Discussion #4

Osman Jamil

Nikita Khlystov

PART 1: Library generation (lecture 5) & transformation (lecture 7)

Review – 3 different kinds of libraries

I. Here are some data pertaining to the *Arabidopsis* genome (Source: Wortman, J.R. et al. *Plant Physiology* **132:**461-468 (2003)):

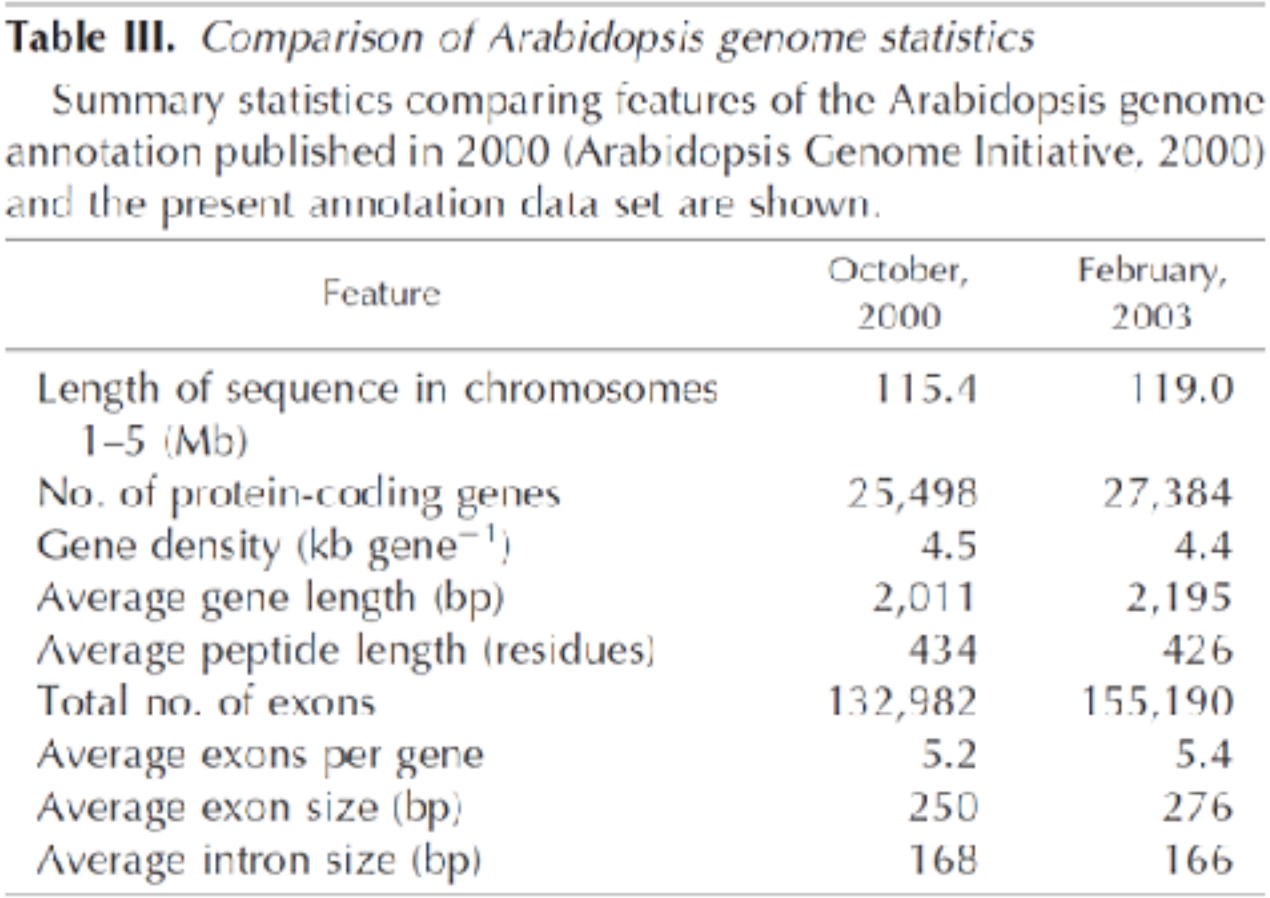


Table III. *Comparison of Arabidopsis genome statistics*

Summary statistics comparing features of the Arabidopsis genome annotation published in 2000 (Arabidopsis Genome Initiative, 2000) and the present annotation data set are shown.

Feature	October, 2000	February, 2003
Length of sequence in chromosomes 1–5 (Mb)	115.4	119.0
No. of protein-coding genes	25,498	27,384
Gene density (kb $gene^{-1}$)	4.5	4.4
Average gene length (bp)	2,011	2,195
Average peptide length (residues)	434	426
Total no. of exons	132,982	155,190
Average exons per gene	5.2	5.4
Average exon size (bp)	250	276
Average intron size (bp)	168	166

(Please use the 2003 annotation data in answering the questions below)

A) If *Arabidopsis* were a prokaryote, given the average peptide length, what would the average gene length be (including the stop codon)?

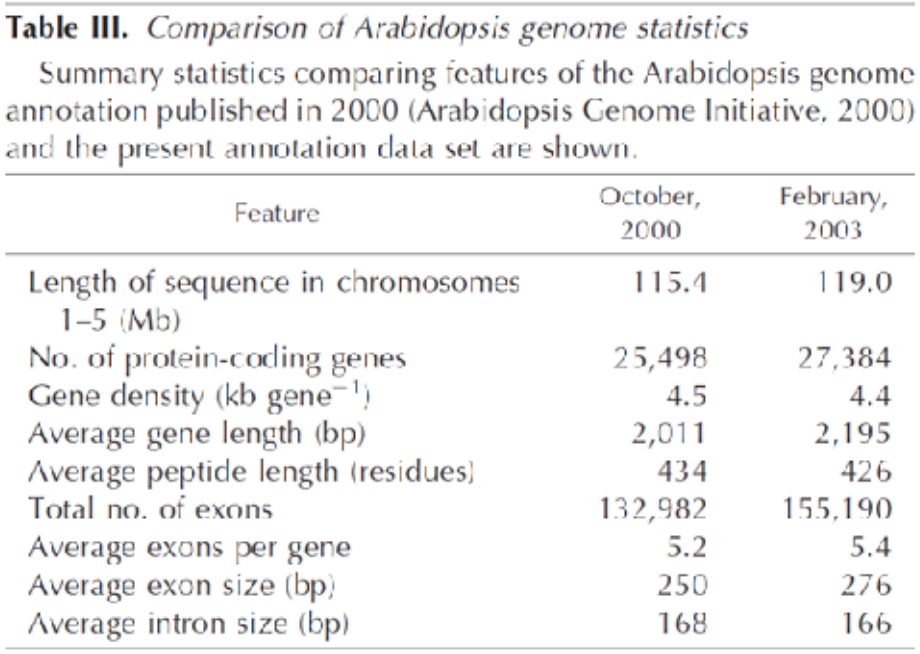


Table III. *Comparison of Arabidopsis genome statistics*

Summary statistics comparing features of the Arabidopsis genome annotation published in 2000 (Arabidopsis Genome Initiative, 2000) and the present annotation data set are shown.

Feature	October, 2000	February, 2003
Length of sequence in chromosomes 1–5 (Mb)	115.4	119.0
No. of protein-coding genes	25,498	27,384
Gene density (kb gene^{-1})	4.5	4.4
Average gene length (bp)	2,011	2,195
Average peptide length (residues)	434	426
Total no. of exons	132,982	155,190
Average exons per gene	5.2	5.4
Average exon size (bp)	250	276
Average intron size (bp)	168	166

B) As discussed in class, *Agrobacterium tumefaciens* can infect plants, integrating a stretch of so-called T-DNA at random into the host's genome. This has been used to generate "knockout" libraries in *Arabidopsis*: if the T-DNA lands within some gene, that gene can usually no longer be expressed in active form. Assuming that T-DNA insertion at any point in a gene sequence (including introns) leads to a functional knockout, how many different insertional mutants should we generate to obtain a knockout library covering 90% of all *Arabidopsis* genes? How about 99%?

We can use the usual formula for library coverage:

$$N = \frac{\ln(1 - p)}{\ln(1 - f)}$$

In this case, we need to think about what *f* stands for: instead of cloning out fragments of some average length from the genome, we are inserting at random into the genome, containing genes with some average length. Therefore, in this case, *f* will just be the ratio of the average gene length to the length of the genome, so the number of mutants we need is

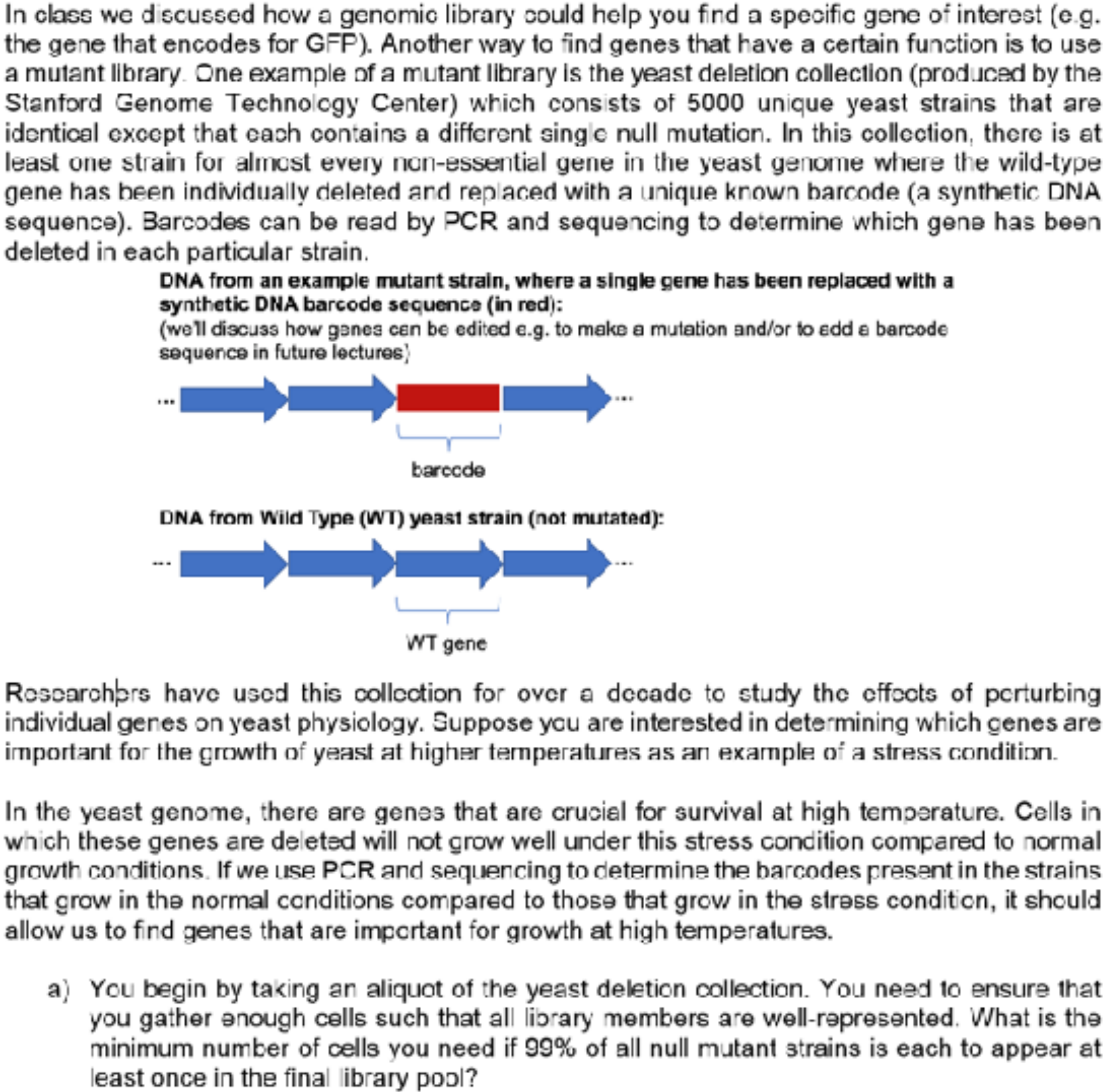
- For 90% coverage: N ln(1 – 0.9)/ln(1 – 2,195/119,000,000) **125,000**
- For 99% coverage: N = ln(1 0.99)/ln(1 2,195/119,000,000) = **250,000**

- C) As discussed in class, if we want to isolate a particular gene from a genomic library, we can calculate N, the number of clones we need to screen, using the following equation:

$$N = \frac{\ln(1-p)}{\ln(1-f)}$$

  where *p* is the probability of finding the gene in one of our clones and *f* is the ratio of the average insert length to the length of the genome. One tacit assumption here is that the insert length is significantly longer than the length of our target gene: if the inserts are too short, intuitively, the chance of finding a clone containing the full-length gene sequence goes down. If we do not make this assumption, and let *g* be the ratio of the target gene length to length of the genome, how does the above equation change?

Question 2:



In class we discussed how a genomic library could help you find a specific gene of interest (e.g. the gene that encodes for GFP). Another way to find genes that have a certain function is to use a mutant library. One example of a mutant library is the yeast deletion collection (produced by the Stanford Genome Technology Center) which consists of 5000 unique yeast strains that are identical except that each contains a different single null mutation. In this collection, there is at least one strain for almost every non-essential gene in the yeast genome where the wild-type gene has been individually deleted and replaced with a unique known barcode (a synthetic DNA sequence). Barcodes can be read by PCR and sequencing to determine which gene has been deleted in each particular strain.

DNA from an example mutant strain, where a single gene has been replaced with a synthetic DNA barcode sequence (in red):

(we'll discuss how genes can be edited e.g. to make a mutation and/or to add a barcode sequence in future lectures)

DNA from Wild Type (WT) yeast strain (not mutated):

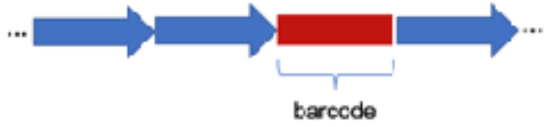
Researchers have used this collection for over a decade to study the effects of perturbing individual genes on yeast physiology. Suppose you are interested in determining which genes are important for the growth of yeast at higher temperatures as an example of a stress condition.

In the yeast genome, there are genes that are crucial for survival at high temperature. Cells in which these genes are deleted will not grow well under this stress condition compared to normal growth conditions. If we use PCR and sequencing to determine the barcodes present in the strains that grow in the normal conditions compared to those that grow in the stress condition, it should allow us to find genes that are important for growth at high temperatures.

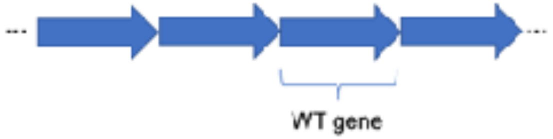
a) You begin by taking an aliquot of the yeast deletion collection. You need to ensure that you gather enough cells such that all library members are well-represented. What is the minimum number of cells you need if 99% of all null mutant strains is each to appear at least once in the final library pool?

In class we discussed how a genomic library could help you find a specific gene of interest (e.g. the gene that encodes for GFP). Another way to find genes that have a certain function is to use a mutant library. One example of a mutant library is the yeast deletion collection (produced by the Stanford Genome Technology Center) which consists of 5000 unique yeast strains that are identical except that each contains a different single null mutation. In this collection, there is at least one strain for almost every non-essential gene in the yeast genome where the wild-type gene has been individually deleted and replaced with a unique known barcode (a synthetic DNA sequence). Barcodes can be read by PCR and sequencing to determine which gene has been deleted in each particular strain.

DNA from an example mutant strain, where a single gene has been replaced with a synthetic DNA barcode sequence (in red):
(we'll discuss how genes can be edited e.g. to make a mutation and/or to add a barcode sequence in future lectures)



DNA from Wild Type (WT) yeast strain (not mutated):



Researchers have used this collection for over a decade to study the effects of perturbing individual genes on yeast physiology. Suppose you are interested in determining which genes are important for the growth of yeast at higher temperatures as an example of a stress condition.

In the yeast genome, there are genes that are crucial for survival at high temperature. Cells in which these genes are deleted will not grow well under this stress condition compared to normal growth conditions. If we use PCR and sequencing to determine the barcodes present in the strains that grow in the normal conditions compared to those that grow in the stress condition, it should allow us to find genes that are important for growth at high temperatures.

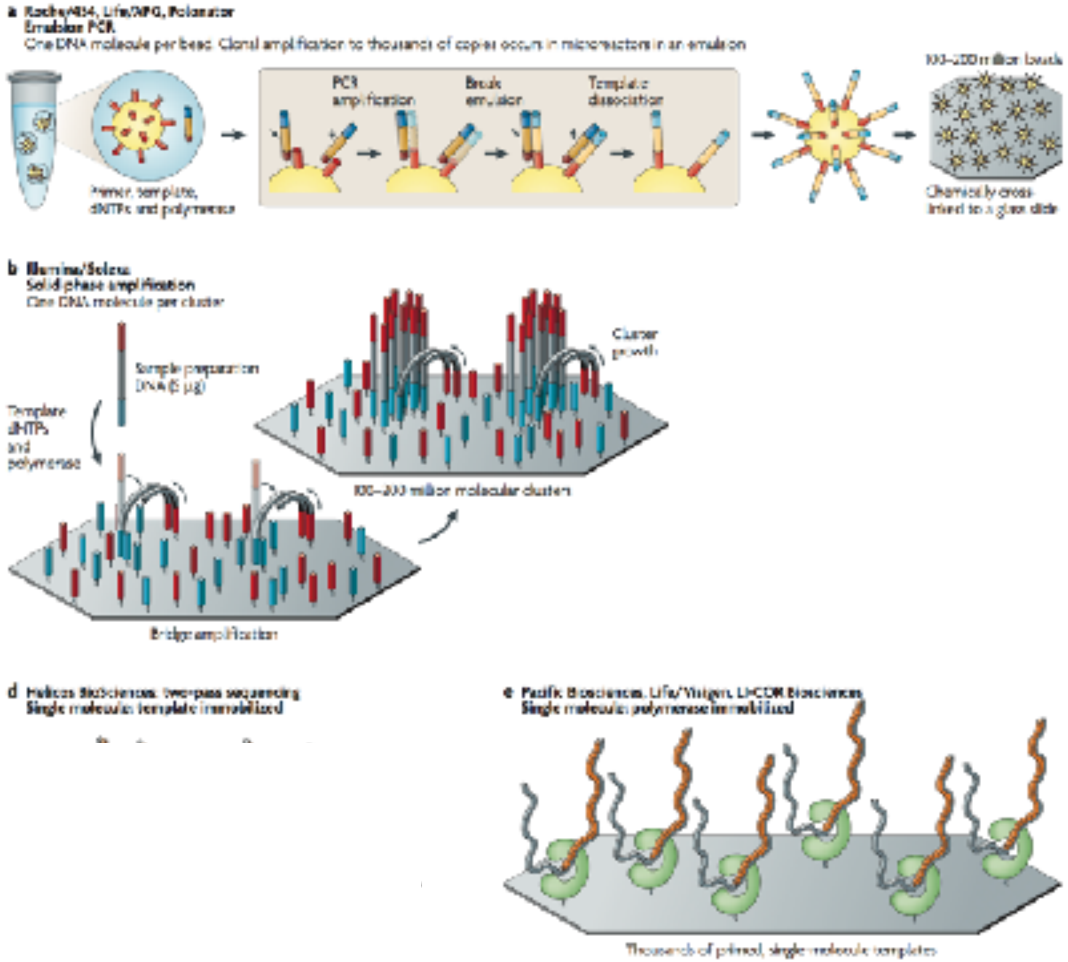
b) We are interested in quantifying the abundance of each library member (i.e. each barcode) in the cell population before and after exposure to the stress condition. To amplify the barcodes in each pool, we would like to begin with, on average, 1000 copies of each barcode to ensure that inefficiency in PCR and sequencing does not alter each member's representation. What concentration of total library DNA is needed to attain those 1000 copies? Assume these copies are contained in a volume of 50 microliters.

c) Over-amplifying a DNA library can cause depletion of primers, which leads to the formation of chimeric products and other undesirable effects. Practically, this means limiting the final DNA concentration achieved by PCR to less than 1 nM, which is enough for subsequent sequencing. How many cycles of PCR are necessary to amplify the initial DNA to this concentration?

d) You now sequence the pool of barcodes obtained from the cell population before and after exposure to high temperatures. How can you tell which genes are important for cell survival under these conditions? Are you able to say something about the relative importance of these genes?

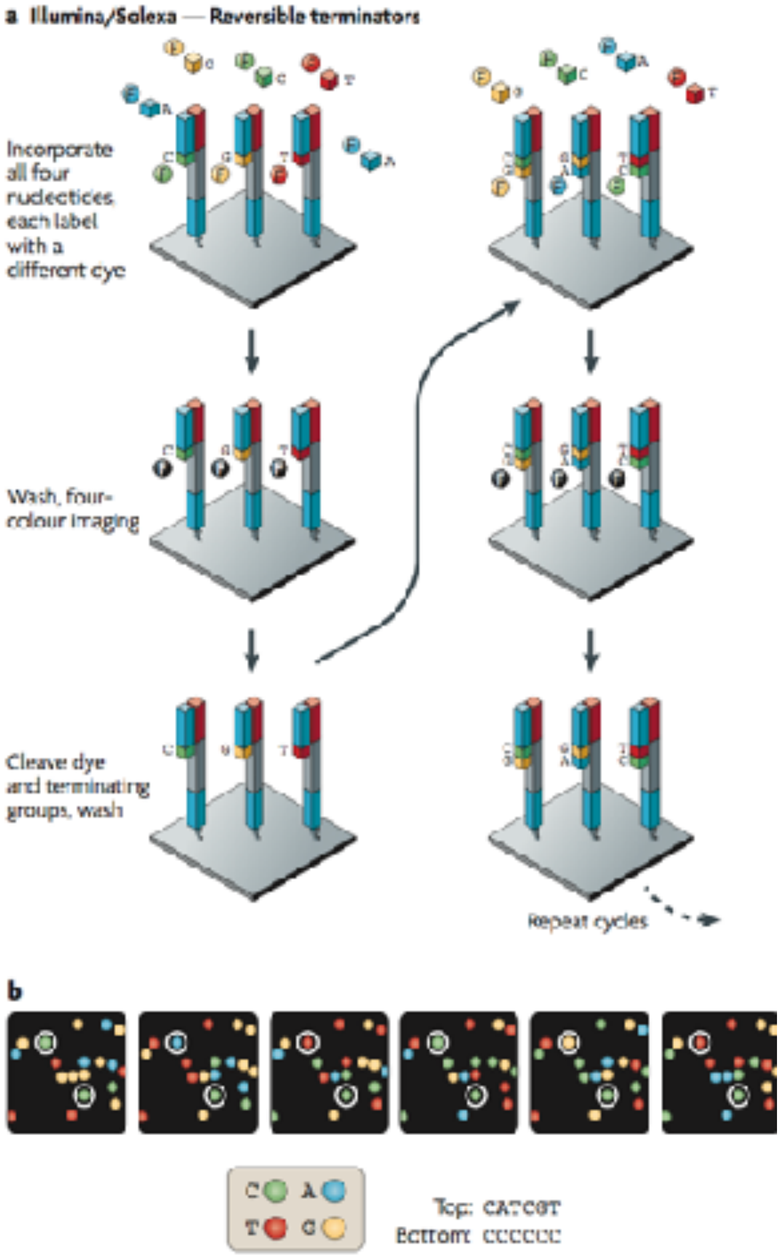
PART 2: DNA synthesis (Lecture 5) & sequencing (Lecture 6 & 7)

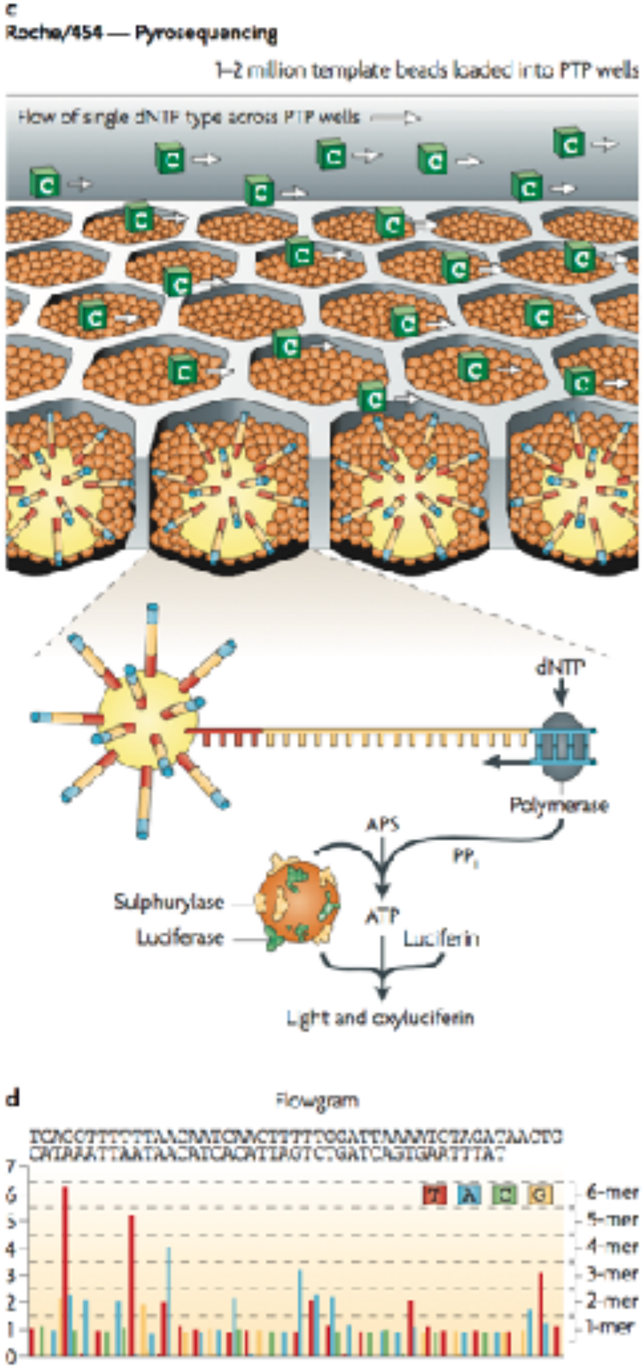
Review of DNA sequencing → template immobilization


a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a glass slide
b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Bridge amplification
100–200 million molecular clusters
Cluster growth
d Helicos BioSciences: two-pass sequencing
Single molecule: template immobilized
e Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized
Thousands of primed, single-molecule templates

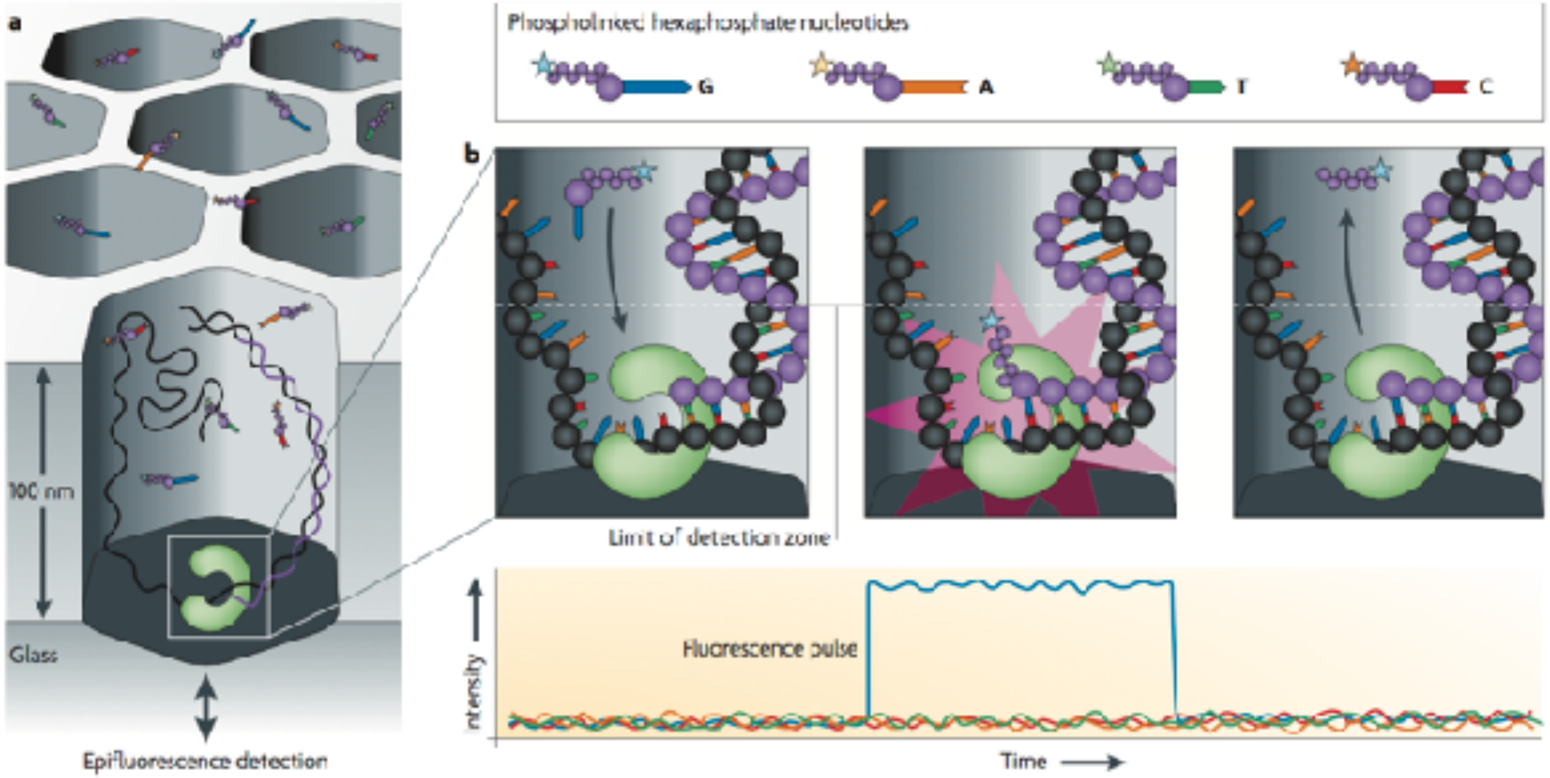
lization strategies. In emulsion PCR (emPCR) **(a)**, a reaction mixture consisting of ... mixed to encapsulate bead–DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. EmPCR beads can be chemically attached to a glass slide or deposited into PicoTiterPlate wells (FIG. 3c). Solid-phase amplification **(b)** is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Three approaches are shown for immobilizing single-molecule templates to a solid support: immobilization by a primer **(c)**; immobilization by a template **(d)**; and immobilization of a polymerase **(e)**. dNTP, 2'-deoxyribonucleoside triphosphate.

Review of DNA sequencing → detection of bases incorporated


a Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
b
C A
T G
Top: CATCGT
Bottom: CCCCCC


c Roche/454 — Pyrosequencing
1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
C
dNTP
Polymerase
APS
PP$_i$
Sulphurylase
Luciferase
ATP
Luciferin
Light and oxyluciferin
d
Flowgram
TCAGGTTTTTTAACAATCAACTTTTTGGATTAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT
0
1
2
3
4
5
6
7
T A C G
6-mer
5-mer
4-mer
3-mer
2-mer
1-mer

Review of DNA sequencing → detection of bases incorporated



Pacific Biosciences — Real-time sequencing

Example #3: Genome sequencing

(A) Why do you need to amplify the DNA template in order for Illumina and 454 sequencing to work? **(2 pts)**

(B) Given that multiple molecules of the same DNA template are being detected at the same time in Illumina and 454 sequencing, what are two strategies used to synchronize the polymerases? **(4 pts)**

(C) The need to synchronize the polymerase creates two problems for Illumina that are solved by newer sequencing technologies that read single molecules of DNA. First, the Illumina read length is limited to ~150 bp, whereas single molecule sequencers can generate much longer reads. Please speculate on the reason why the read length is limited for Illumina. **(2 pts)**

(D) In class, PacBio sequencing was presented as an example of Next-Generation Sequencing that offers impressive throughput thanks to DNA polymerase's rapid elongation rate. This form of sequencing is based on the SMRT method (Eid *et al.*, Science, 2009), where a zero-mode waveguide (100 nm in height, 70 nm in diameter) enables fluorophore detection only at the very bottom <30 nanometers of the well. For this application, the Φ29 DNA polymerase from *Bacillus subtilis* phage was chosen for its robust properties, which include an elongation rate of 200 bases per second. Please answer the following regarding DNA polymerases used for PacBio sequencing:

(i) How many DNA polymerases are required at the bottom of the well for SMRT sequencing to work properly? **(1 pt)**

(ii) If a solution of polymerase is dispersed across the many ZMW wells of a sequencing chip, what concentration of polymerase (in mg/ml, given a molar mass of 66,714 g/mol) will allow for one polymerase per well, on average? **(4 pts)**

(iii) The characteristic diffusion length of a small molecule is $l \sim \sqrt{4\mathcal{D}t}$, where $\mathcal{D}$ ~ 10^{-5} cm^2/s is roughly its diffusivity in aqueous solution. Given the polymerase elongation rate above, demonstrate using the relevant time scales why detection is physically possible. Assume diffusion only in the vertical dimension. **(5 pts)**

(A) In class, PacBio sequencing was presented as an example of Next-Generation Sequencing that offers impressive throughput thanks to DNA polymerase's rapid elongation rate. This form of sequencing is based on the SMRT method (Eid *et al.*, Science, 2009), where a zero-mode waveguide (100 nm in height, 70 nm in diameter) enables fluorophore detection only at the very bottom <30 nanometers of the well. For this application, the Φ29 DNA polymerase from *Bacillus subtilis* phage was chosen for its robust properties, which include an elongation rate of 200 bases per second. Please answer the following regarding DNA polymerases used for PacBio sequencing:

i. How many DNA polymerases are required at the bottom of the well for SMRT sequencing to work properly? **(1 pt)**

1 polymerase per well, otherwise conflicting readout

ii. If a solution of polymerase is dispersed across the many ZMW wells of a sequencing chip, what concentration of polymerase (in mg/ml, given a molar mass of 66,714 g/mol) will allow for one polymerase per well, on average? **(4 pts)**

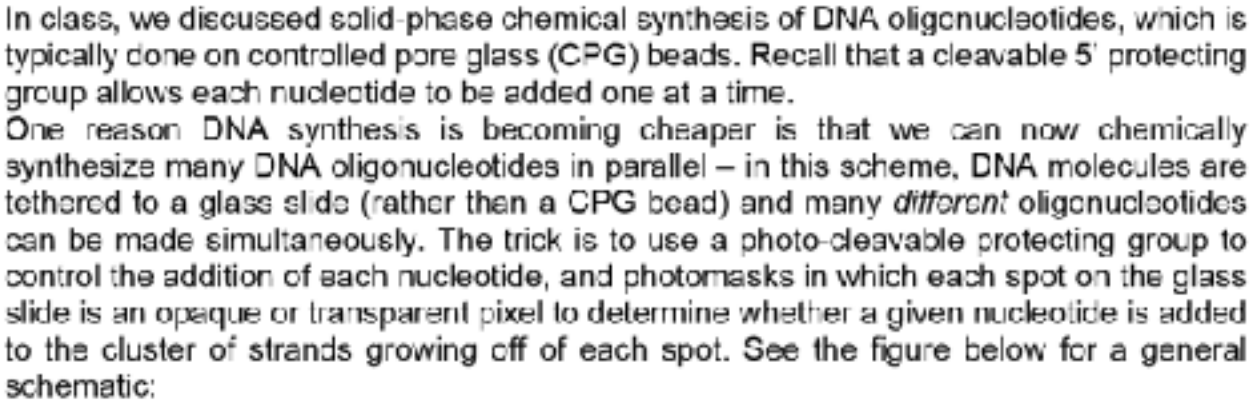
(2 pts) Volume of each well = $\frac{1}{4}\pi d^2 h$ = 0.25 x 3.1415 x (70 nm)^2 x 100 nm = 3.85e5 nm^3 = 3.85e-16 ml
(1 pt) 1 polymerase = 66714 g/mol * 1000 mg/g * 1/6.022e23 molecules/mol = 1.108e-16 mg
(1 pt) 1 polymerase per well = 1.108e-16/3.85e-16 = **0.288 mg/ml**

iii. The characteristic diffusion length of a small molecule is $l \sim \sqrt{4Dt}$, where $D \sim 10^{-5}$ cm^2/s is roughly its diffusivity in aqueous solution. Given the polymerase elongation rate above, demonstrate using the relevant time scales why detection is physically possible. Assume diffusion only in the vertical dimension. **(5 pts)**

(1 pt) Characteristic length = detection height of ZMW ~ 30 nanometers
(1 pt) $t = l^2/4D$ = (30e-6 cm)^2 / (4*1e-5 cm^2/s) = **225 nanoseconds** spent within detection volume
(note the problem statement has incorrect units for diffusivity, no points deducted for discrepancies based on this error)

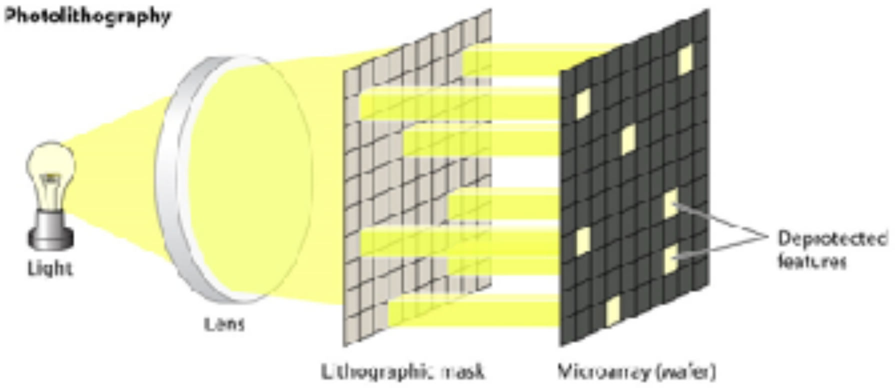
(2 pts) 200 bases per second = **5 milliseconds per base**
(1 pt) ∴ **time spent at polymerase >>> time spent within detection volume**

Example #4: DNA synthesis

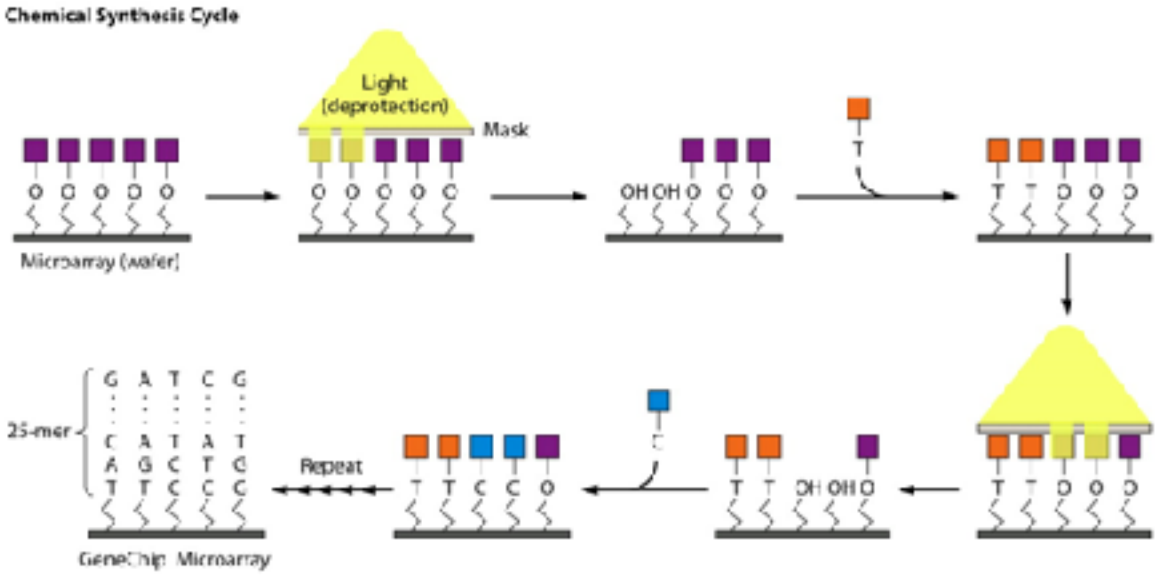


In class, we discussed solid-phase chemical synthesis of DNA oligonucleotides, which is typically done on controlled pore glass (CPG) beads. Recall that a cleavable 5' protecting group allows each nucleotide to be added one at a time.

One reason DNA synthesis is becoming cheaper is that we can now chemically synthesize many DNA oligonucleotides in parallel – in this scheme, DNA molecules are tethered to a glass slide (rather than a CPG bead) and many *different* oligonucleotides can be made simultaneously. The trick is to use a photo-cleavable protecting group to control the addition of each nucleotide, and photomasks in which each spot on the glass slide is an opaque or transparent pixel to determine whether a given nucleotide is added to the cluster of strands growing off of each spot. See the figure below for a general schematic:



Photolithography



Chemical Synthesis Cycle

(A) To simplify matters, let’s imagine we need to chemically synthesize four unique oligos: CGTA, ACGT, TACG, and GTAC
Using the CPG beads method (so each oligo synthesized separately), how many rounds of chemical synthesis, each adding a single protected base, need to be done to make these 4 oligos? **(2 pts)**

(B) If we instead chemically synthesize these 4 oligos in parallel (using the glass slide method shown above), how many rounds of chemical synthesis, each one adding a single protected base, need to be done? **(2 pts)**

(C) The error rate for synthesis on a glass slide is higher than normal. In order to fish out a correctly synthesized product, unique oligonucleotide tags known as “bar codes” can be added to the end of each DNA molecule to generate a library for next-generation sequencing. In this technique, termed ‘dial-out PCR’, PCR primers are then used to amplify the correct sequence that is identified in their sequencing data. It has been shown that a correct sequence can be ‘dialed’ out even if it is as rare as 1/100,000 in the synthesized mixture. After 40 rounds of PCR, how much of an enrichment factor has been achieved? **(5 pts)**