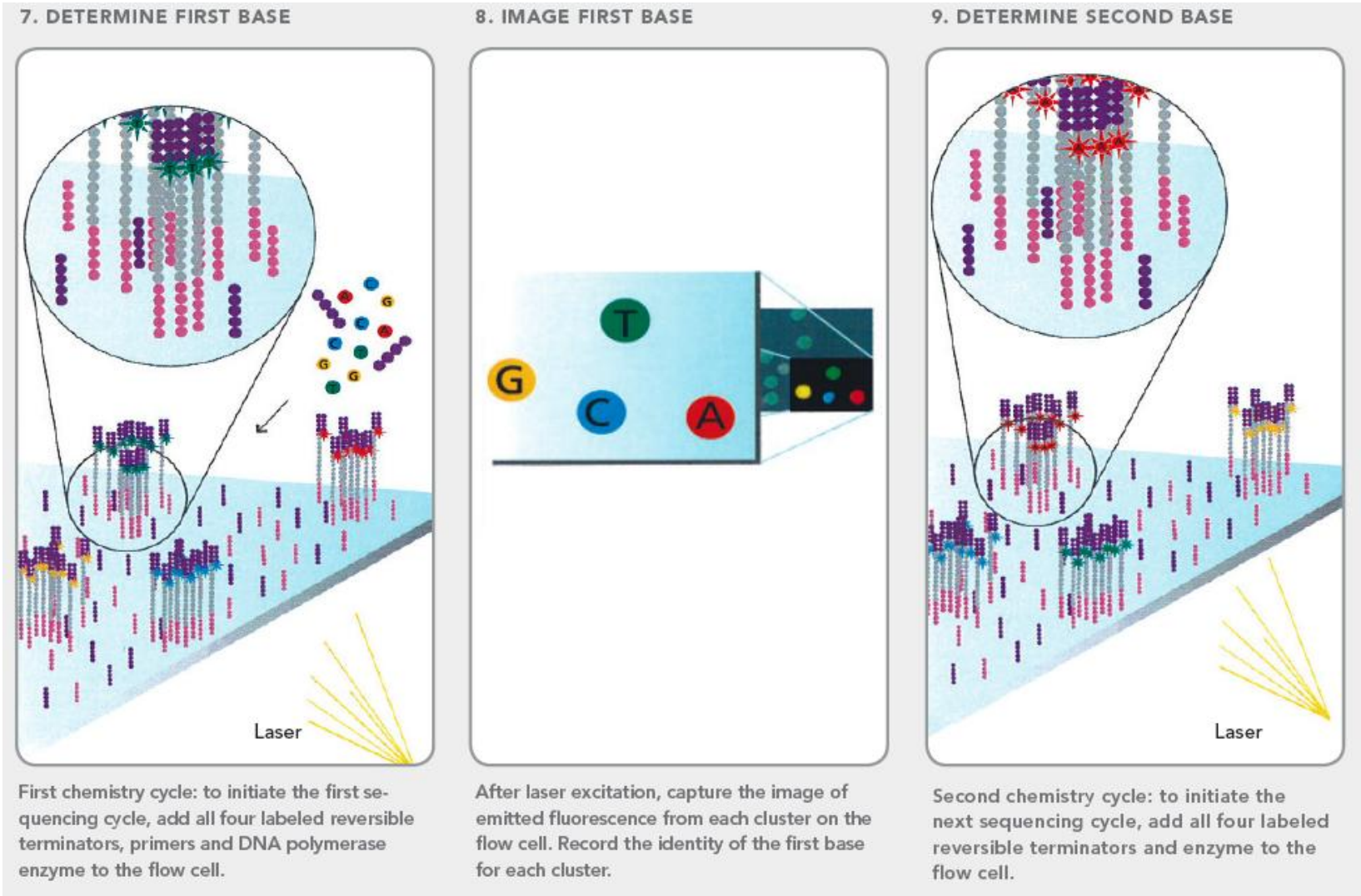




Illumina (Solexa) Workflow



Illumina (Solexa) Workflow

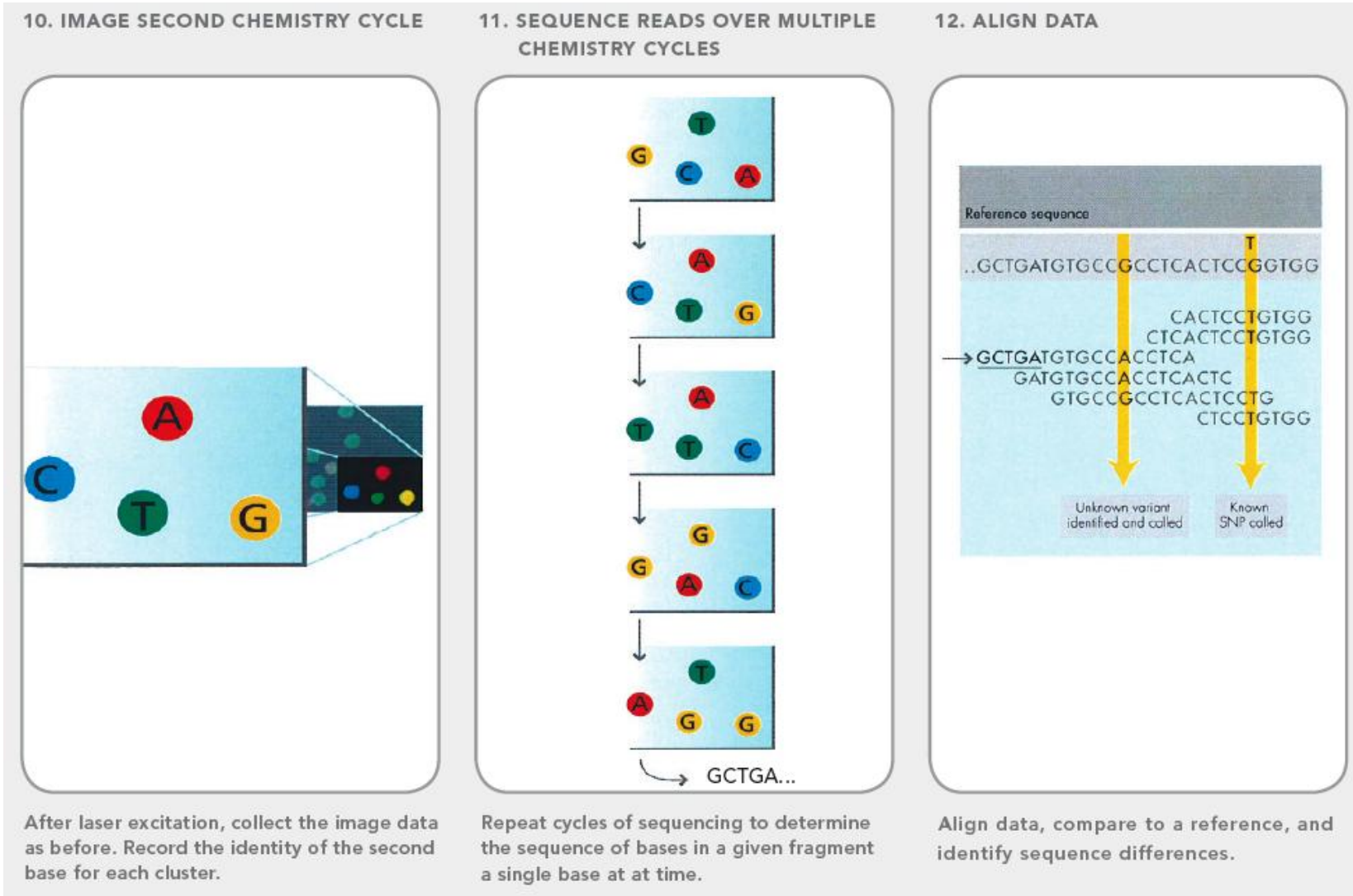


First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

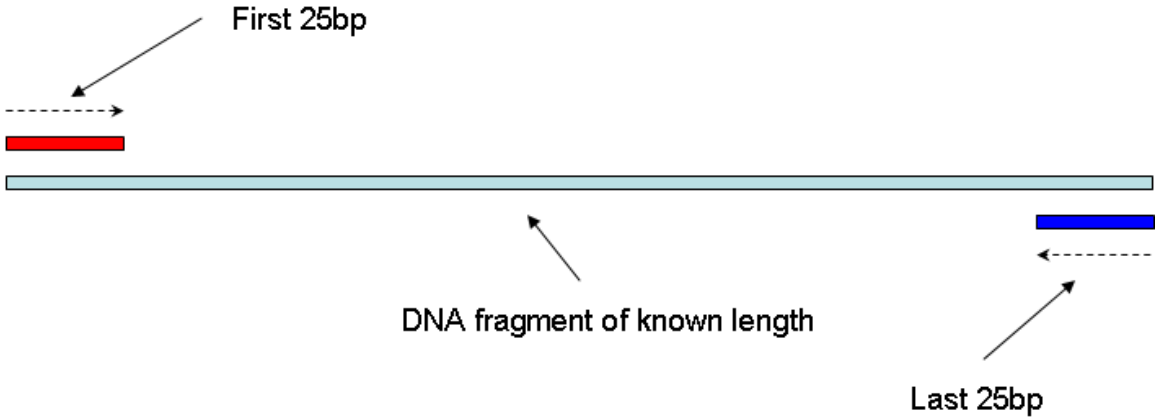
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Illumina (Solexa) Workflow

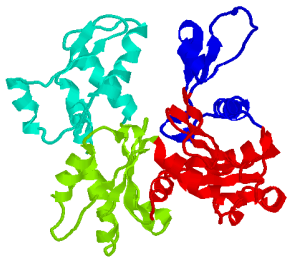


Pair-end Reads

- Paired-end sequencing (Mate pairs)
  - Sequence two ends of a fragment of known size.



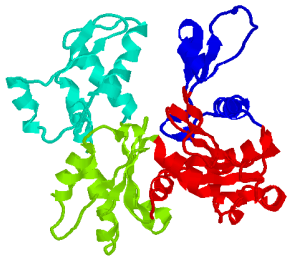
  - Currently fragment length (insert size) can range from 200 bps – 10,000 bps



Comparison of existing methods

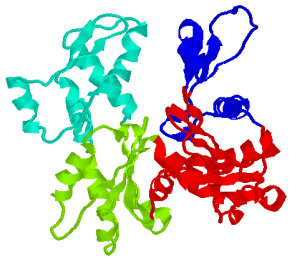
	Feature generation	Sequencing by synthesis
454	Emulsion PCR	Polymerase (pyrosequencing)
Solexa	Bridge PCR	Polymerase (reversible terminators)
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)
Polonator	Emulsion PCR	Ligase (nonamers)
HeliScope	Single molecule	Polymerase (asynchronous extensions)

	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length
454	~$60	$500,000	Yes	Indel	250 bp
Solexa	~$2	$430,000	Yes	Subst.	36 bp
SOLiD	~$2	$591,000	Yes	Subst.	35 bp
Polonator	~$1	$155,000	Yes	Subst.	13 bp
HeliScope	~$1	$1,350,000	Yes	Del	30 bp



Lecture Outline

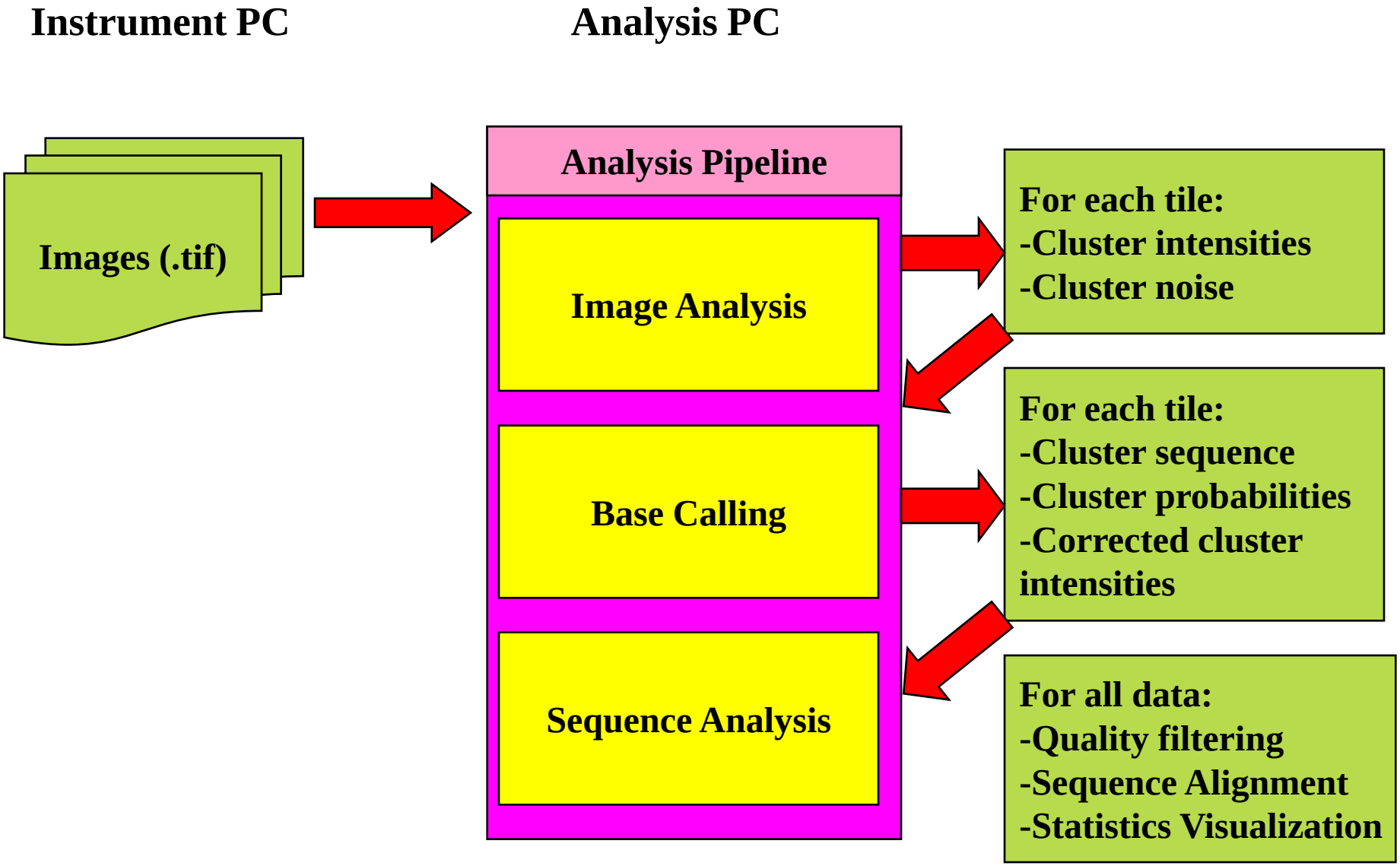
- Introduction to sequencing
- Next-generation sequencers
- Role of bioinformatics in sequencing
- Theory of sequence assembly
- Celera assembler
- Assembly of short reads

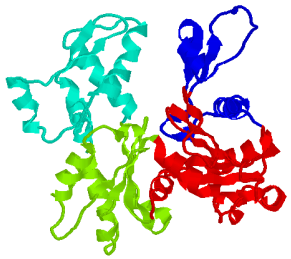


Analysis tasks

- Initial analysis: base calling
- Mapping to a reference genome
- *De novo* or assisted genome assembly
- SNP detection
- Transcriptome profiling
- DNA methylation studies
- CHIP-Seq

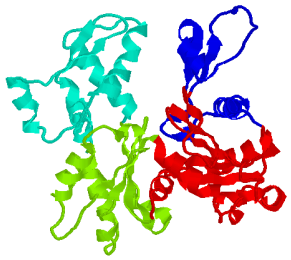
Initial Data Analysis workflow



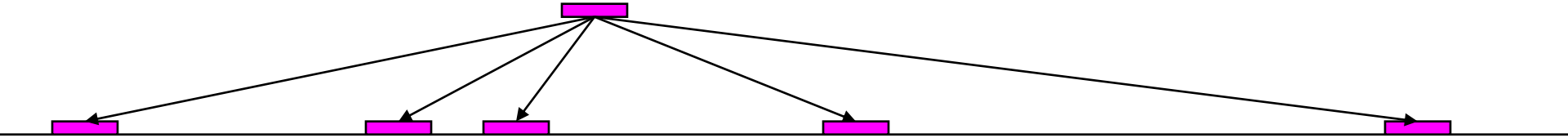


Short read mapping

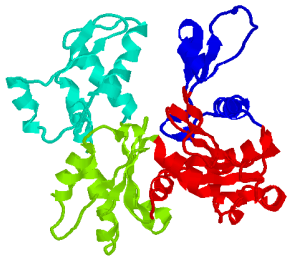
- Input:
  - A reference genome
  - A collection of many 25-100bp tags
  - User-specified parameters
- Output:
  - One or more genomic coordinates for each tag
- In practice, only 70-75% of tags successfully map to the reference genome.



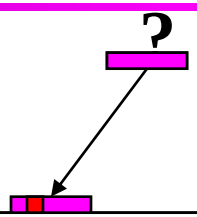
Multiple mapping



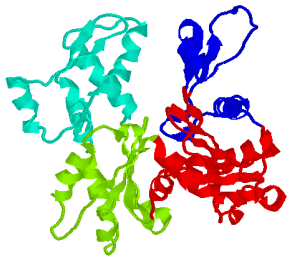
- A single tag may occur more than once in the reference genome.
- The user may choose to ignore tags that appear more than *n* times.
- As *n* gets large, you get more data, but also more noise in the data.



Inexact matching

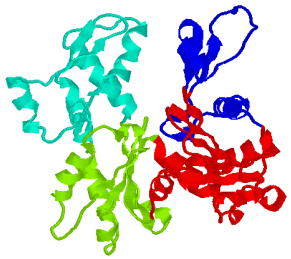


- An observed tag may not exactly match any position in the reference genome.
- Sometimes, the tag *almost* matches
- Such mismatches may represent a SNP or a bad read-out.
- The user can specify the maximum number of mismatches, or a quality score threshold.
- As the number of allowed mismatches goes up, the number of mapped tags increases, but so does the number of incorrectly mapped tags.



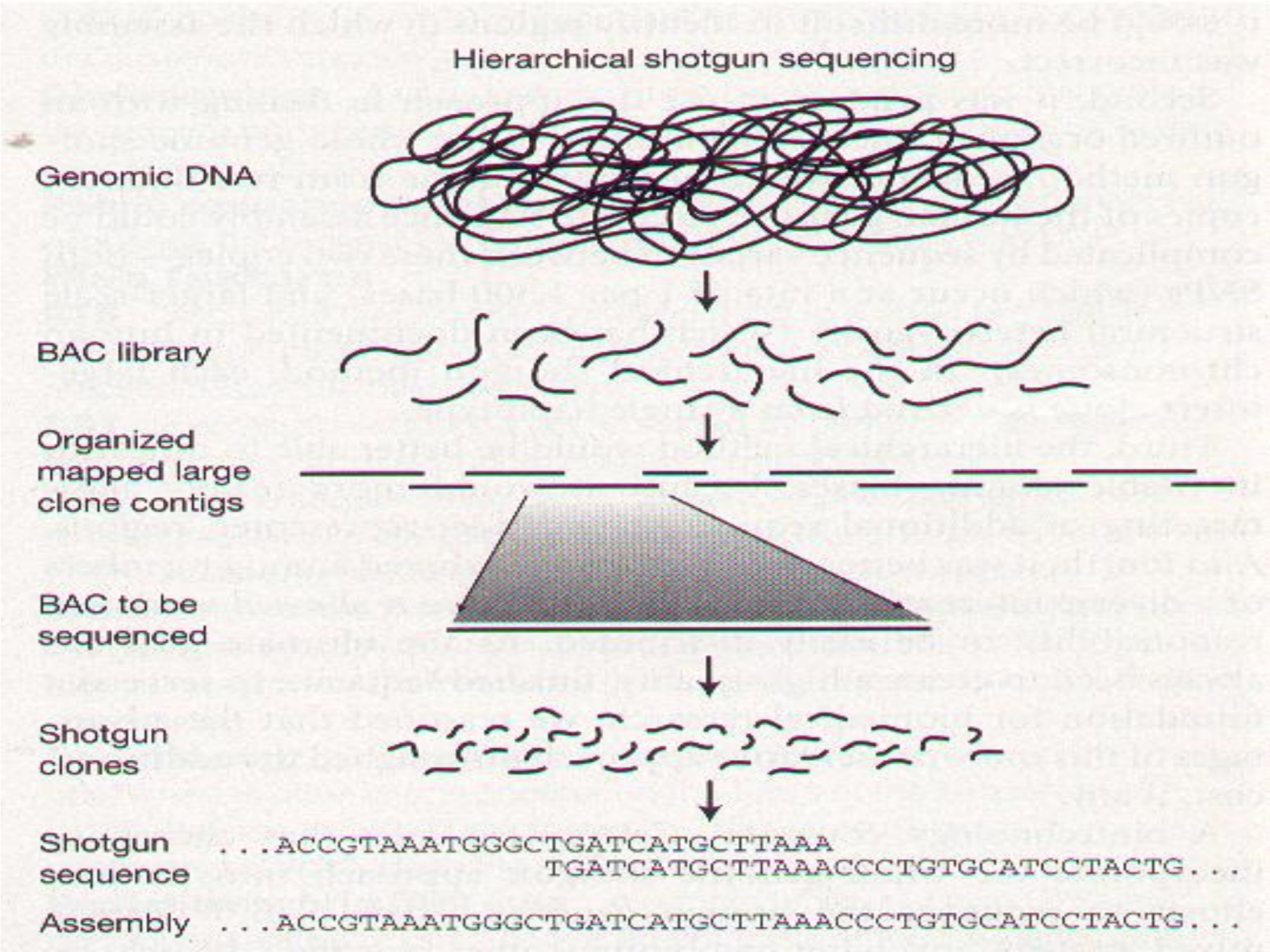
Short-read analysis software

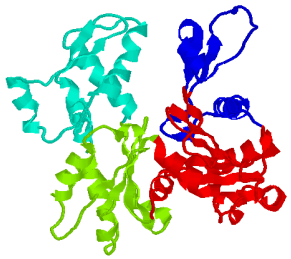
Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240



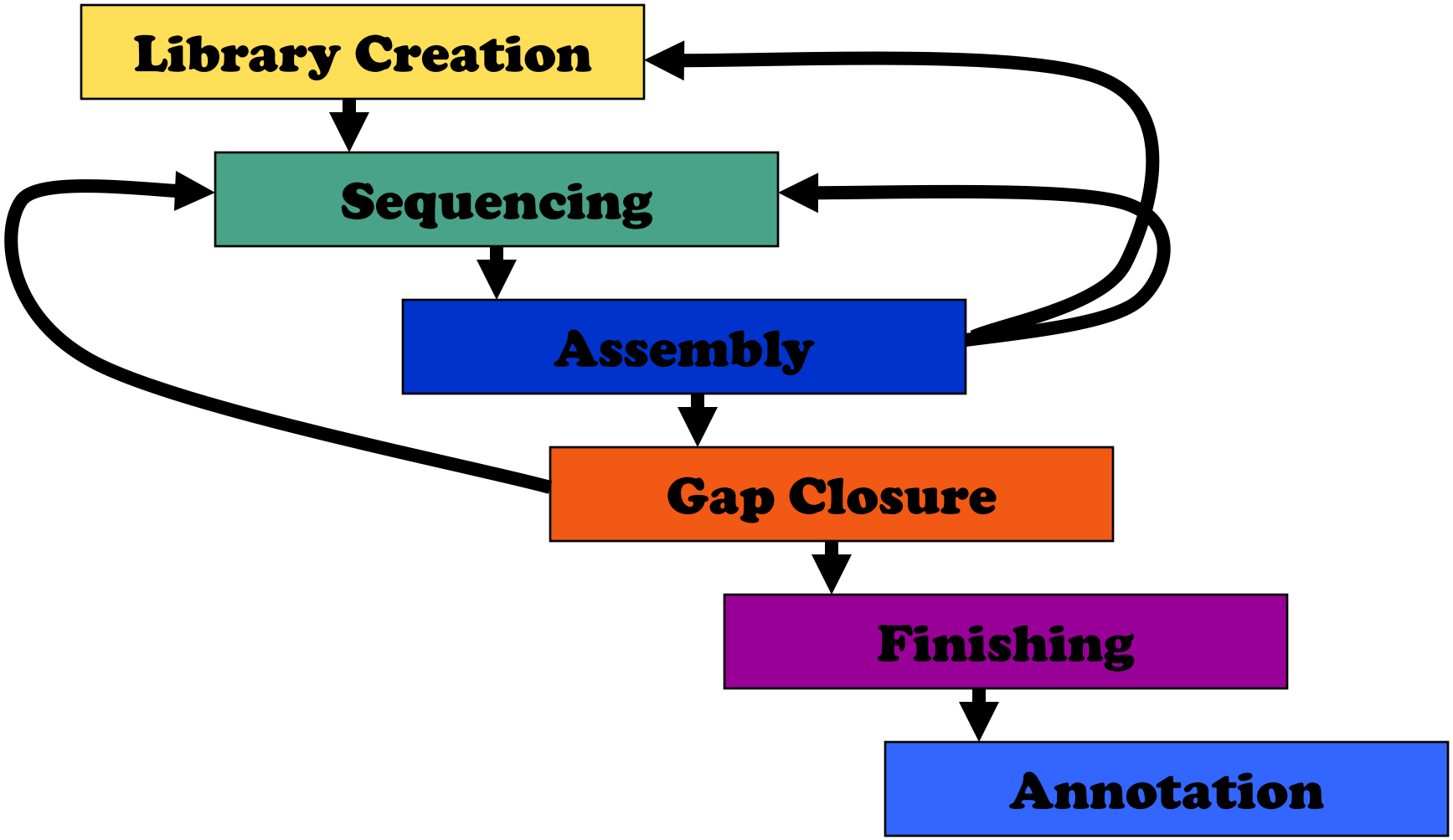
Lecture Outline

- Introduction to sequencing
- Next-generation sequencers
- Role of bioinformatics in sequencing
- Theory of sequence assembly
- Celera assembler
- Assembly of short reads


Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...



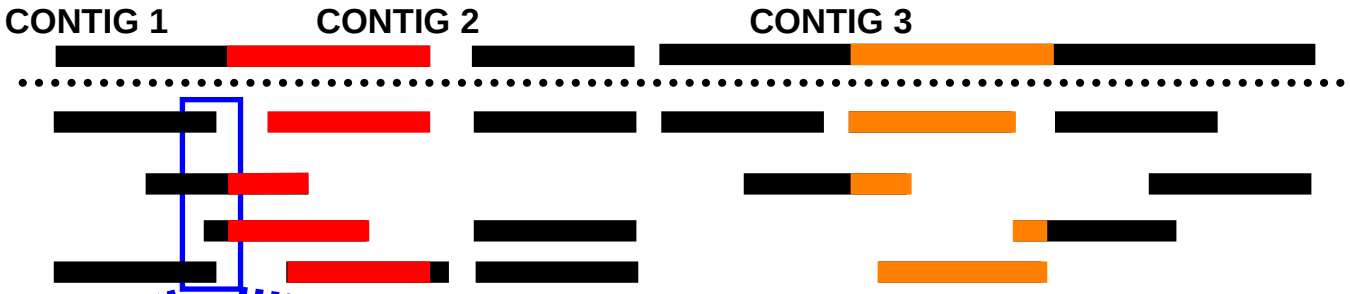
Sequencing Procedure



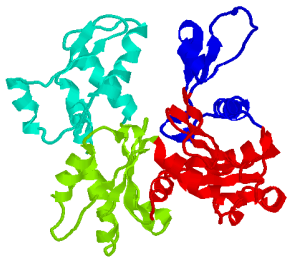
Genome Sequence Analysis - Step One
Assemble Sequences into Contigs

Sequenced fragmented DNA

Assembled DNA Sequence



AAACGCGATCGATCGATCGA
AAACGCGATCGATCGATCGATCGATCGATCGATCGTAG
CGATCGATCGATCGATCGTAG
AAACGCGATCGATCGATCGA



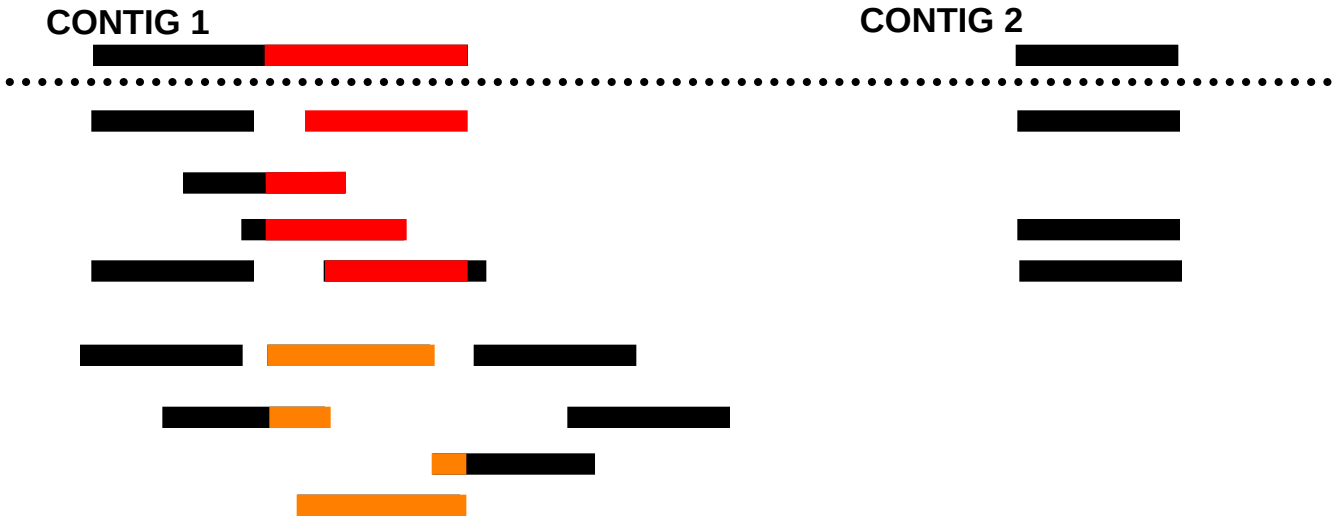
Repeat Problems

- Repeats at read ends can be assembled in multiple ways.

Genome Sequence Analysis - Step One
Initial Problem with Assembly

Sequenced fragmented DNA

Incorrectly Assembled DNA Sequence

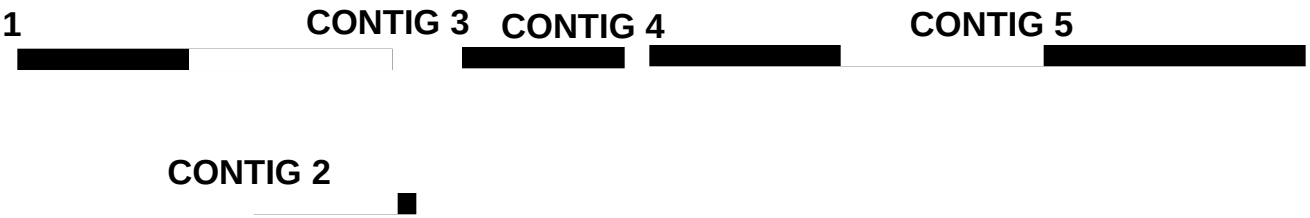


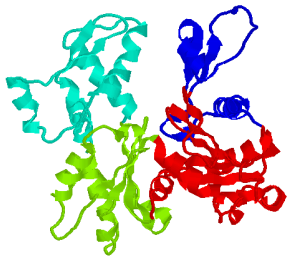
Genome Sequence Analysis - Step One
Need to Mask Repeats

Sequenced fragmented DNA

Masked DNA Sequence

Assembled DNA Sequence

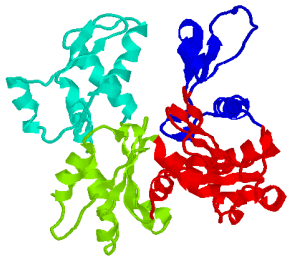




Lander-Waterman Model

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis“ Genomics 2 (3): 231- 239

- Poisson Estimate
- Number of reads
- Average length of a read
- Probability of base read



Lander–Waterman Assumptions

1. Sequencing reads will be randomly distributed in the genome

2. The ability to detect an overlap between two truly overlapping reads does not vary from clone to clone

Lander-Waterman Model

- Probability that a base is not represented: $P_0 = e^{-(LN/G)}$,or $P_0 = e^{-c}$
- L = read length
- N = number of reads
- G = target length (BAC, genome, etc)
- e = 2.718
- Remember that coverage is the total length of acquired sequence divided by the target length or LN/G
- Therefore, coverage is independent of the read length

Sequence coverage

The probability any base is NOT sequenced is given by:

$P_0 = e^{-c}$ where c = fold sequence coverage (c = LN/G),
LN = #bases sequenced, i.e. L = read length and N = # reads
G = length of template
and the constant, e = 2.718 (e = 2.718281828459)

c	$P_0=e^{-c}$	% not sequence	% sequenced (1- P_0)
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

Sum of all gaps

Total Gap Length = Ge^{-c} where c = Fold coverage,
G = target sequence length

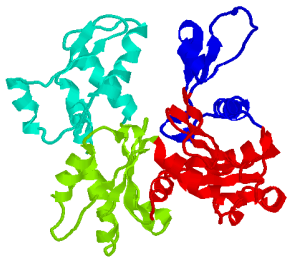
	target size					
	50kb	150kb	300kb	2Mb	4Mb	500Mb
c	Ge^{-c}	Ge^{-c}	Ge^{-c}	Ge^{-c}	Ge^{-c}	Ge^{-c}
1	18,500	55,500	111,000	740,000	1,480,000	185,000,000
2	6,750	20,250	40,500	270,000	540,000	67,500,000
3	2,500	7,500	15,000	100,000	200,000	25,000,000
4	900	2,700	5,400	36,000	72,000	9,000,000
5	335	1,005	2,010	13,400	26,800	3,350,000
6	125	375	750	5,000	10,000	1,250,000
7	45	135	270	1,800	3,600	450,000
8	15	45	90	600	1,200	150,000
9	5	15	30	200	400	50,000
10	2	6	12	90	180	20,000

Number of gaps

Number of Gaps = Ne^{-c}, where N = (G*c/L)

150kb Target Clone:

	Read Length					
	500			600		
c	N	e^{-c}	#Gaps=Ne^{-c}	N	e^{-c}	#Gaps=Ne^{-c}
1	300	0.37	111	250	0.37	93
2	600	0.135	81	500	0.135	68
3	900	0.05	45	750	0.05	38
4	1200	0.018	22	1000	0.018	18
5	1500	0.0067	10	1250	0.0067	8
6	1800	0.0025	5	1500	0.0025	4
7	2100	0.0009	2	1750	0.0009	2
8	2400	0.0003	1	2000	0.0003	1
9	2700	0.0001	0	2250	0.0001	0
10	3000	0.000045	0	2500	0.000045	0



In practice…

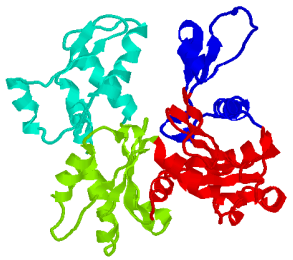
Lander-Waterman is almost always an underestimate

-cloning biases in shotgun libraries

-repeats

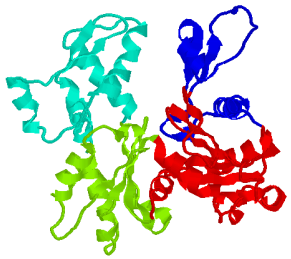
-GC/AT rich regions

-other low complexity regions



Sequence Assembly Algorithms

- Different than similarity searching
- Look for ungapped overlaps at end of fragments
  - (method of Wilbur and Lipman, (SIAM J. Appl. Math. 44; 557-567, 1984)
- High degree of identity over a short region
- Want to exclude chance matches, but not be thrown off by sequencing errors



Sequence Reconstruction Algorithm

- In the shotgun approach to sequencing, small fragments of DNA are reassembled back into the original sequence. This is an example of the Shortest Common Superstring (SCS) problem where we are given fragments and we wish to find the shortest sequence containing all the fragments.
- A superstring of the set *P* is a single string that contains every string in *P* as a substring.
- For example: for

  F1 = GCGC

  F2 = CGCC

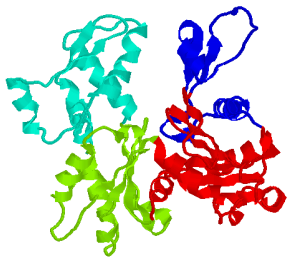
  F3 = GGCG

  The SCS is: GGCGCC

  F1 = GCGC

  F2 = CGCC

  F3 = GGCG



Greedy Algorithm for the Shortest Superstring Problem

- **The shortest superstring problem can be examined as a Hamiltonian path and is shown to be equivalent to the Traveling Salesman problem. The shortest superstring problem is NP-complete.**
- **A greedy algorithm exists that sequentially merges fragments starting with the pair with the most overlap first.**

Let T be the set of all fragments and let S be an empty set.

do {

For the pair (s,t) in T with maximum overlap. [s=t is allowed]

{

If s is different from t, merge s and t.

If s = t, remove s from T and add s to S.

}

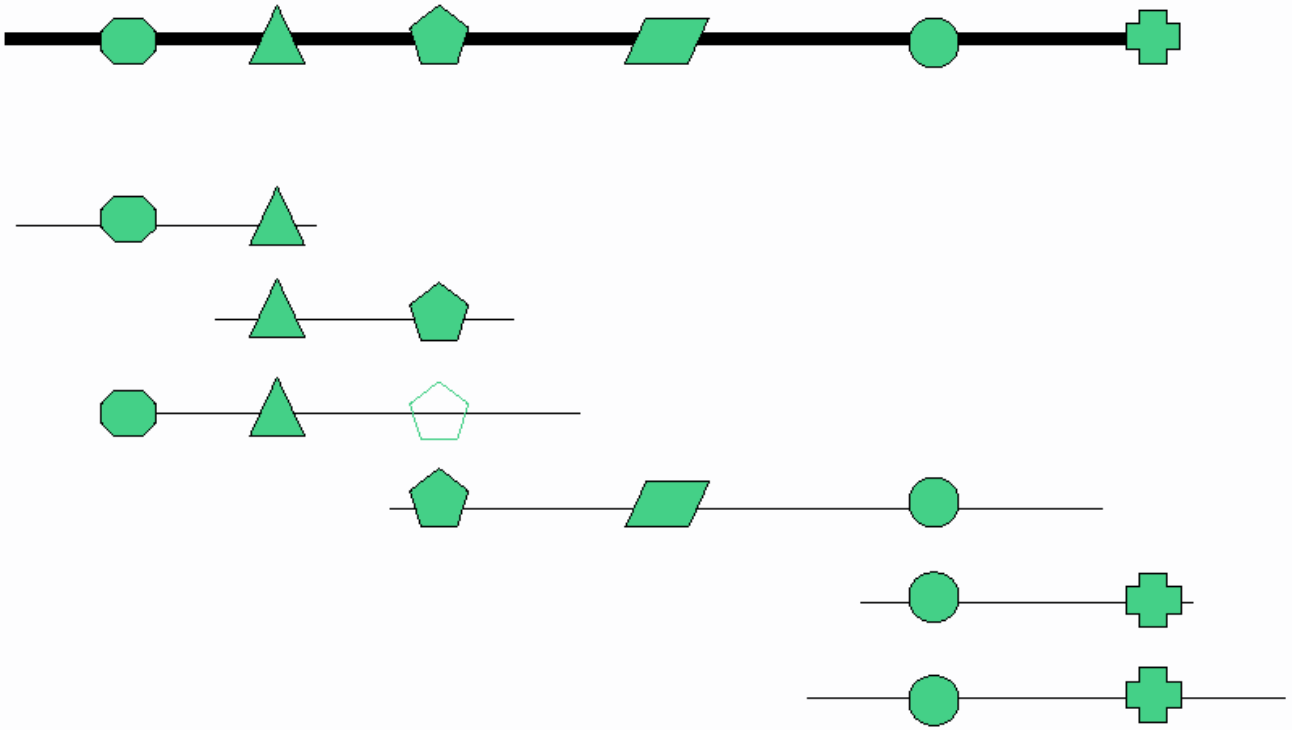
} while (T is not empty);

Output the concatenation of the elements of S.

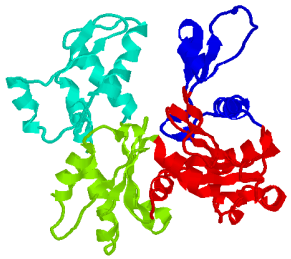
- **This greedy algorithm is of polynomial complexity and ignores the biological problems of: which direction a fragment is orientated, errors in data, insertions and deletions.**

STS content mapping

"Landmark mapping"

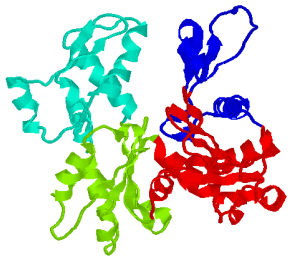


1. Test each clone for the presence or absence of each marker.
2. Assemble "contigs" based on shared marker content.



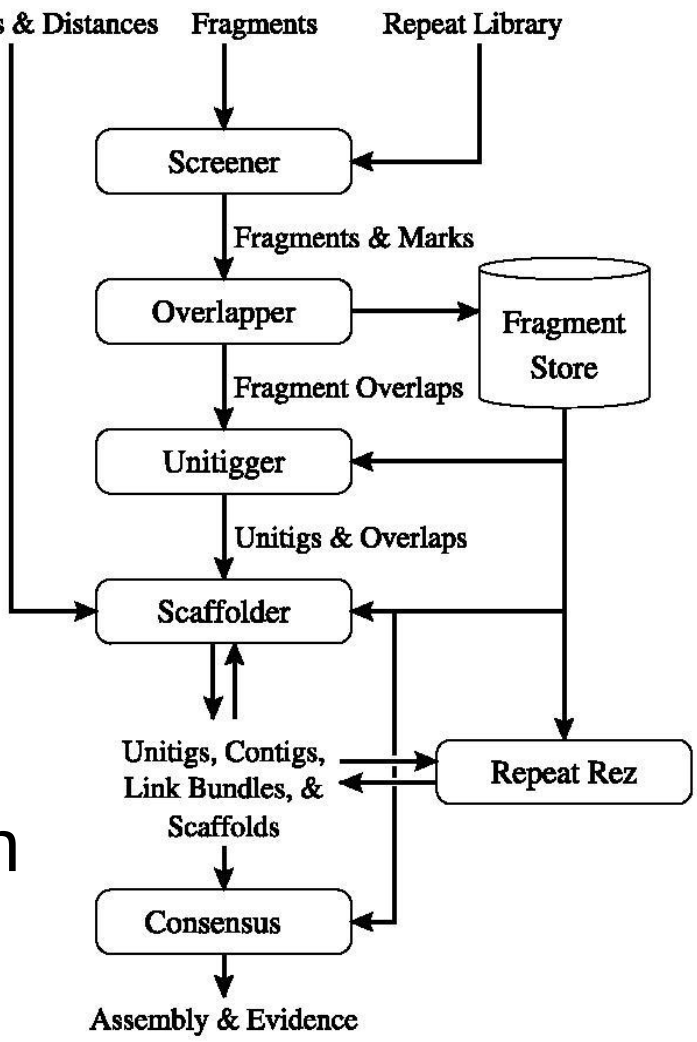
Lecture Outline

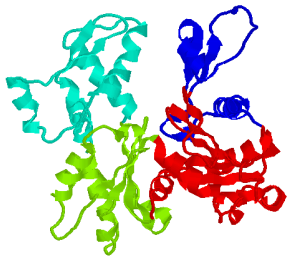
- Introduction to sequencing
- Next-generation sequencers
- Role of bioinformatics in sequencing
- Theory of sequence assembly
- Celera assembler
- Assembly of short reads



Celera Assembler

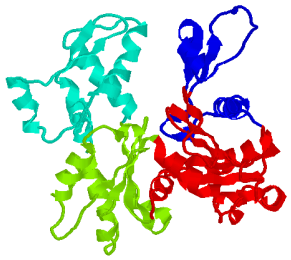
- Designed by Gene Myers, used to assemble the drosophilia, mouse and human genomes
- Steps:
  - Screener
  - Overlapper
  - Unitigger
  - Scaffolder & repeat resolution
  - Consensus





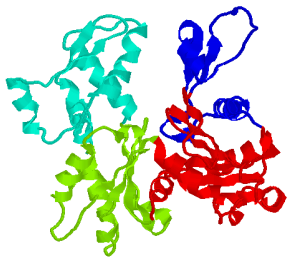
Screening reads

- Reads must be of very high reliability for assembly. Looking for 98%+ accuracy
- Vector contamination. Sequencing requires placing portions of the sequence to be determined in vectors (e.g. BACs or YACs). Need to avoid including any vector sequence
- Can also screen for known common repeats at this stage



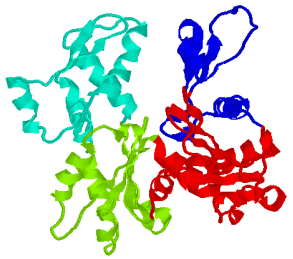
Overlapper

- Compare every fragment to every other
- Criterion: at least 40bp overlap with no more than 6% mismatches
- Probability of a chance overlap so low that all of these are either true overlaps or part of a repeated sequence (“repeat overlap”)
- Key objective is to distinguish between these two possibilities as early as possible in the assembly process.



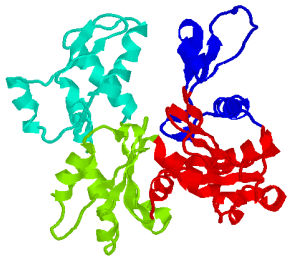
Unitigs

- Do the easy ones to assemble subset first.
- Fragments that have only one possible assembly are combined into longer sequences.
  - Reads which entirely match a subsegment of another
  - Fragment overlaps for which there are no conflicting overlaps
  - For Drosophila, 3.158M fragments collapse into 54,000 unitigs, going from 221M overlaps to



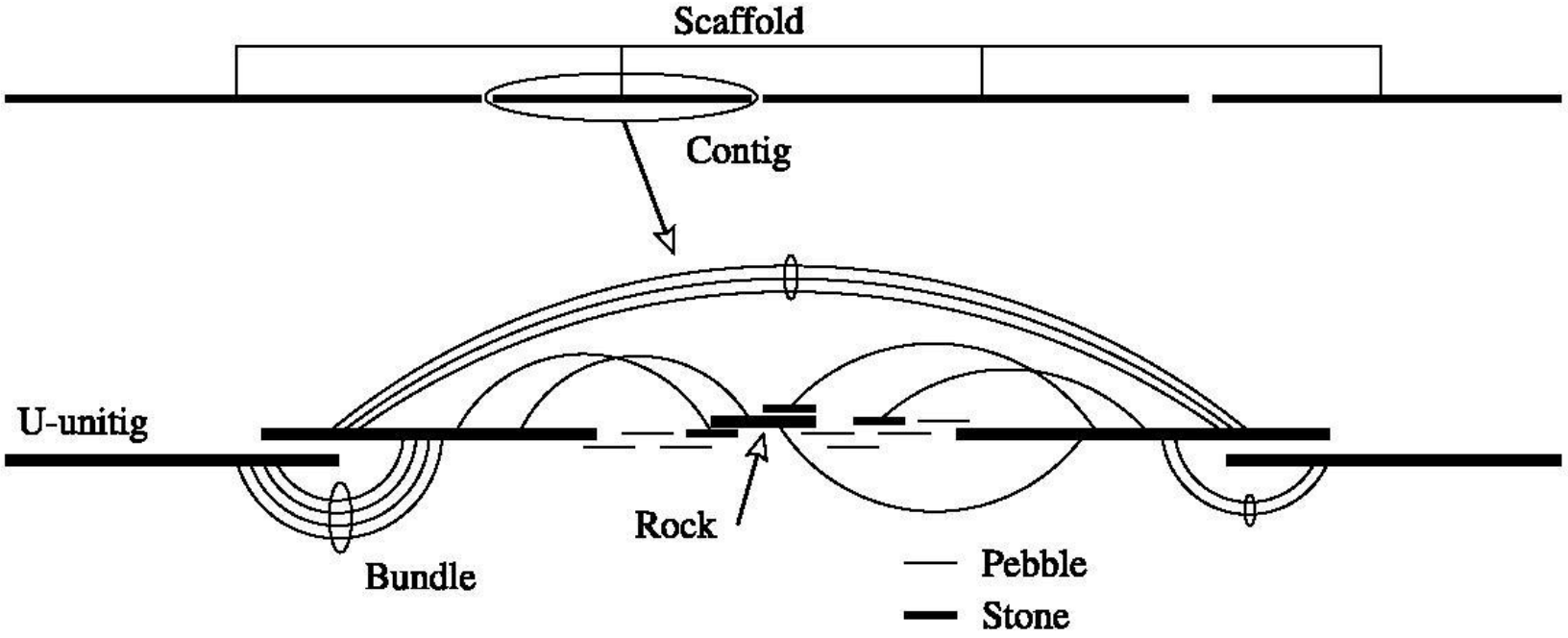
Celera Scaffolding

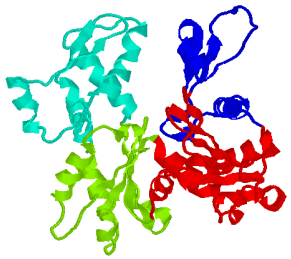
- Scaffold is a set of ordered, oriented contigs with gaps of approximately known size
- When the left and right reads of a mate are in different unitigs, their distance orients the unitigs and estimates the gap size.
- “Bundle” is a consistent set (2 or more) of mate pairs that place a pair of unitigs with respect to each other.



Scaffold picture

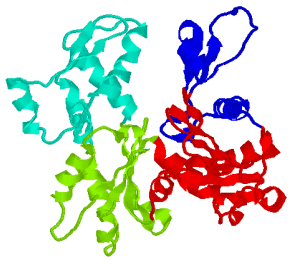
- At this point, errors are only in interiors of long repeating regions





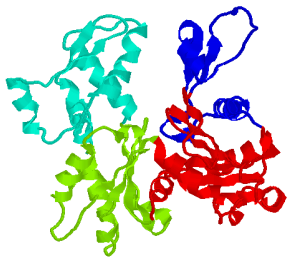
Lecture Outline

- Introduction to sequencing
- Next-generation sequencers
- Role of bioinformatics in sequencing
- Theory of sequence assembly
- Celera assembler
- Assembly of short reads



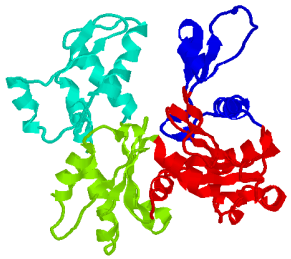
Assembly for short reads

- Challenging to assembly data.
- Short fragment length = very small overlap therefore many false overlaps (while reads are getting longer)
- Sequenced up to 100x coverage, increase in data size
- Pair-end reads are helpful



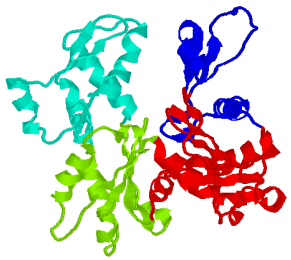
Current approaches

- Euler / De Bruijn approach.
- More suited for short read assembly.
- Implemented in *Velvet*, the mostly used short read assembly method at present (http://www.ebi.ac.uk/%7Ezerbino/velvet/)



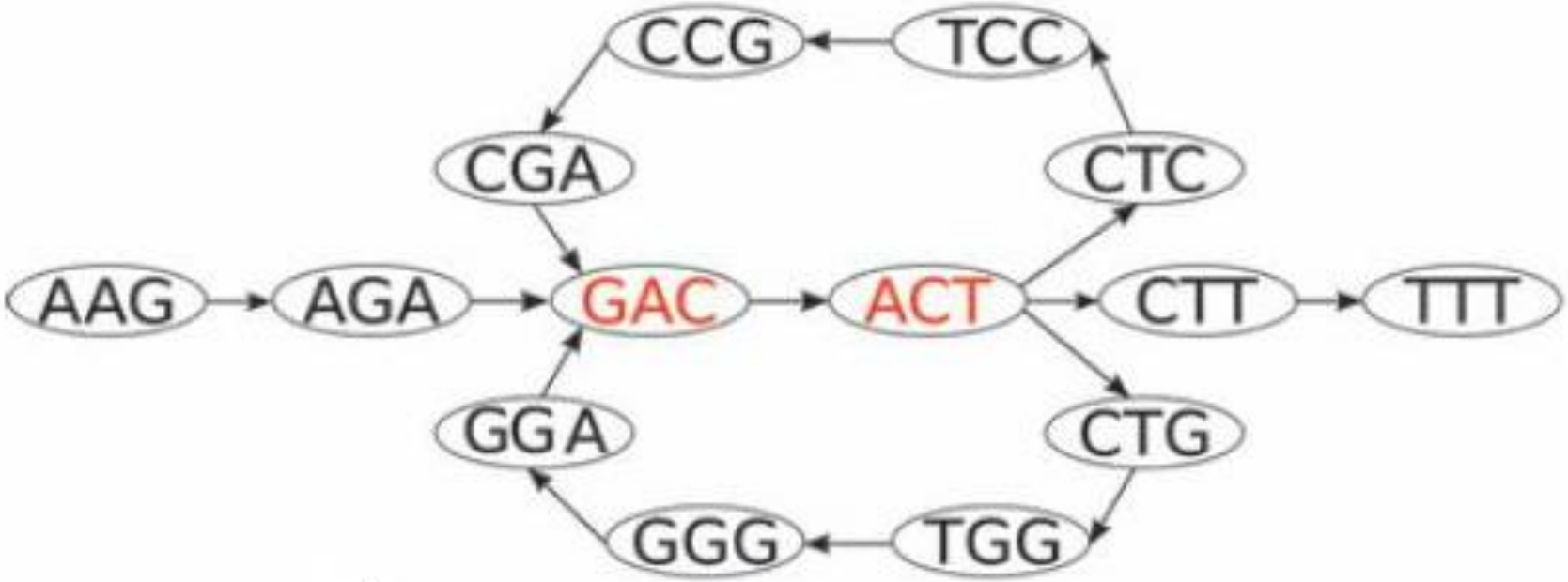
De Bruijn graph method

- Break each read sequence in to overlapping fragments of size *k.* (*k*-mers)
- Form De Bruijn graph such that each (*k-1)*-mer represents a node in the graph.
- Edge exists between node *a* to *b* iff there exists a *k*-mer such that is prefix is *a* and suffix is *b.*
- Traverse the graph in unambiguous path to form contigs.

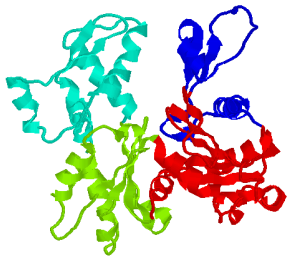


De Bruijn graph

AAGACTCCGACTGGGACTTT

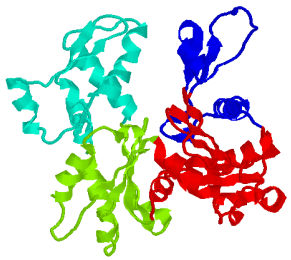


de Bruijn graph of a sequence



Summary

- Is most active research area (for the next 5-10 years)
- Data rich; high quality (digital vs. analog)
- Applicable to many studies
- Promising to personalized medicine
- Intensive developments for bioinformatics
- Fast evolving
- Assembly is challenging
- Using pair-end reads is essential



Homework

Read about the tools at

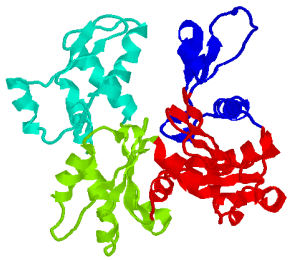
http://seqanswers.com/forums/showthread.php?t=43

Study Celera Assembler at

http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page

Study Verlet at

http://www.ebi.ac.uk/%7Ezerbino/velvet/



Acknowledgments

This file is for the educational purpose only. Some materials (including pictures and text) were taken from the Internet at the public domain.