15:30 - 16:15 – Sean Prosser

New Developments for Natural History Collection Barcoding

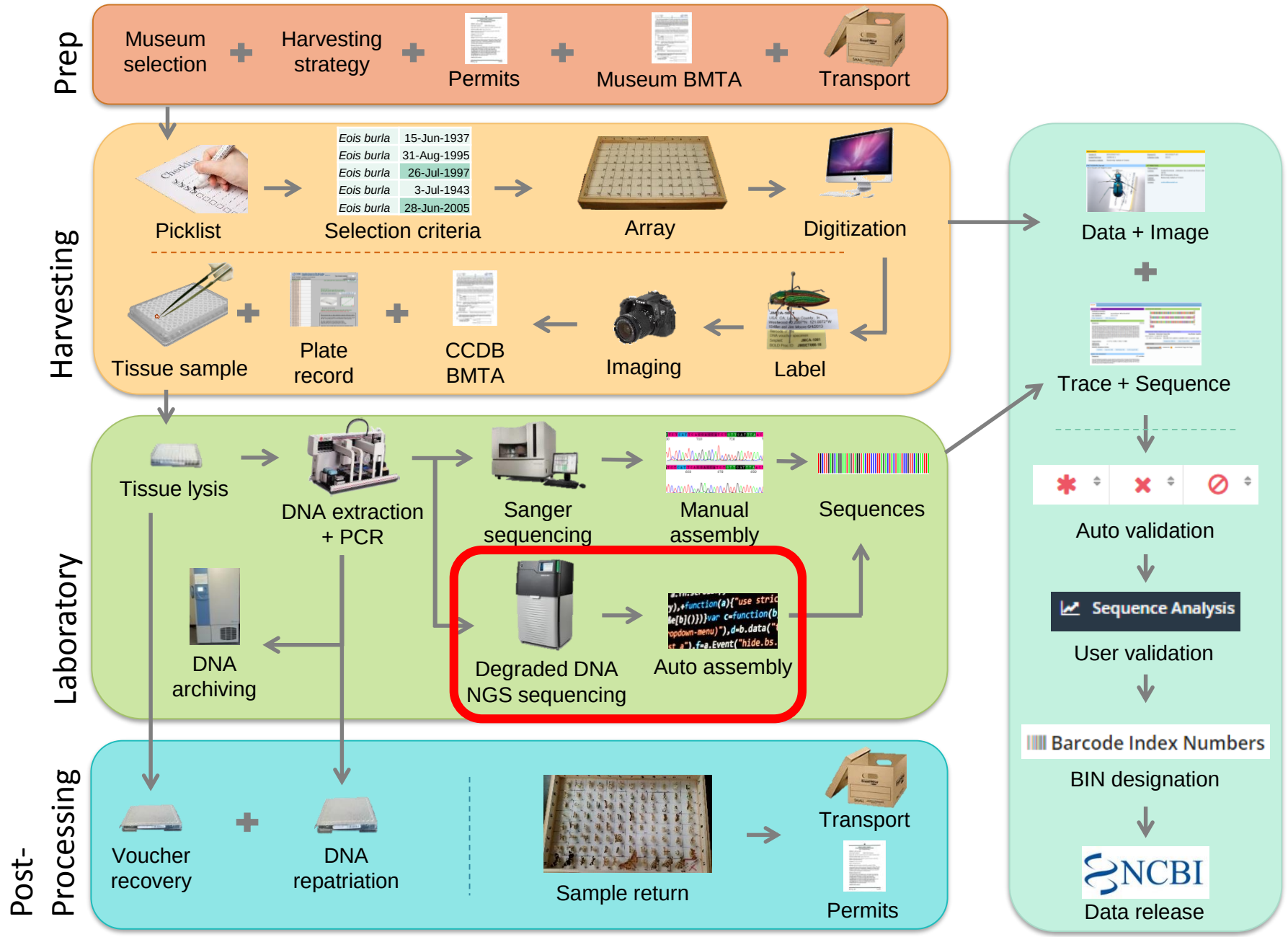
Centre for Biodiversity Genomics

DNA Barcoding Natural History Collections

Recap

Barcoding Museum Specimens

Age	Target Amplicons	Final Sequence Length	Method	No. Reactions (PCR/SEQ)
Fresh – 15 yrs	658 bp	658 bp	**Sanger**	1/1
15 - 60 yrs	307 bp, 407 bp	658 bp	**Sanger**	2/4
60-240+ yrs	15 amplicons ranging from 119 - 366 bp	658 bp	**Sanger or NGS**	2/1

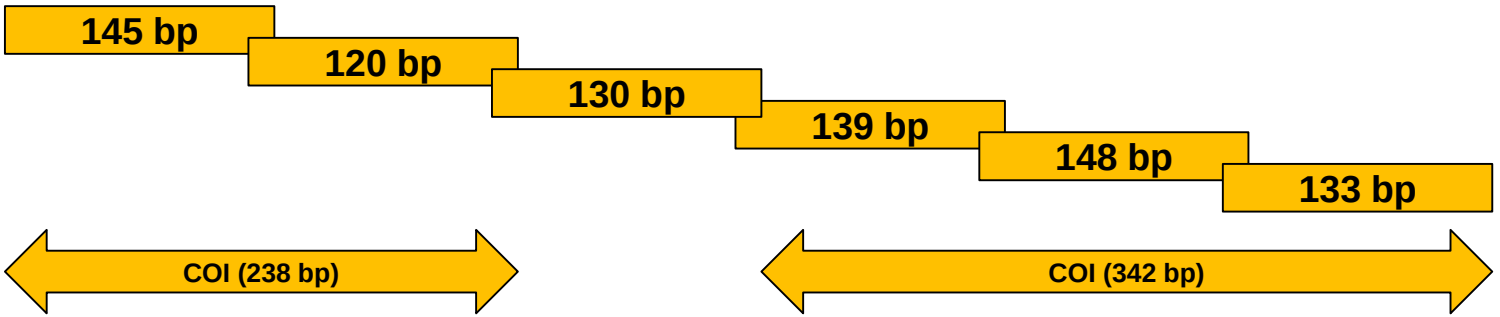

Prep
Museum selection
Harvesting strategy
Permits
Museum BMTA
Transport
Harvesting
Picklist
Eois burla 15-Jun-1937
Eois burla 31-Aug-1995
Eois burla 26-Jul-1997
Eois burla 3-Jul-1943
Eois burla 28-Jun-2005
Selection criteria
Array
Digitization
Tissue sample
Plate record
CCDB BMTA
Imaging
Label
Laboratory
Tissue lysis
DNA extraction + PCR
Sanger sequencing
Manual assembly
Sequences
DNA archiving
Degraded DNA NGS sequencing
Auto assembly
Post-Processing
Voucher recovery
DNA repatriation
Sample return
Transport
Permits
BOLD Upload
Data + Image
Trace + Sequence
Auto validation
Sequence Analysis
User validation
Barcode Index Numbers
BIN designation
NCBI
Data release
BOLD Workbench

New Developments

- Primers
- PCR Protocols
- MID-Tagging
- NGS Data Assembly and Analysis
- NGS Platforms

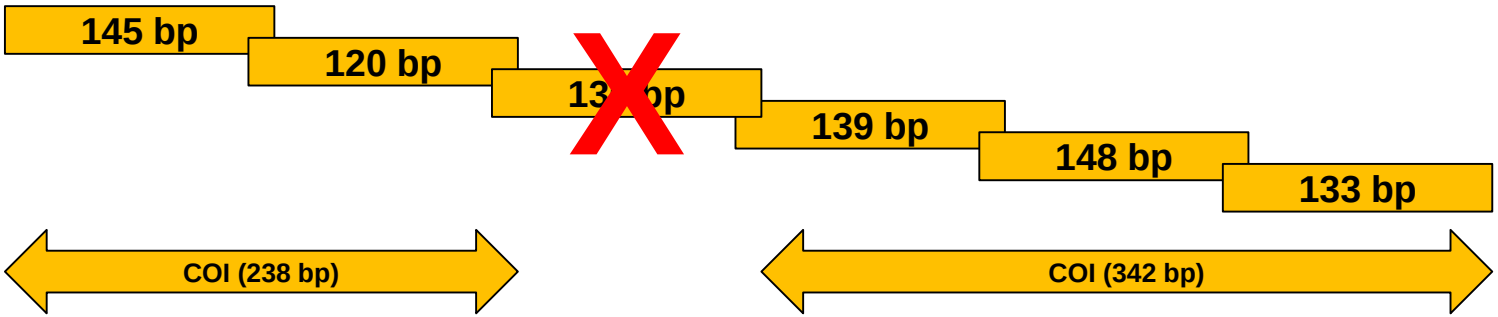
Primers

Barcoding Museum Specimens



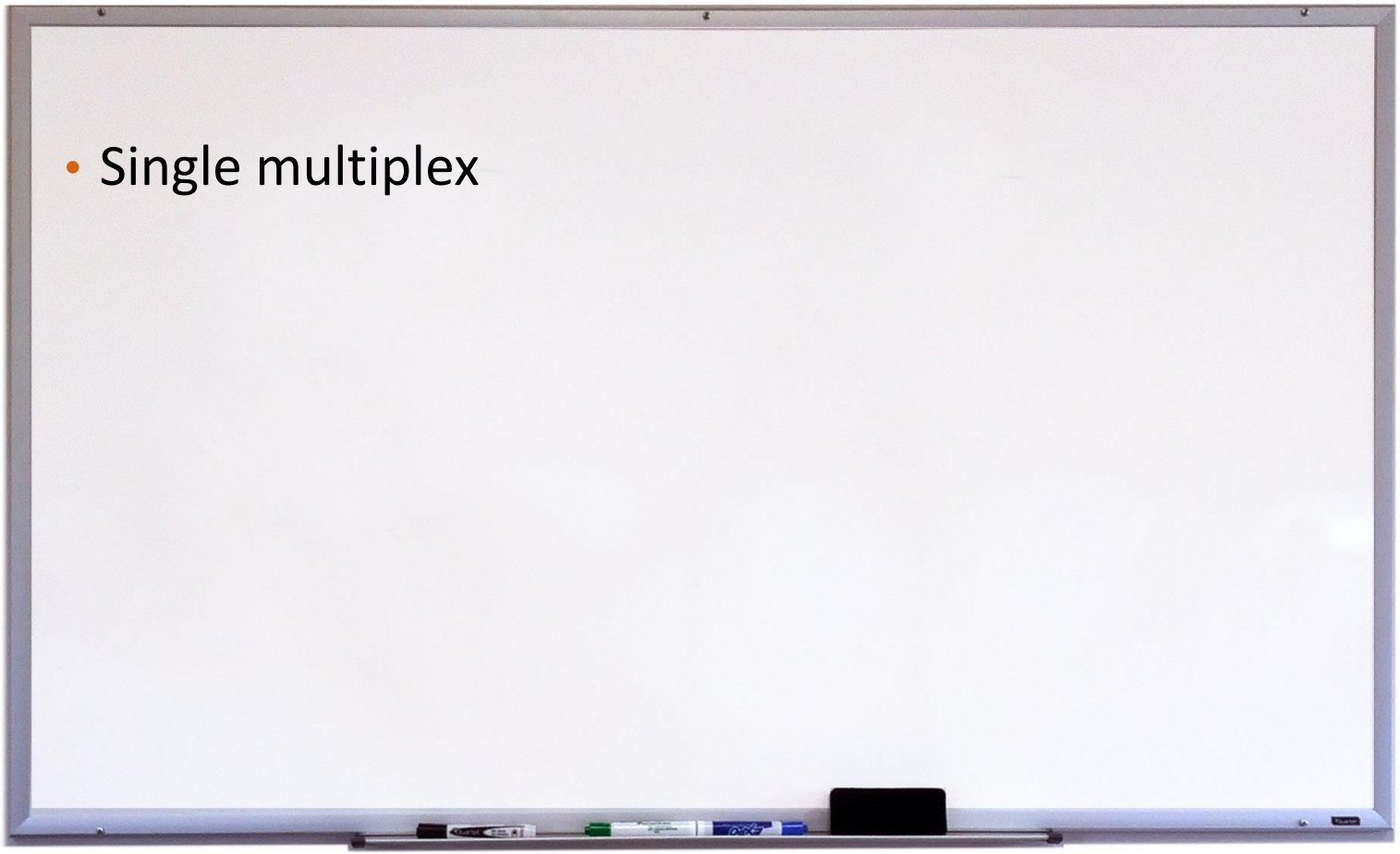
Primers

Barcoding Museum Specimens



- **Older specimens → degraded DNA**

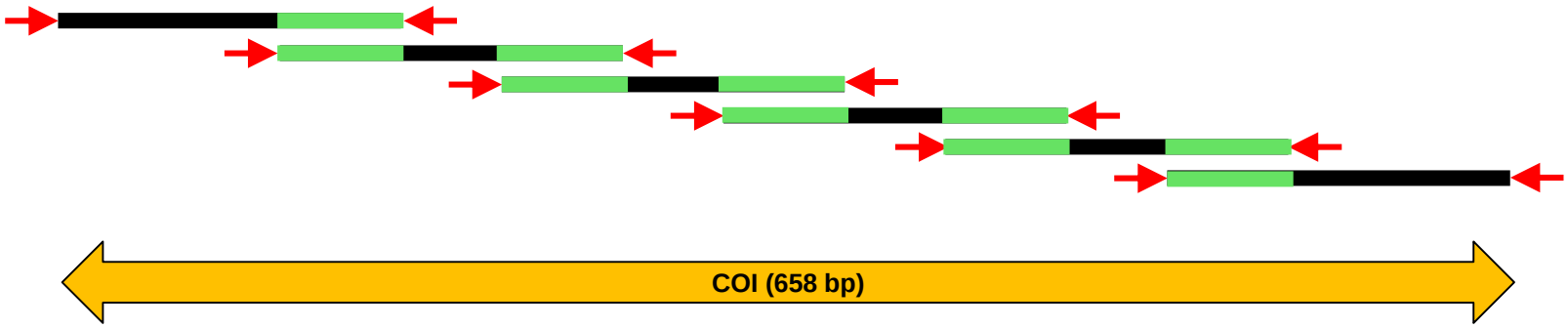
PCR Protocols



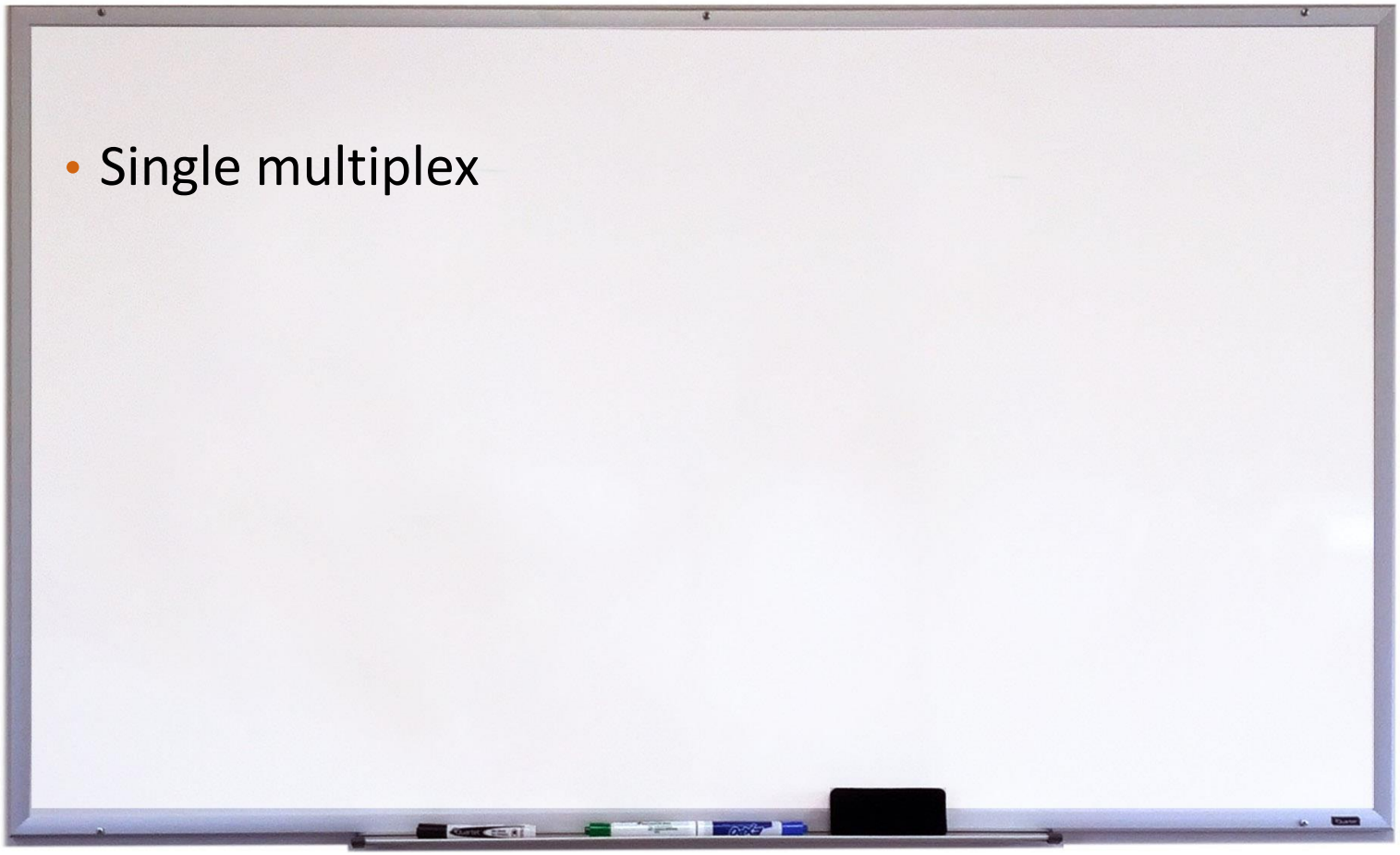
PCR Protocols

Single Multiplex

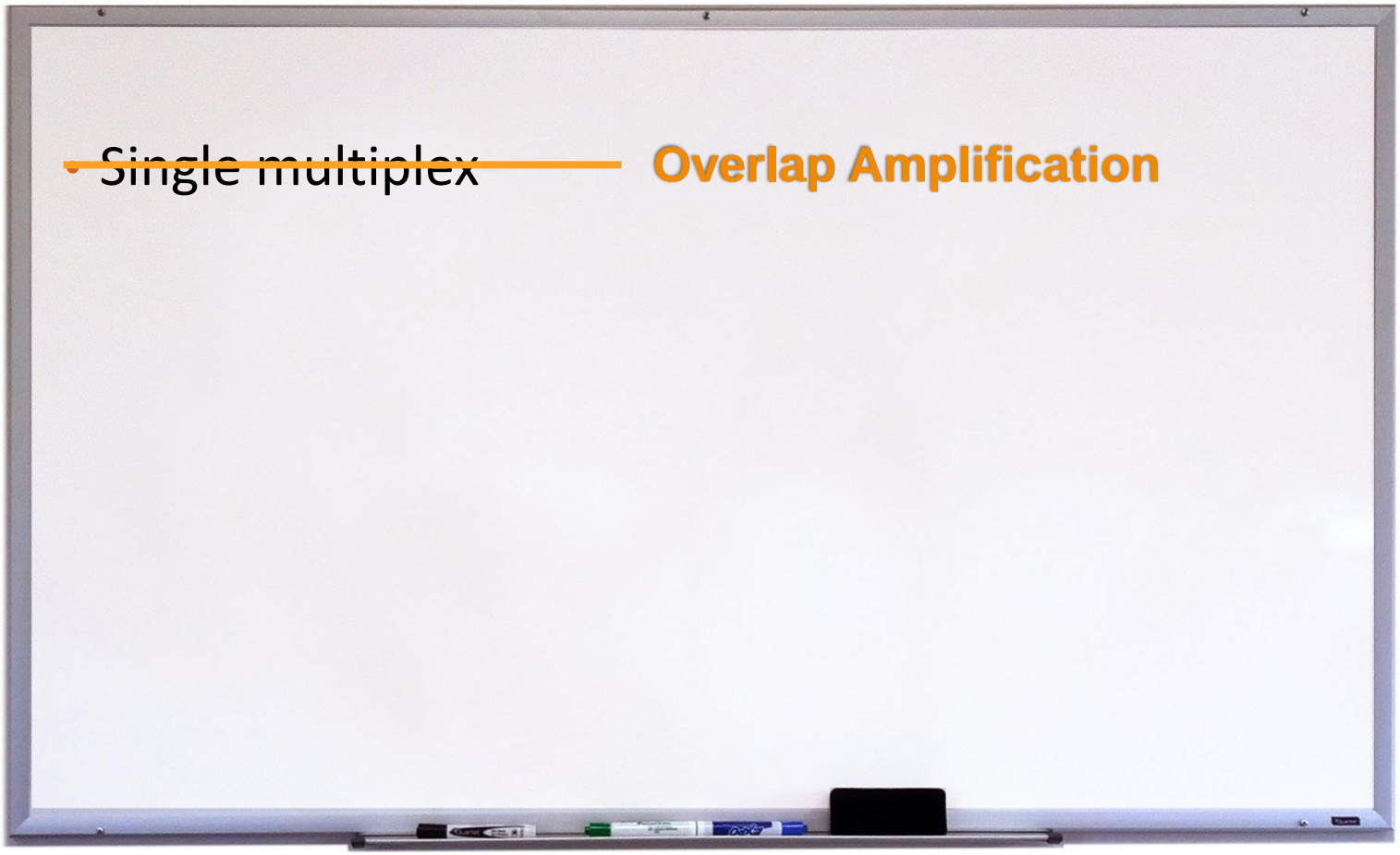
- Preferential amplification/sequencing of overlap regions



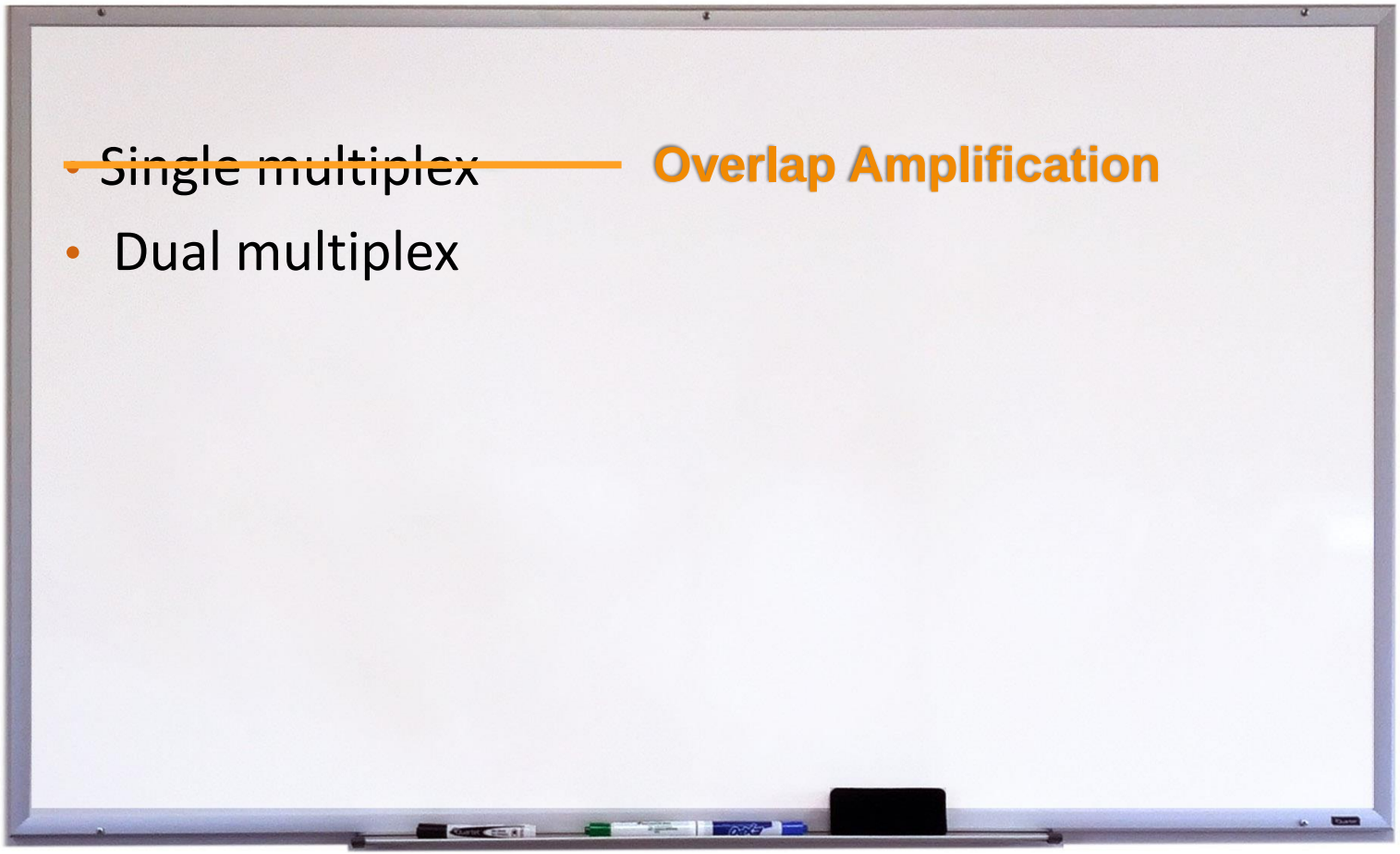
PCR Protocols



PCR Protocols



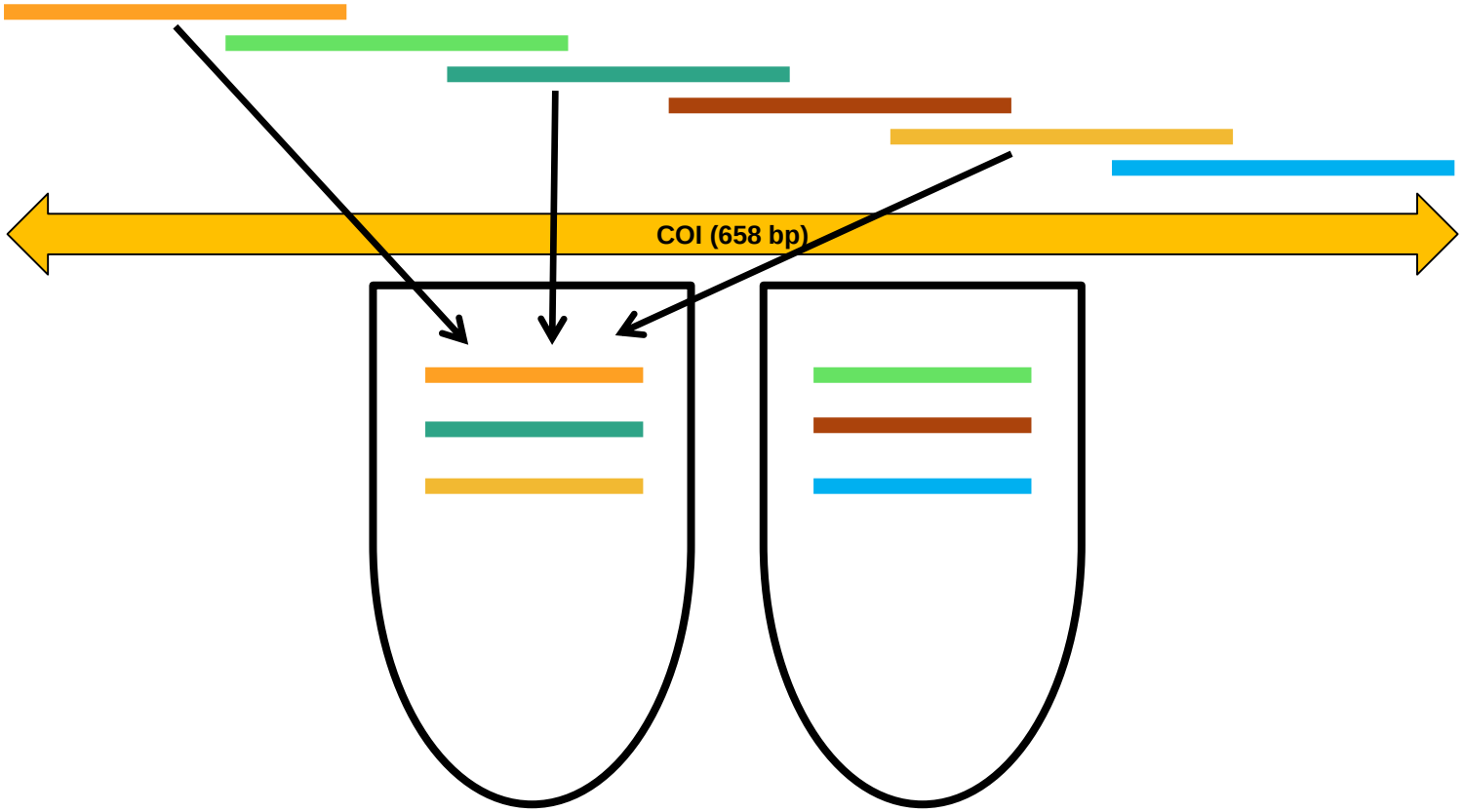
PCR Protocols



PCR Protocols

Dual Multiplex

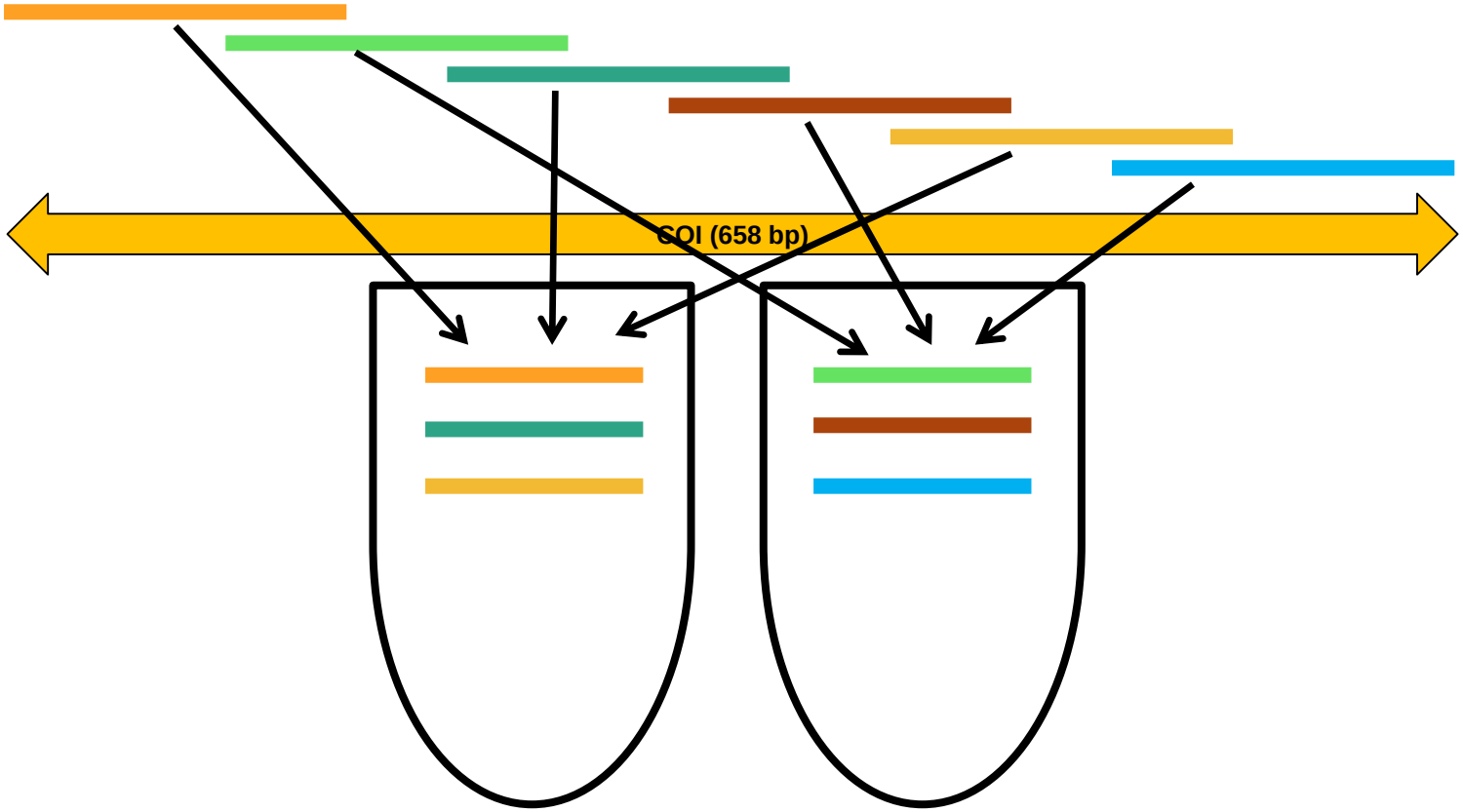
- Amplify several different DNA fragments simultaneously



PCR Protocols

Dual Multiplex

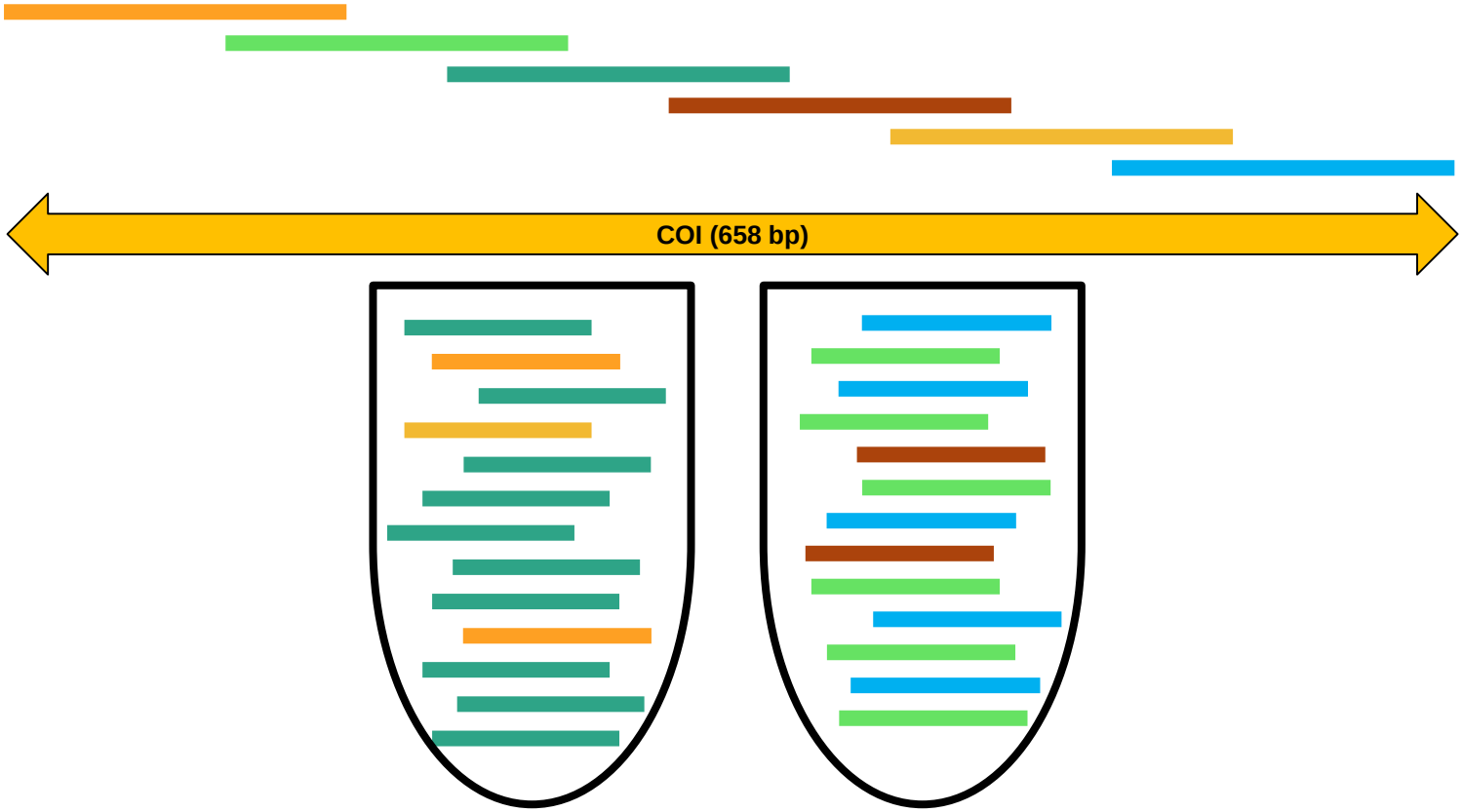
- Amplify several different DNA fragments simultaneously



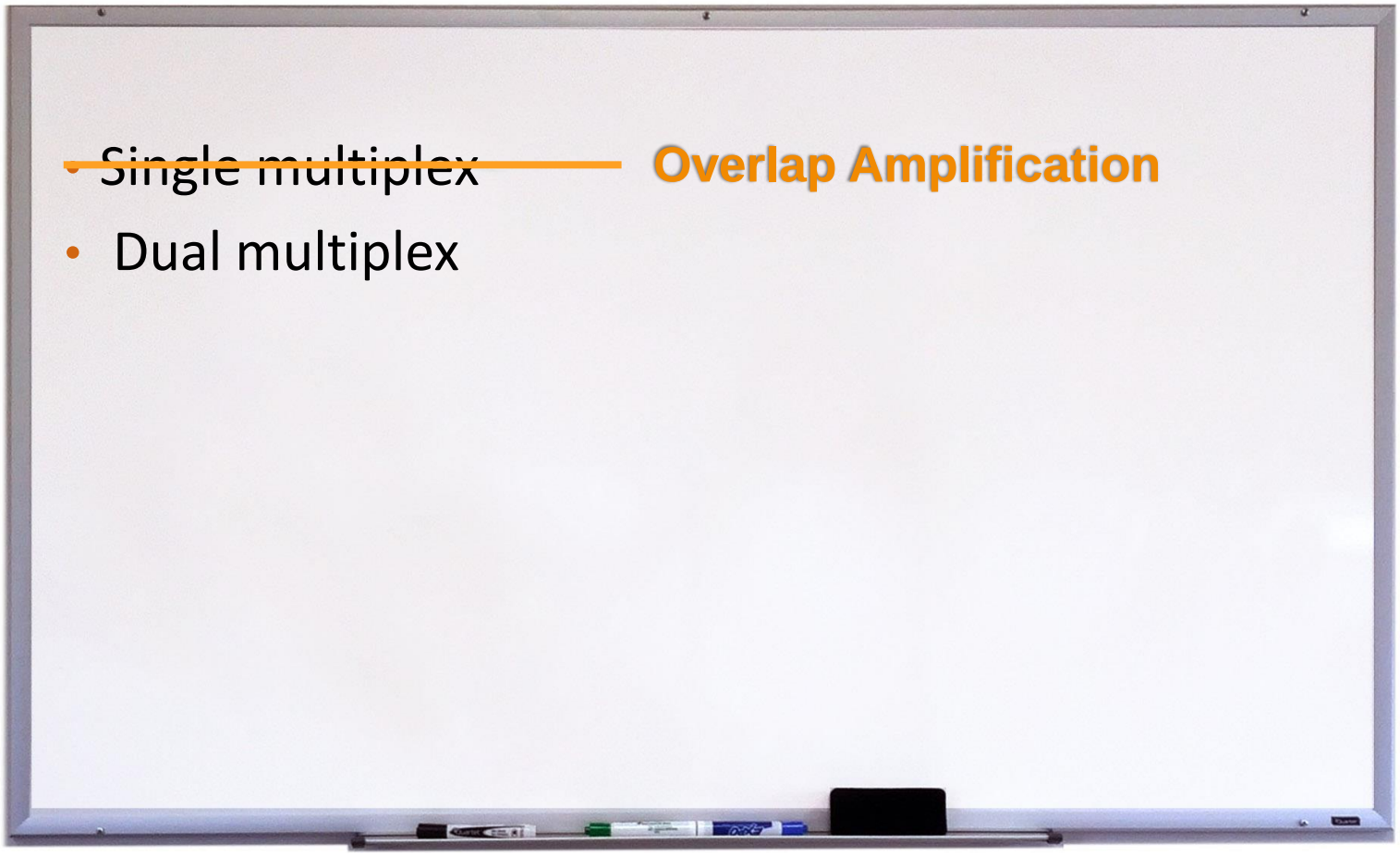
PCR Protocols

Dual Multiplex

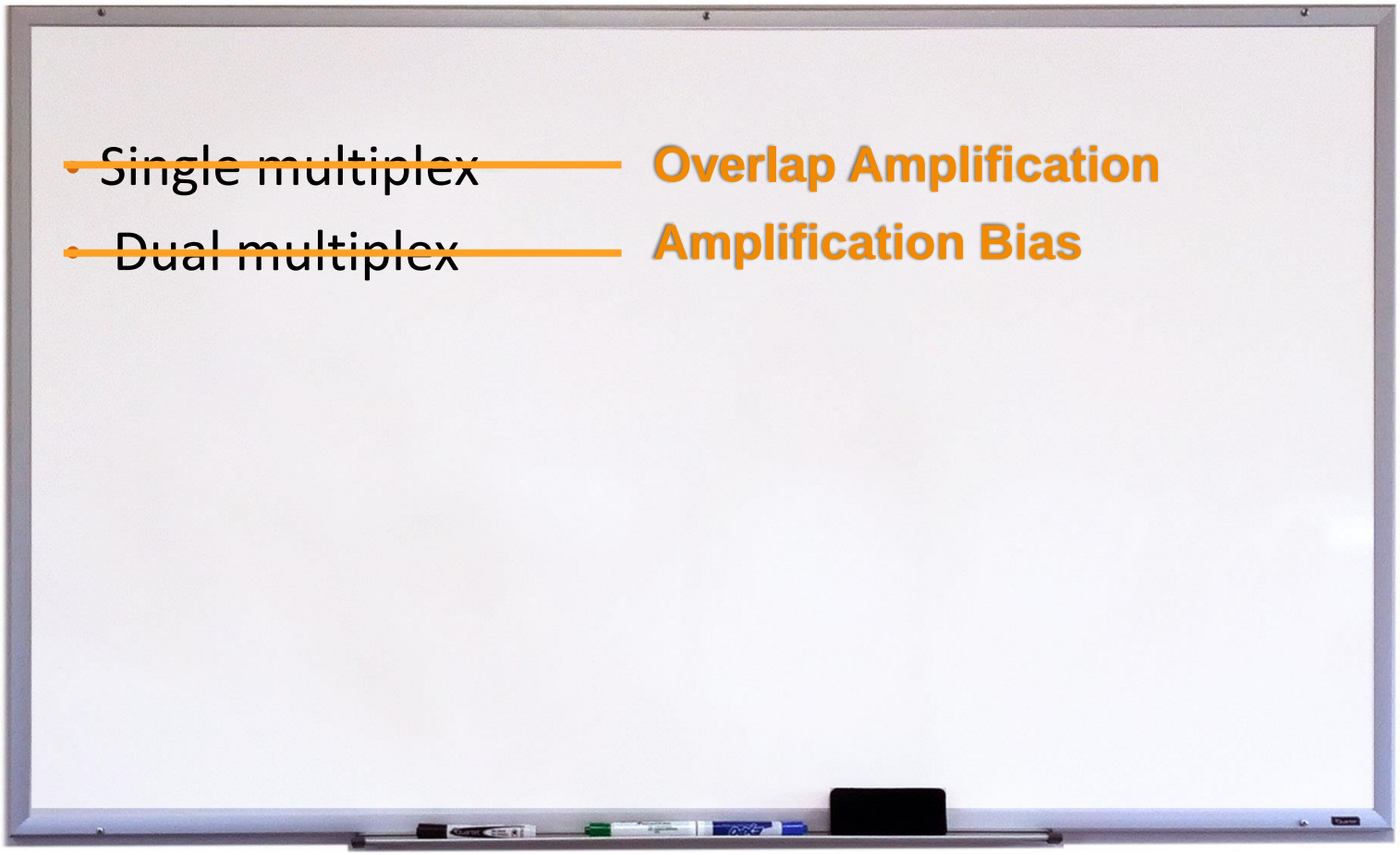
- Amplification bias



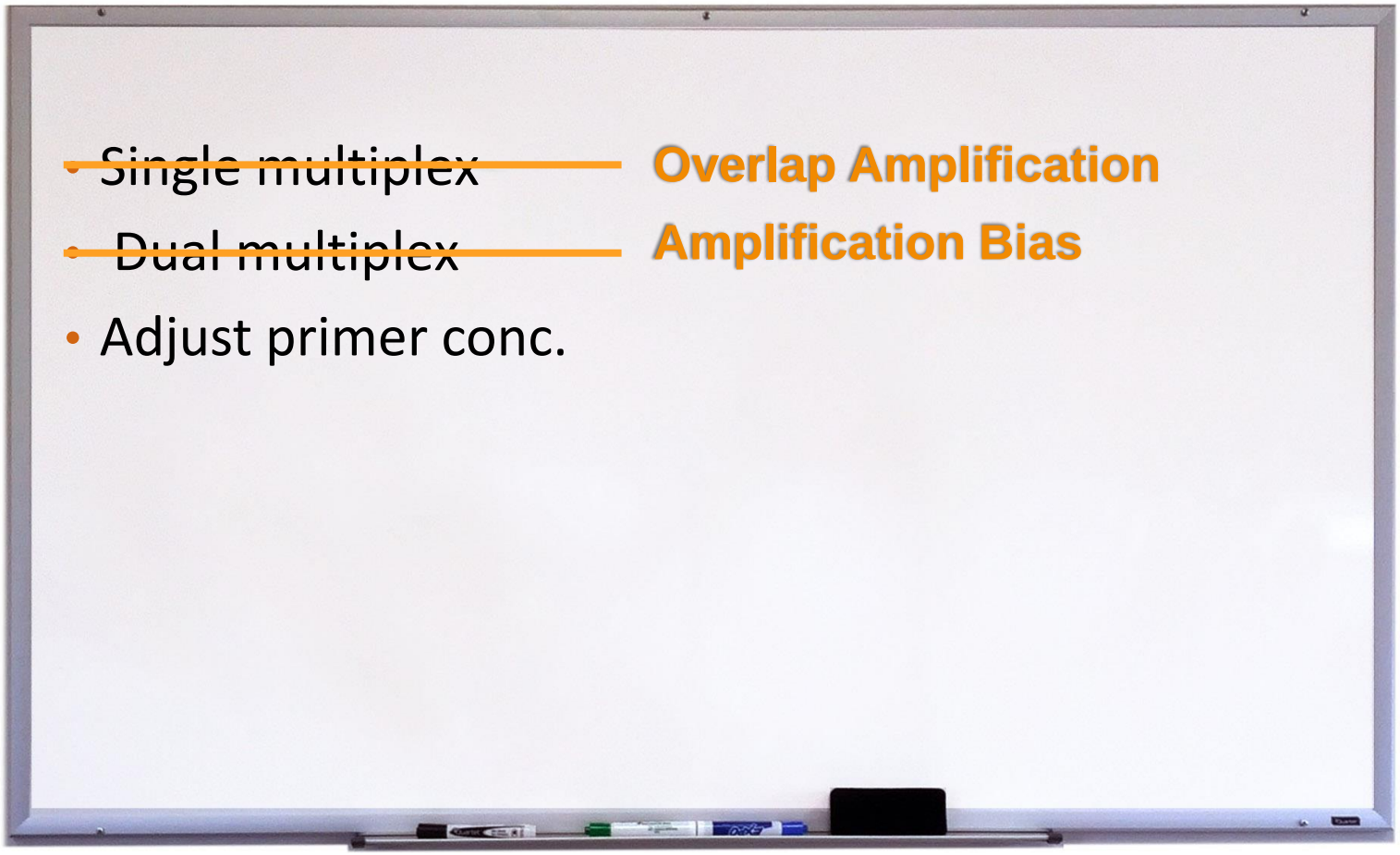
PCR Protocols



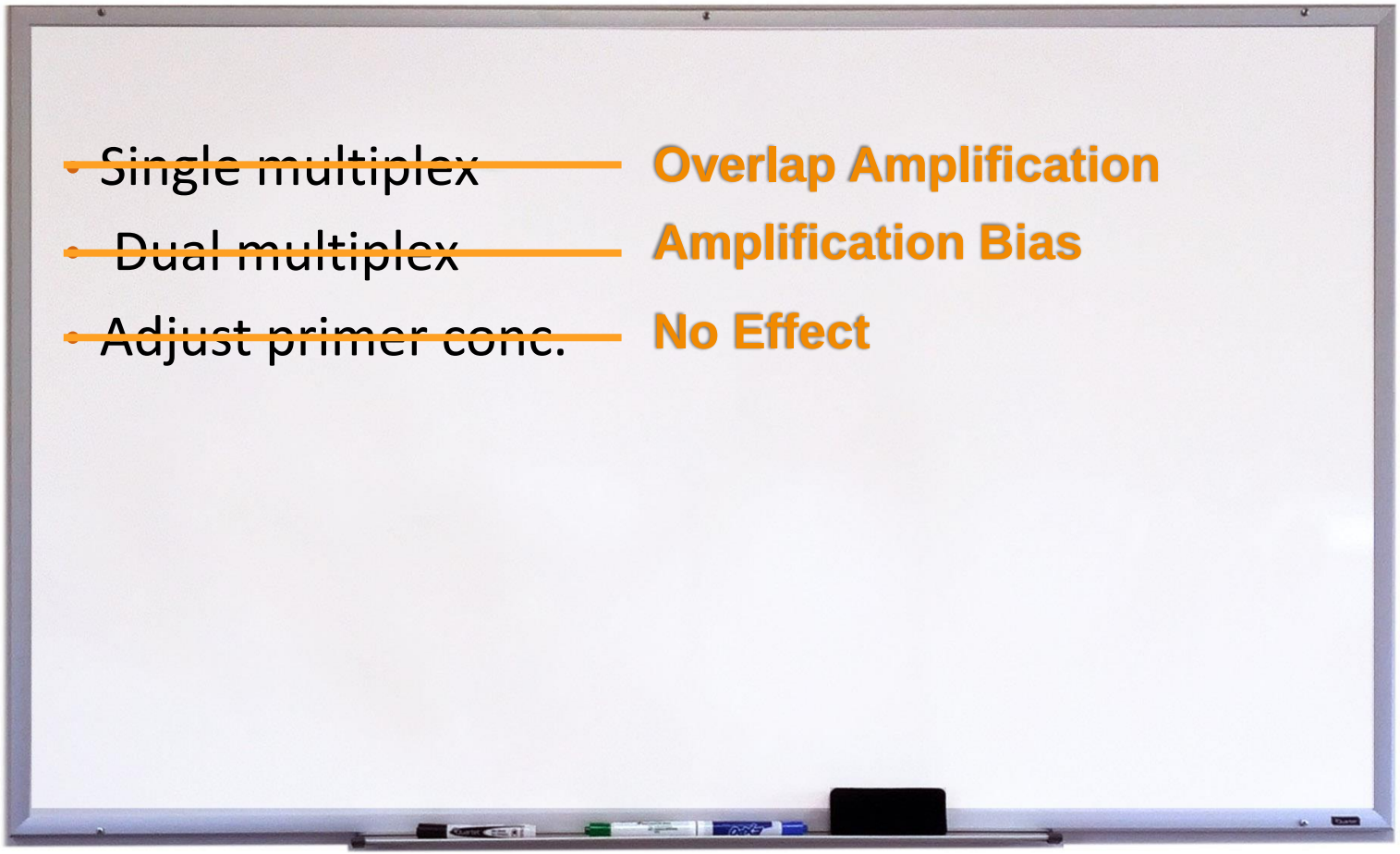
PCR Protocols



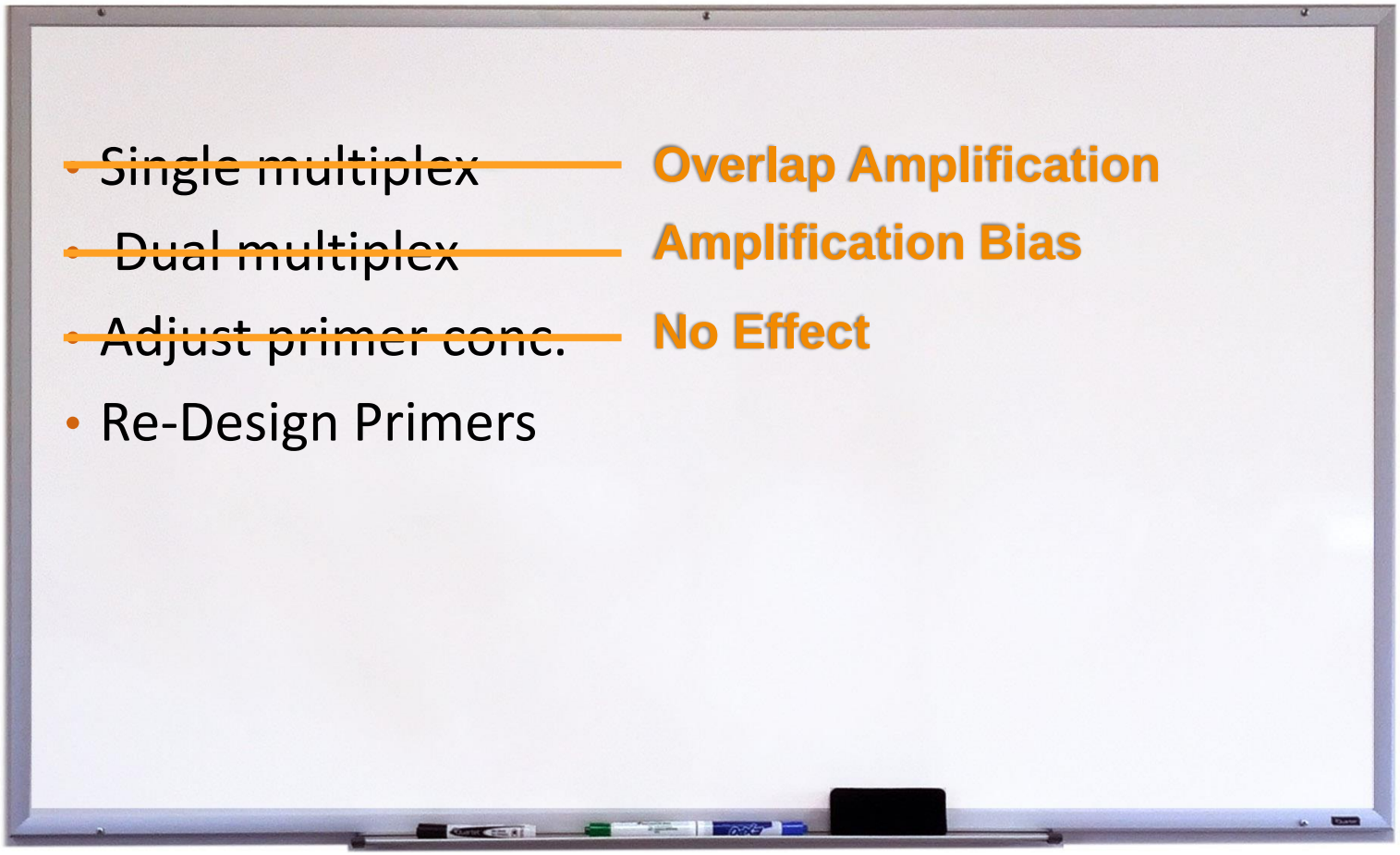
PCR Protocols



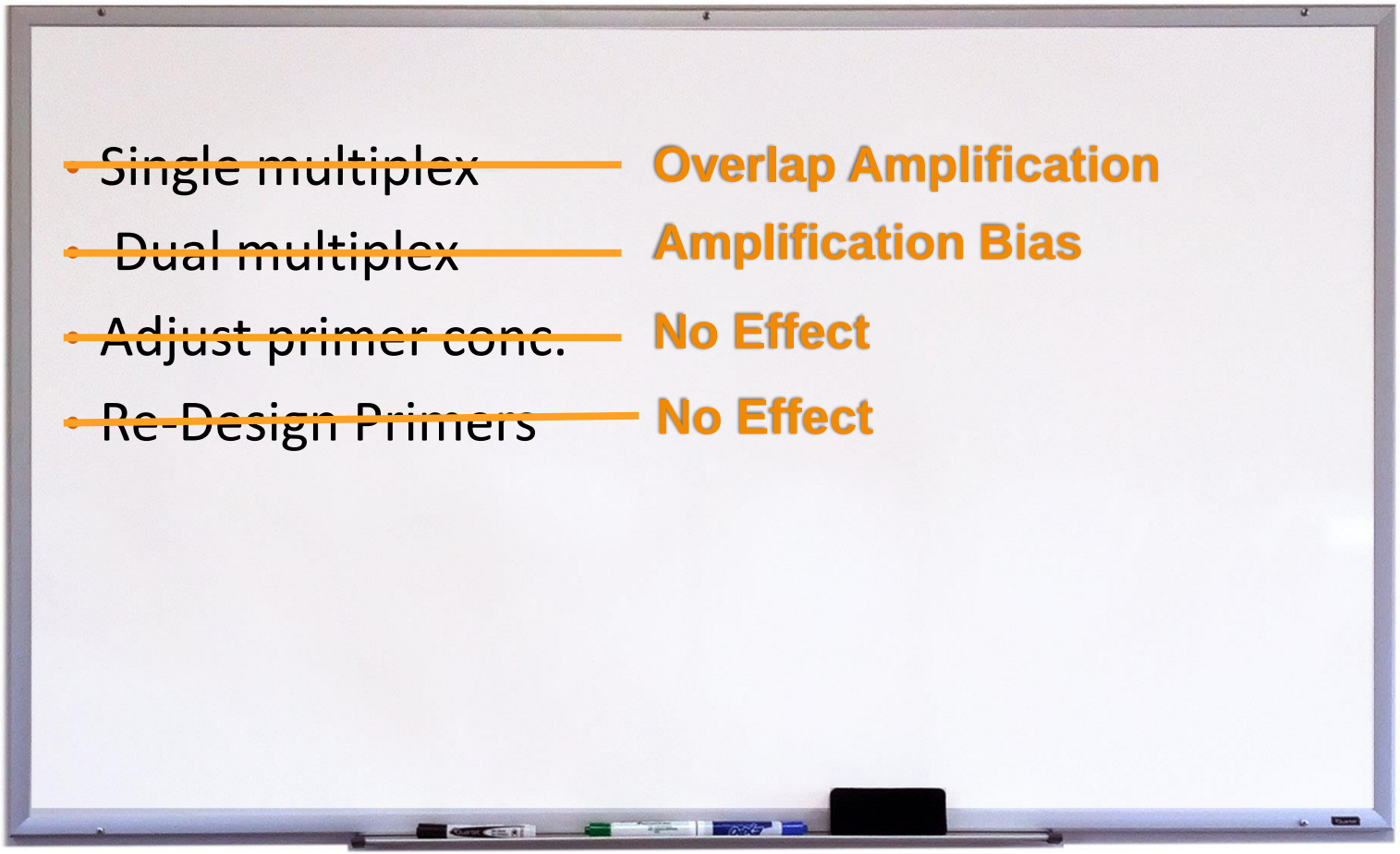
PCR Protocols



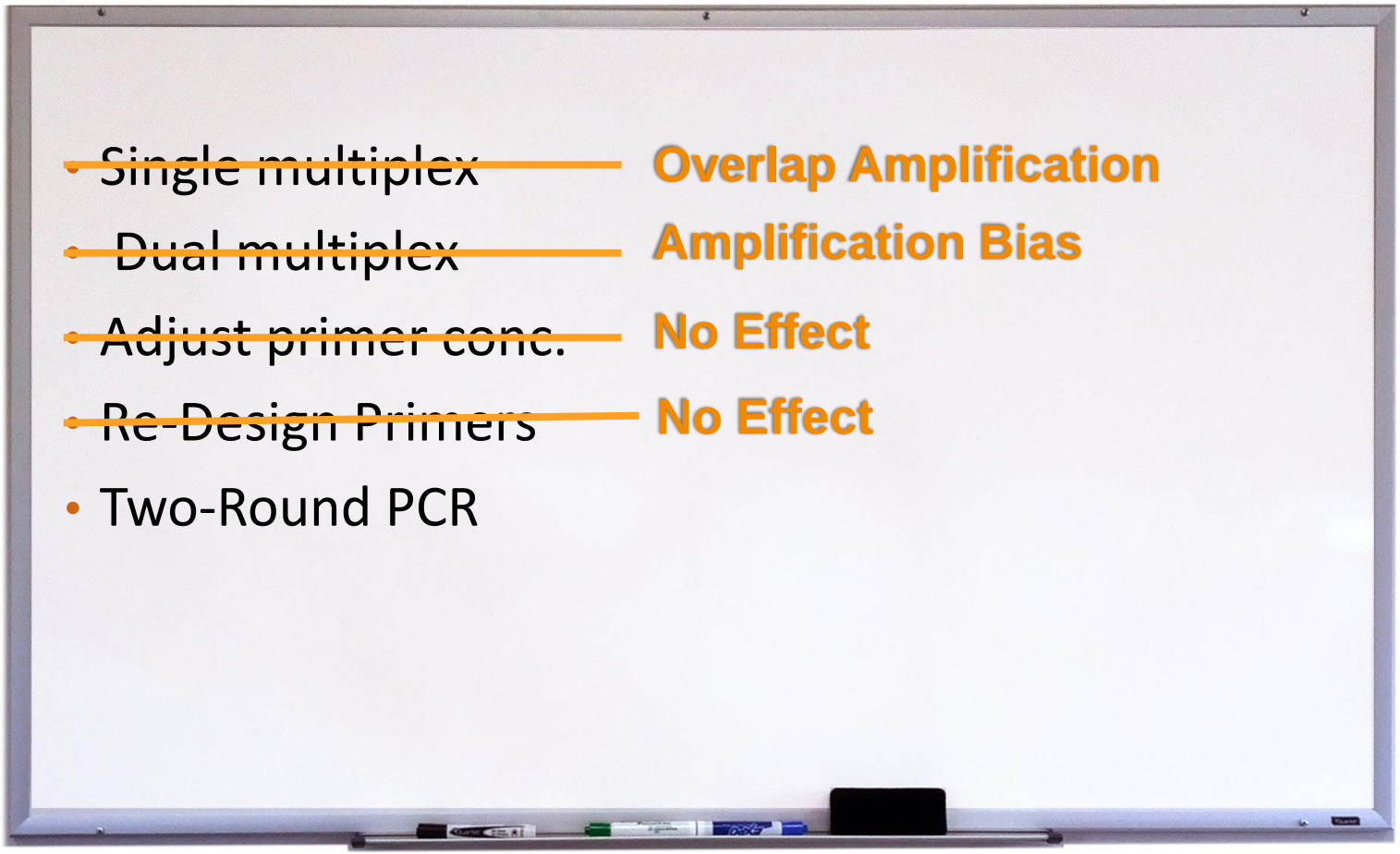
PCR Protocols



PCR Protocols



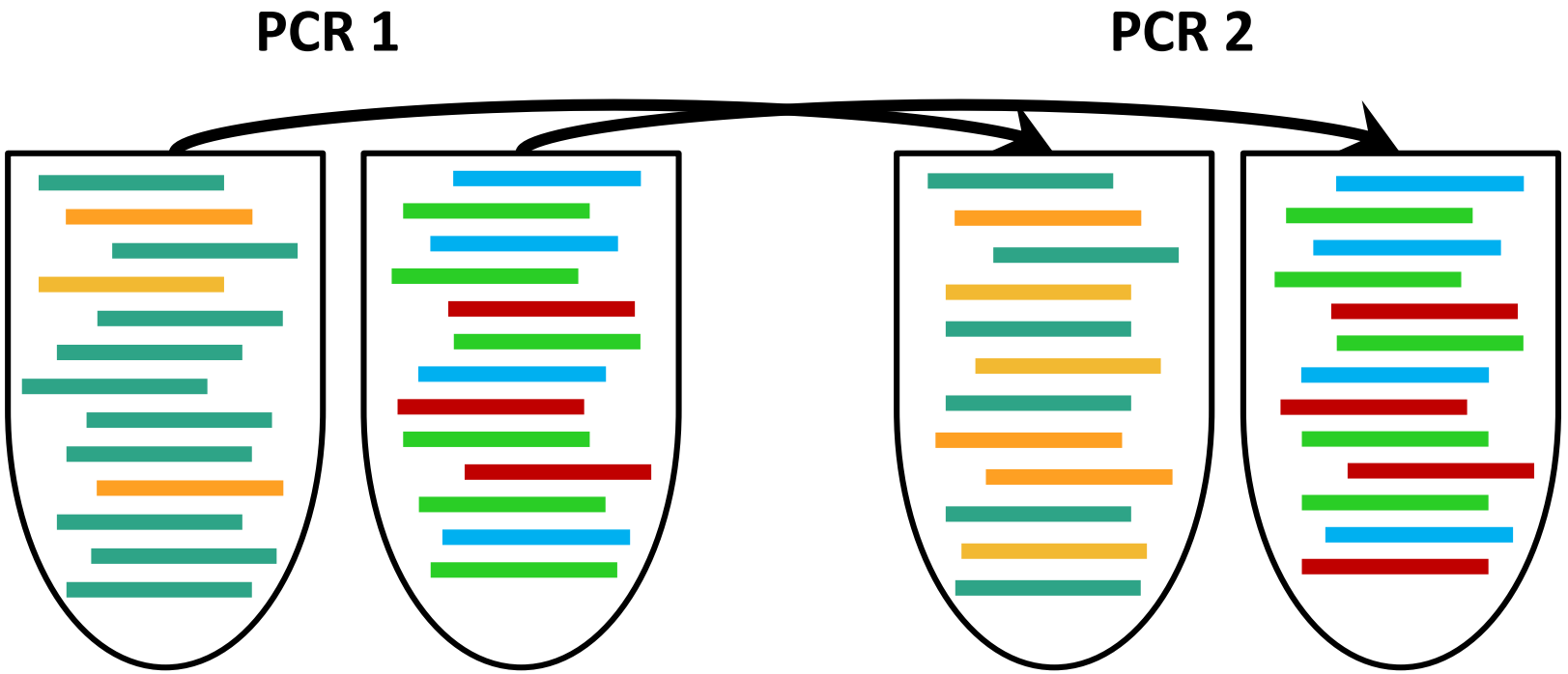
PCR Protocols



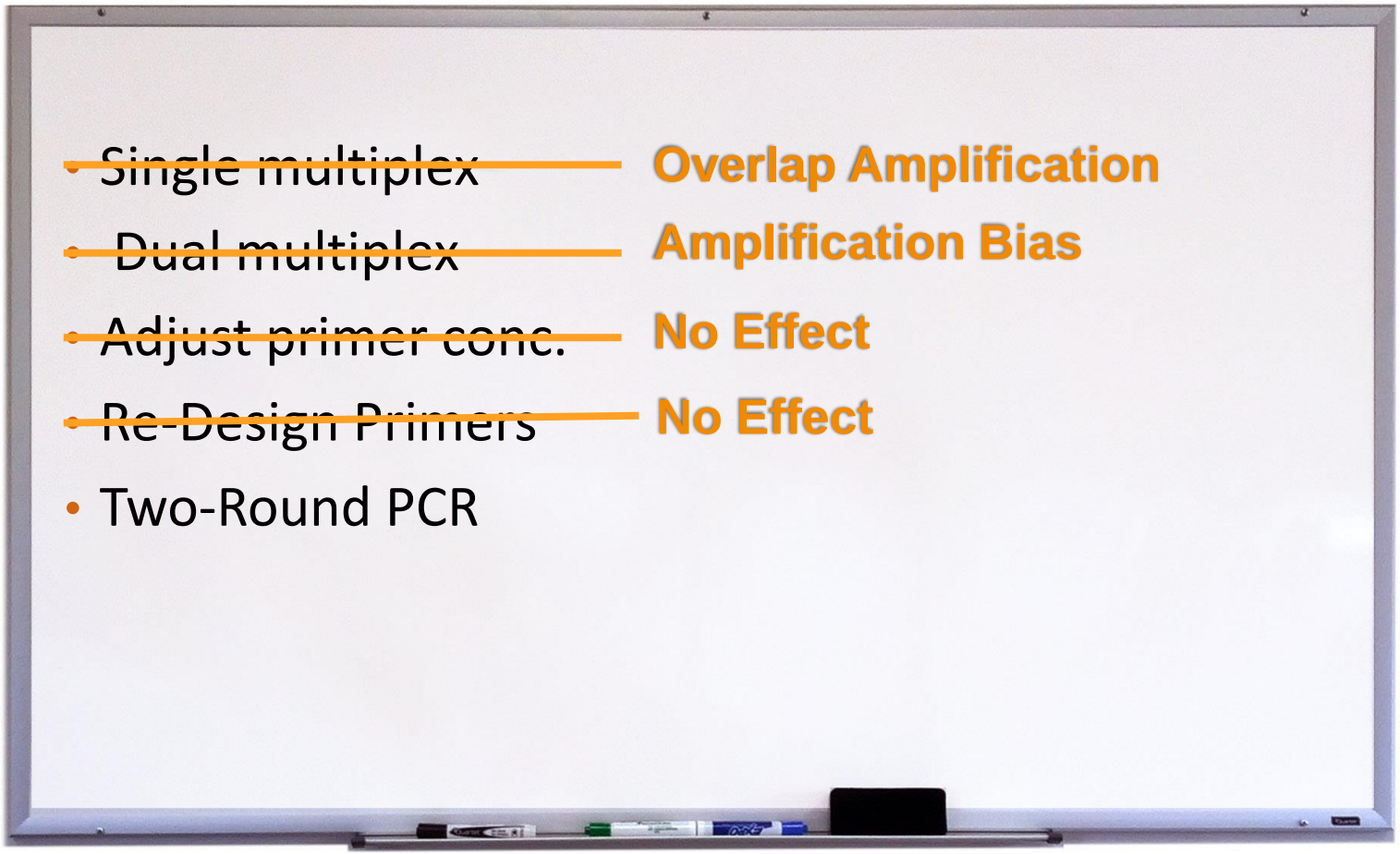
PCR Protocols

Two-Round PCR

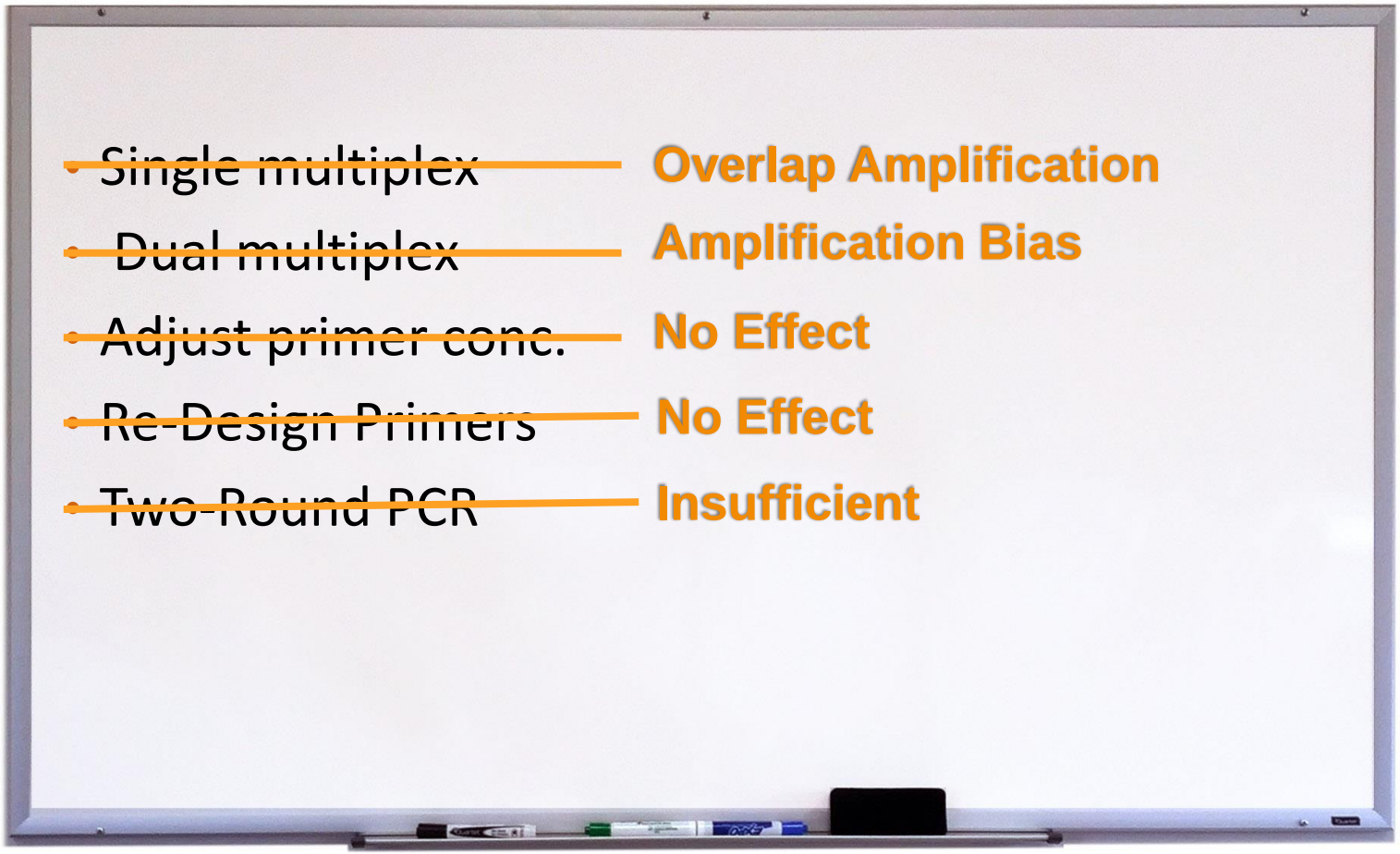
- 2 PCR to help reduce amplification bias → insufficient



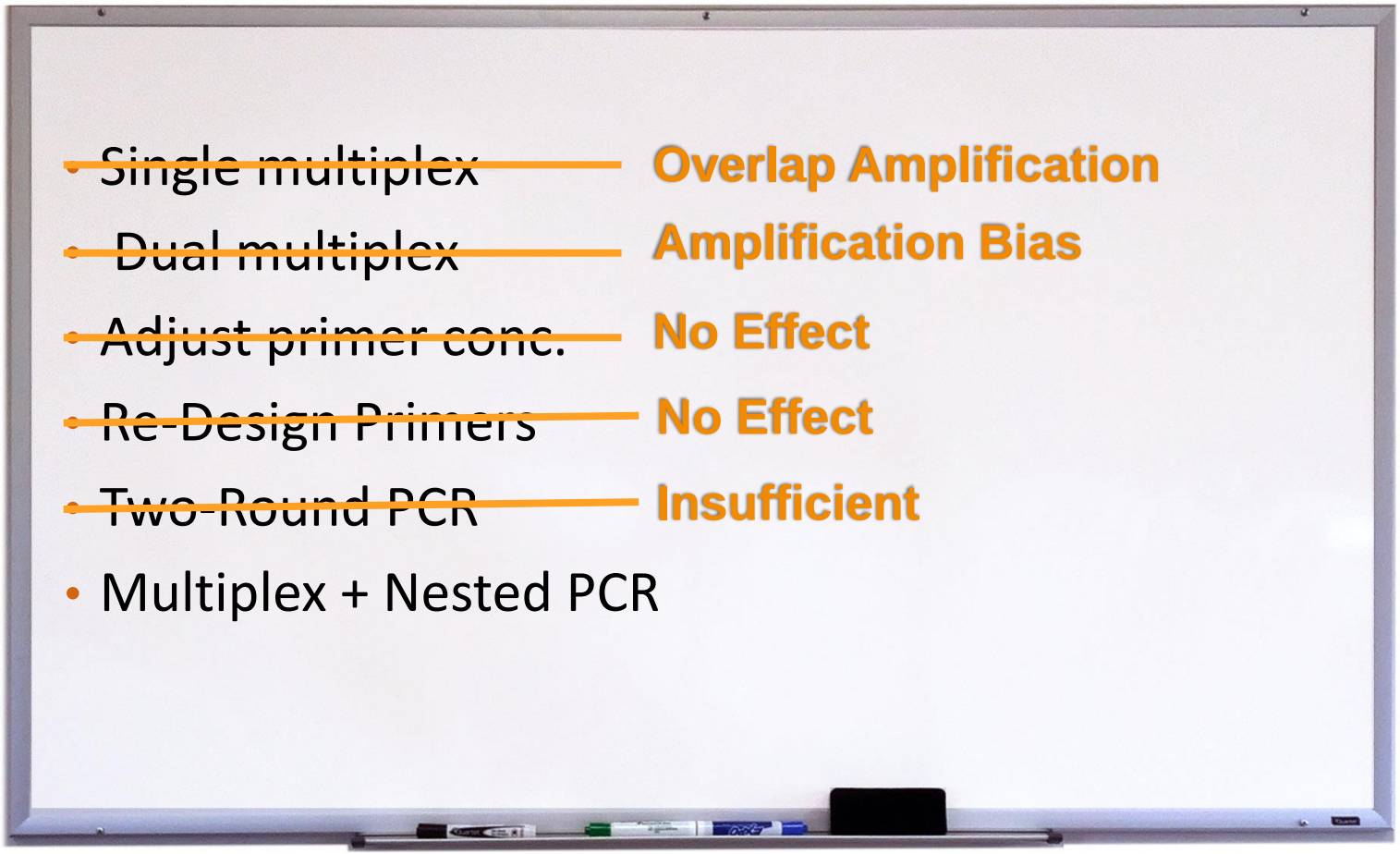
PCR Protocols



PCR Protocols

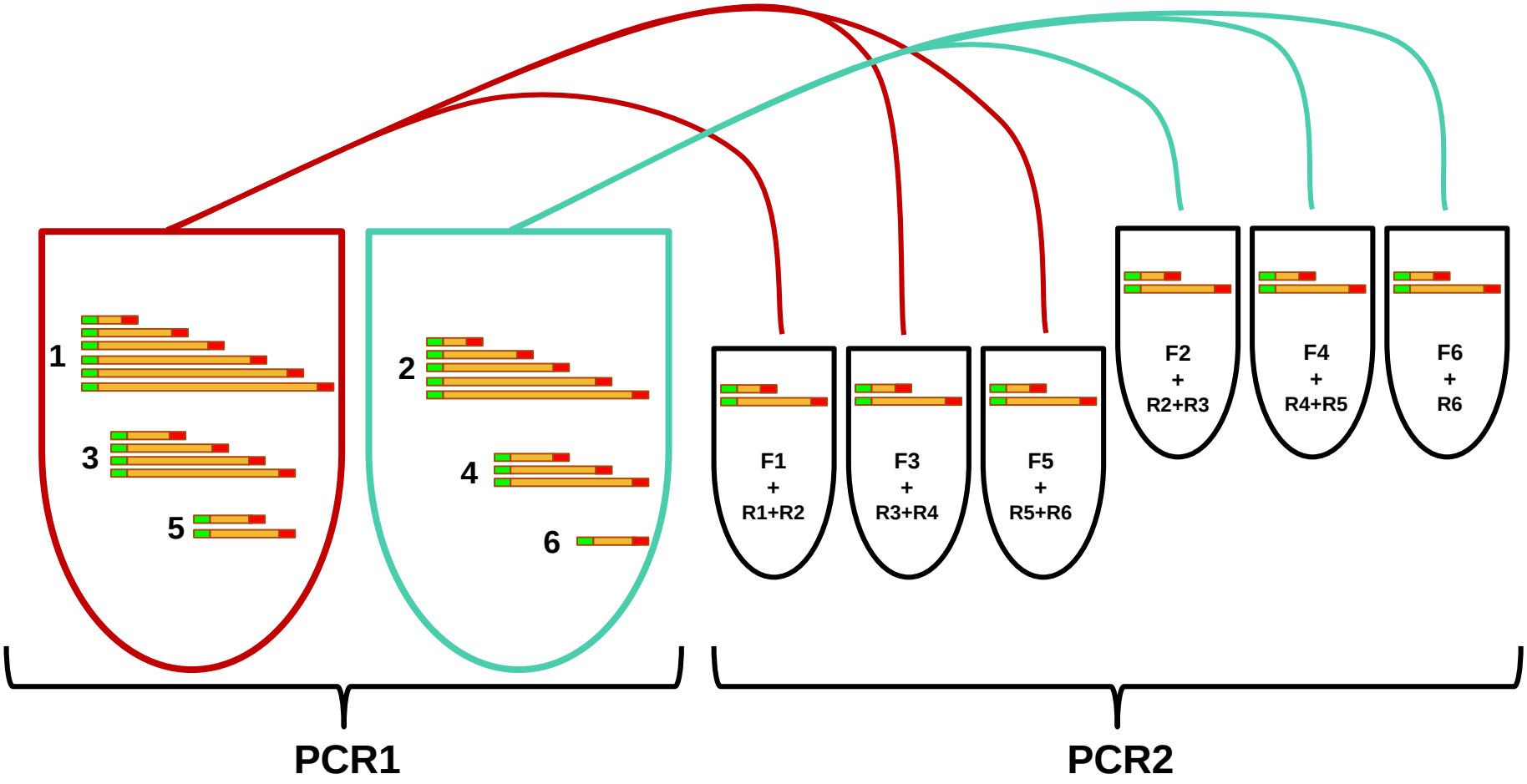


PCR Protocols



PCR Protocols

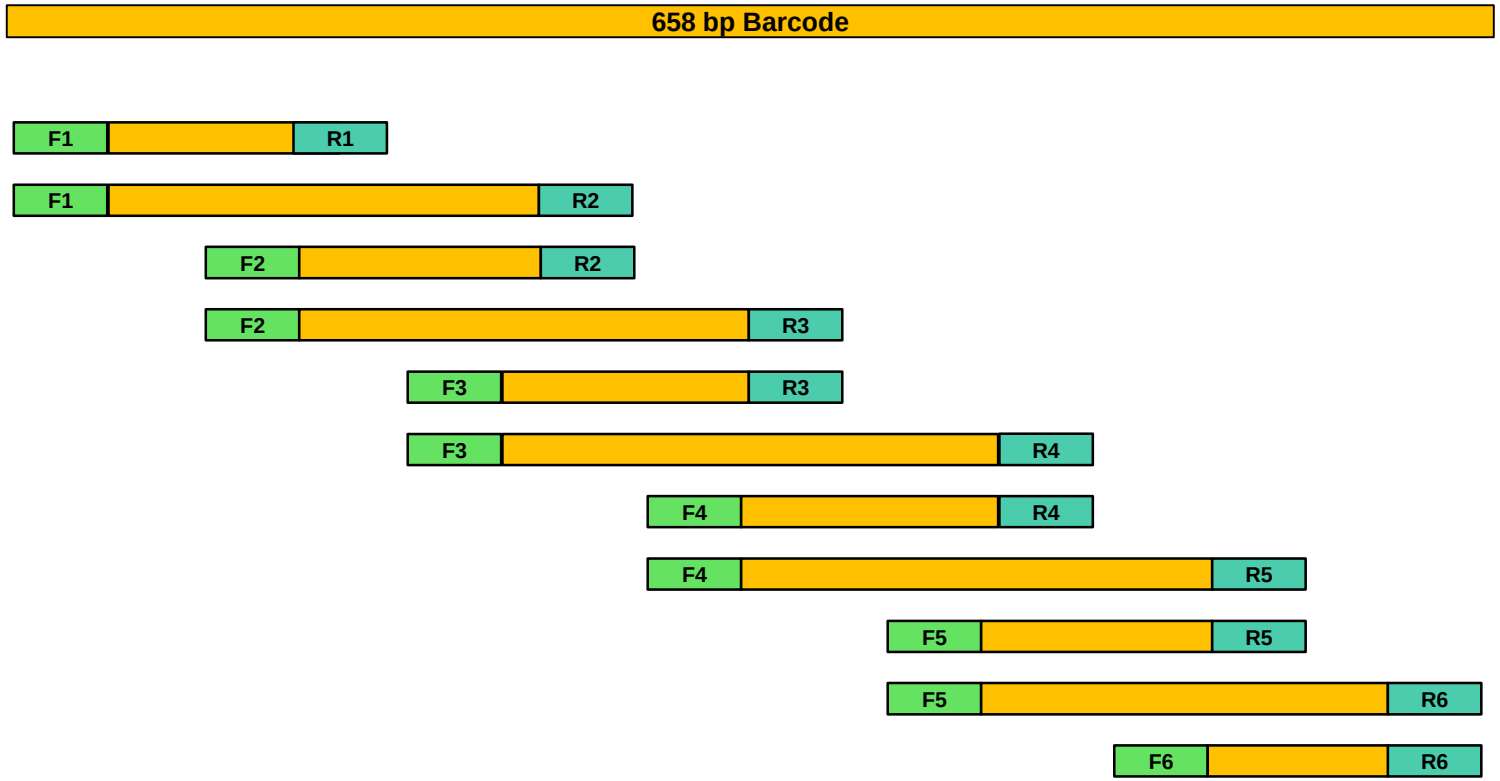
Multiplex + Nested PCR



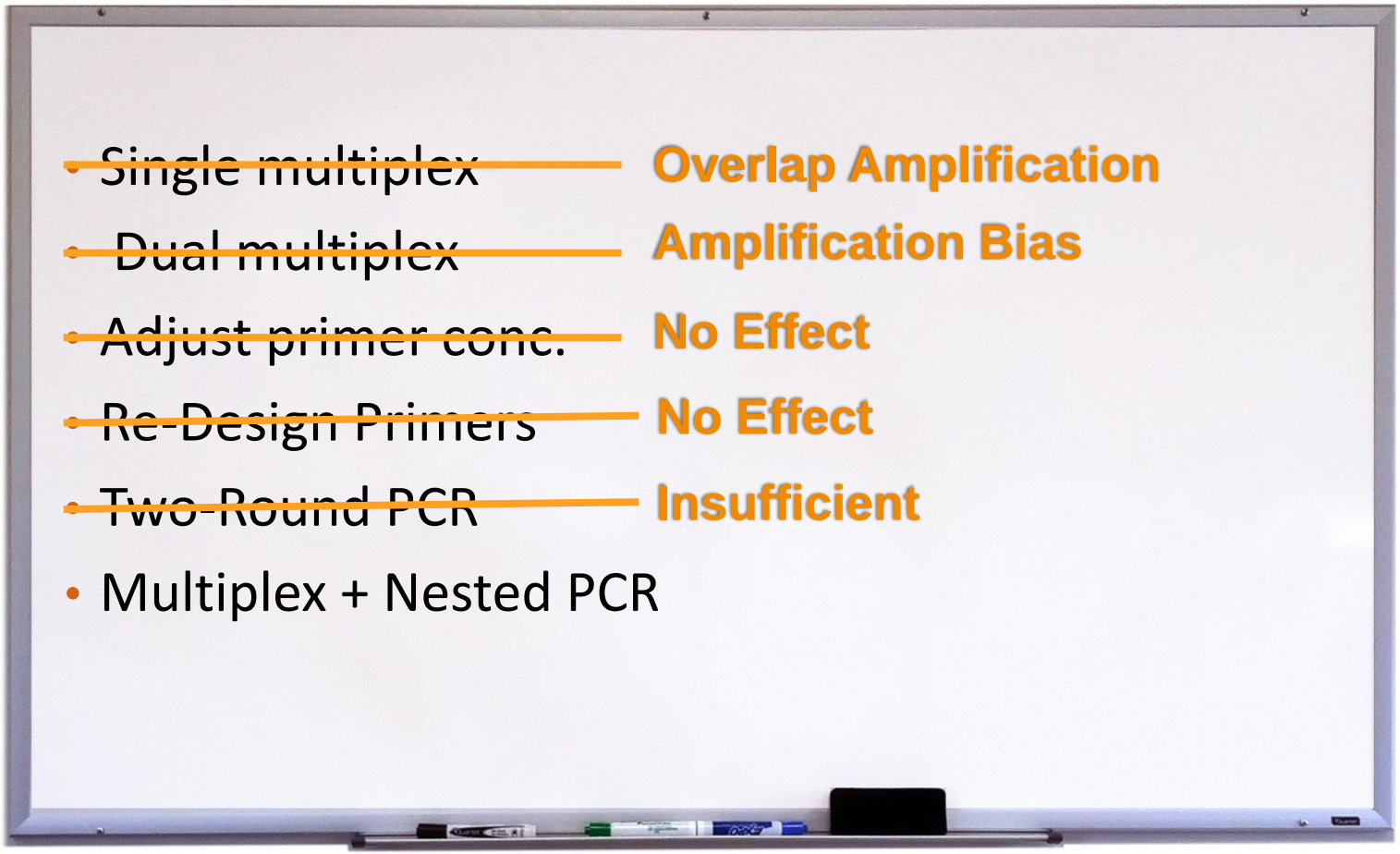
PCR Protocols

Multiplex + Nested PCR

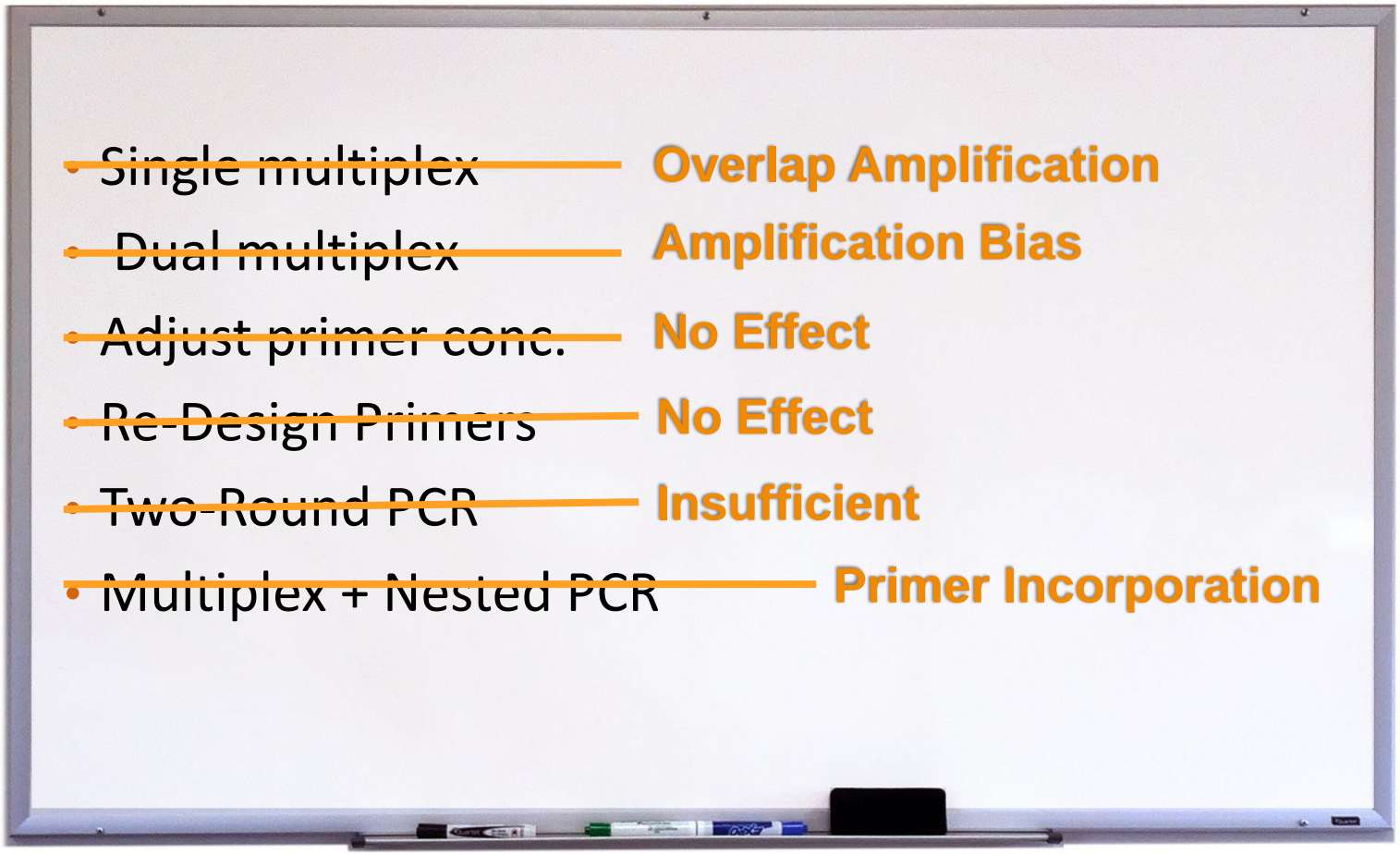
- Redundancy to increase chances of recovery



PCR Protocols

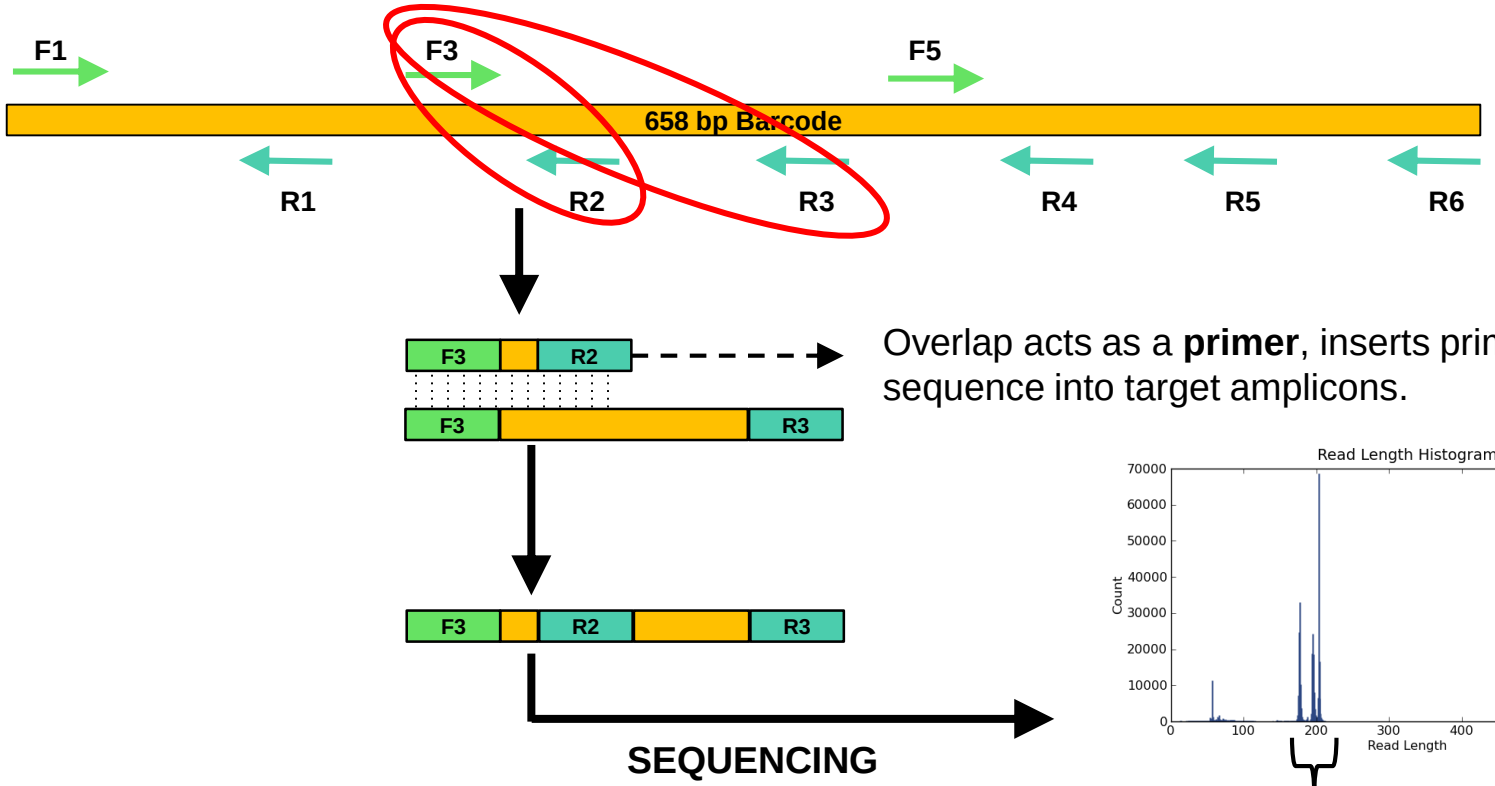


PCR Protocols



PCR Protocols

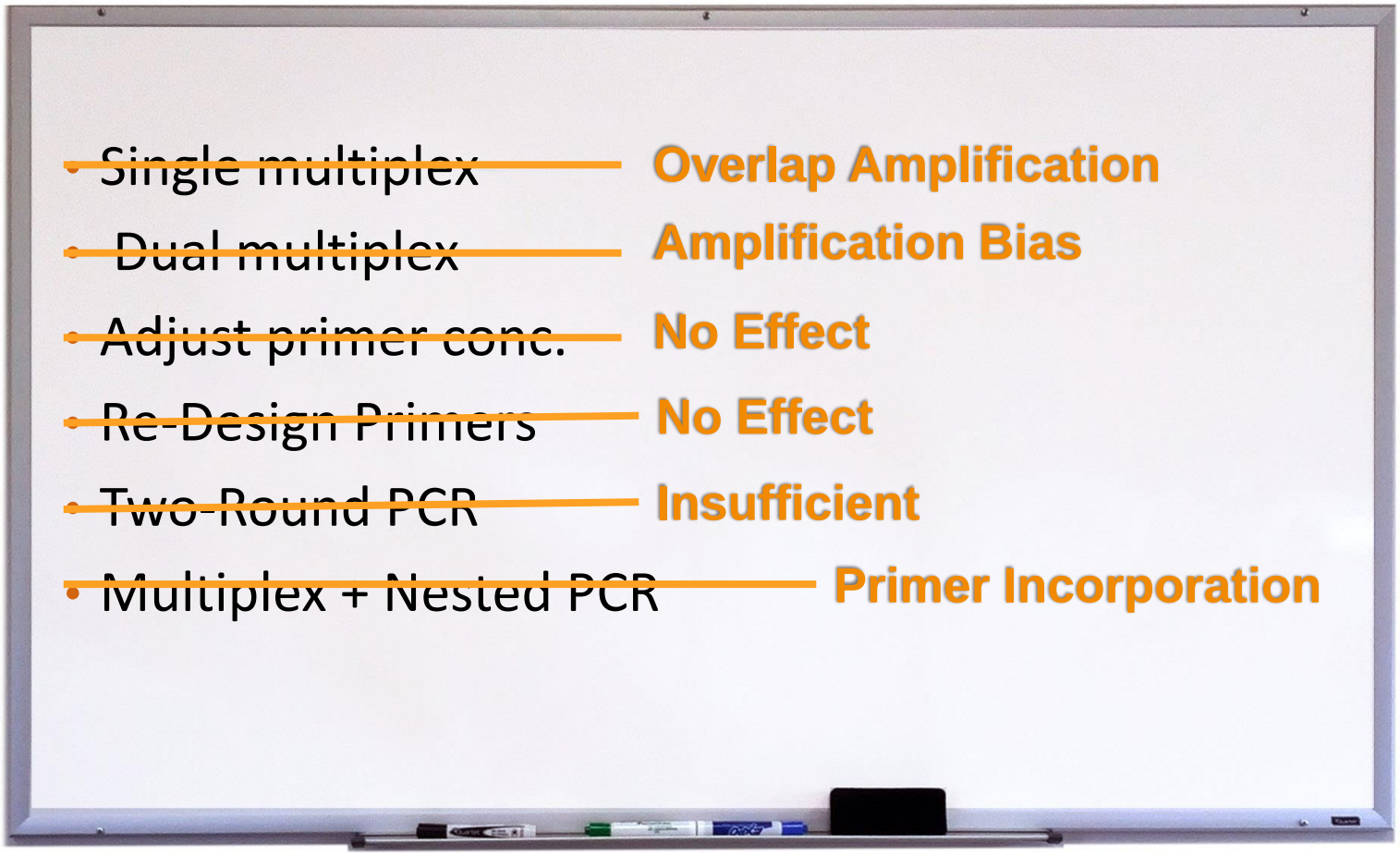
Multiplex + Nested PCR



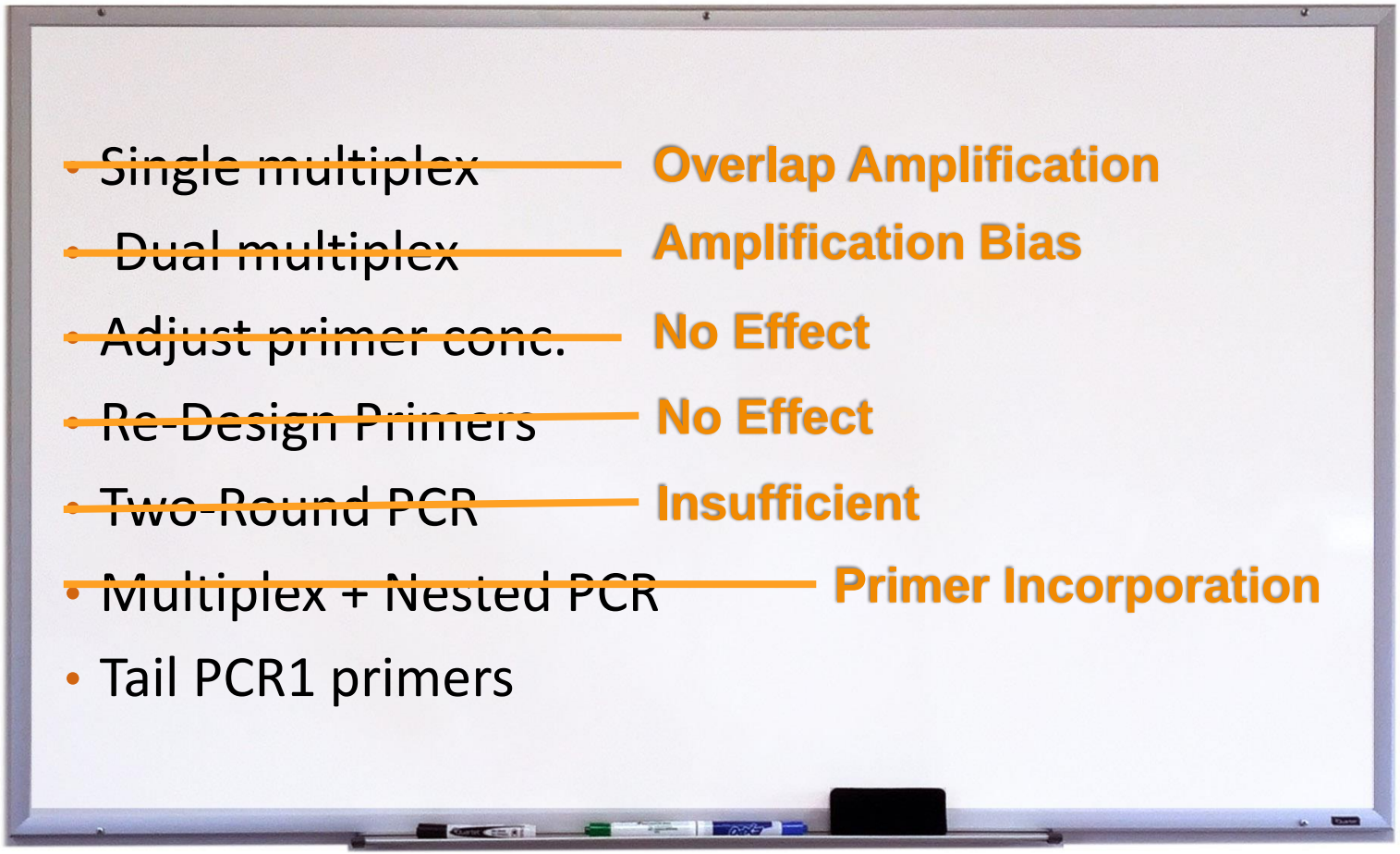
Overlap acts as a **primer**, inserts primer sequence into target amplicons.

X (chimeric reads)

PCR Protocols

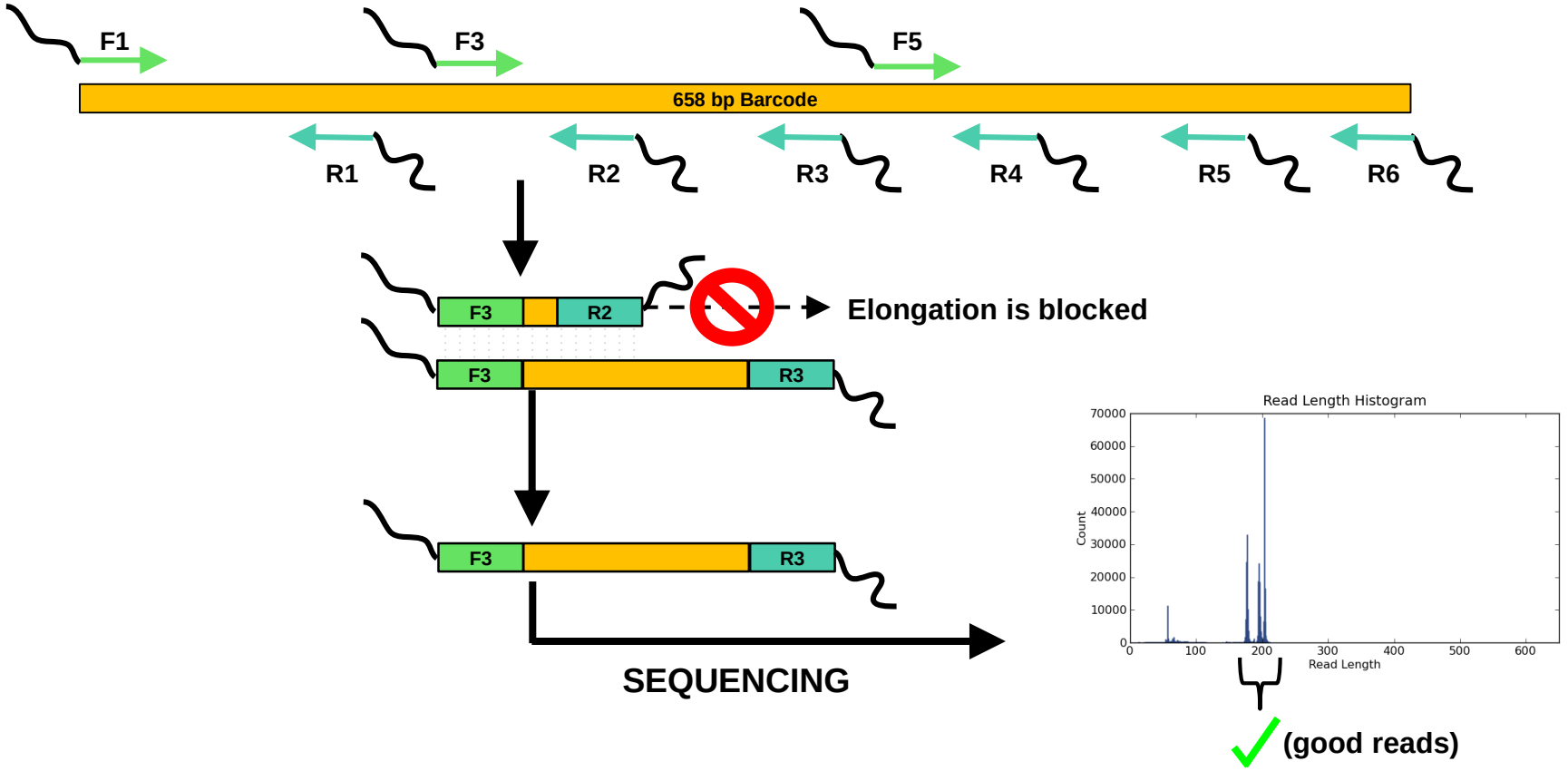


PCR Protocols

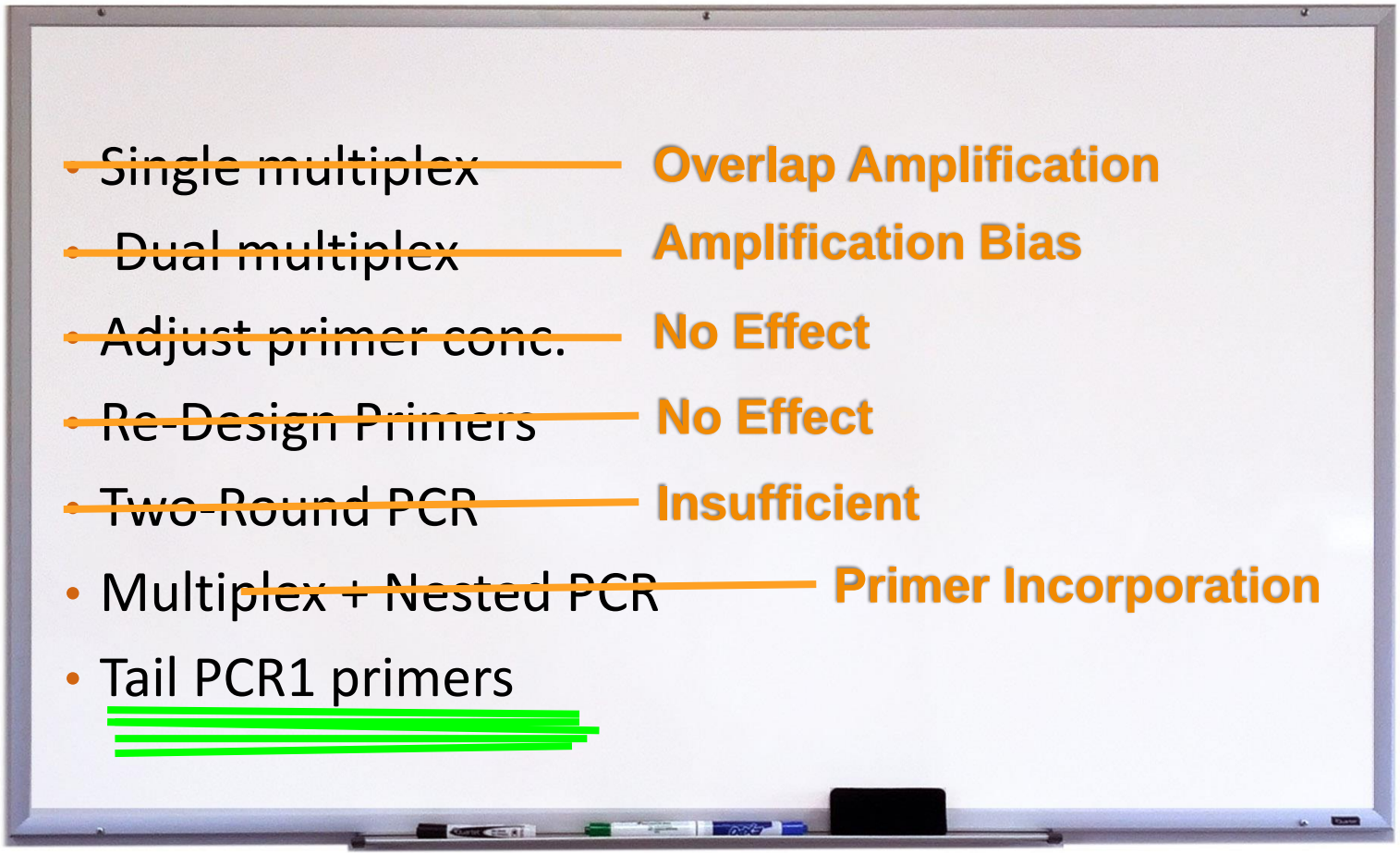


PCR Protocols

Tail PCR1 Primers

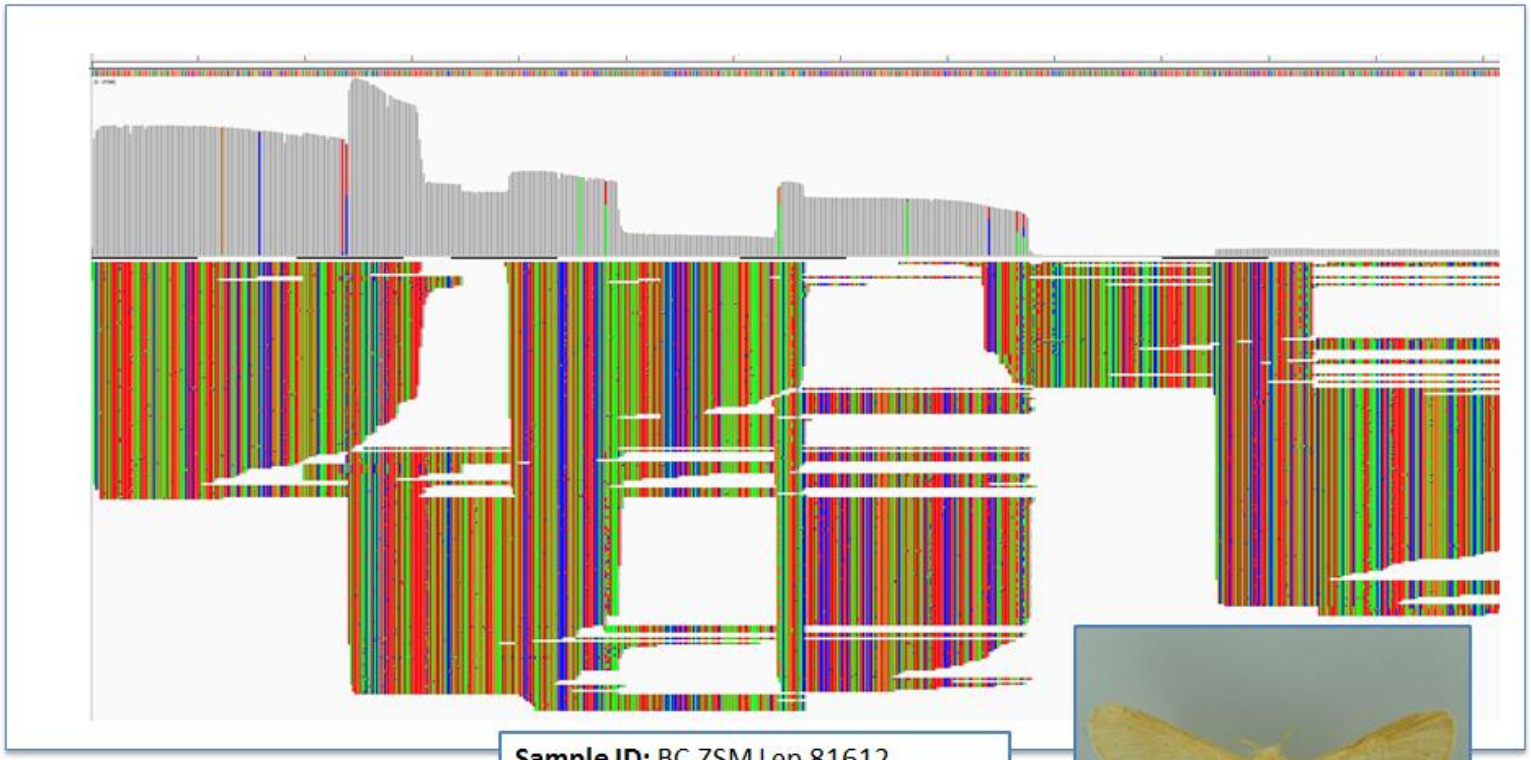


PCR Protocols



PCR Protocols

Reads Assembled into a Full-Length Barcode



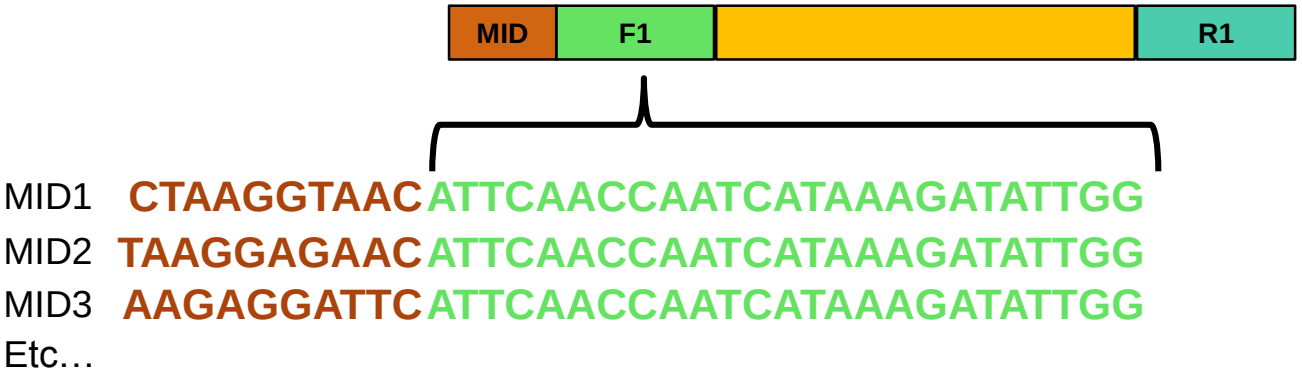
Sample ID: BC ZSM Lep 81612
Identification: *Lasiocampa decolorata*
Status: Holotype
Age: 182 yrs

MID-Tagging

Multiplex **ID**entifier tags

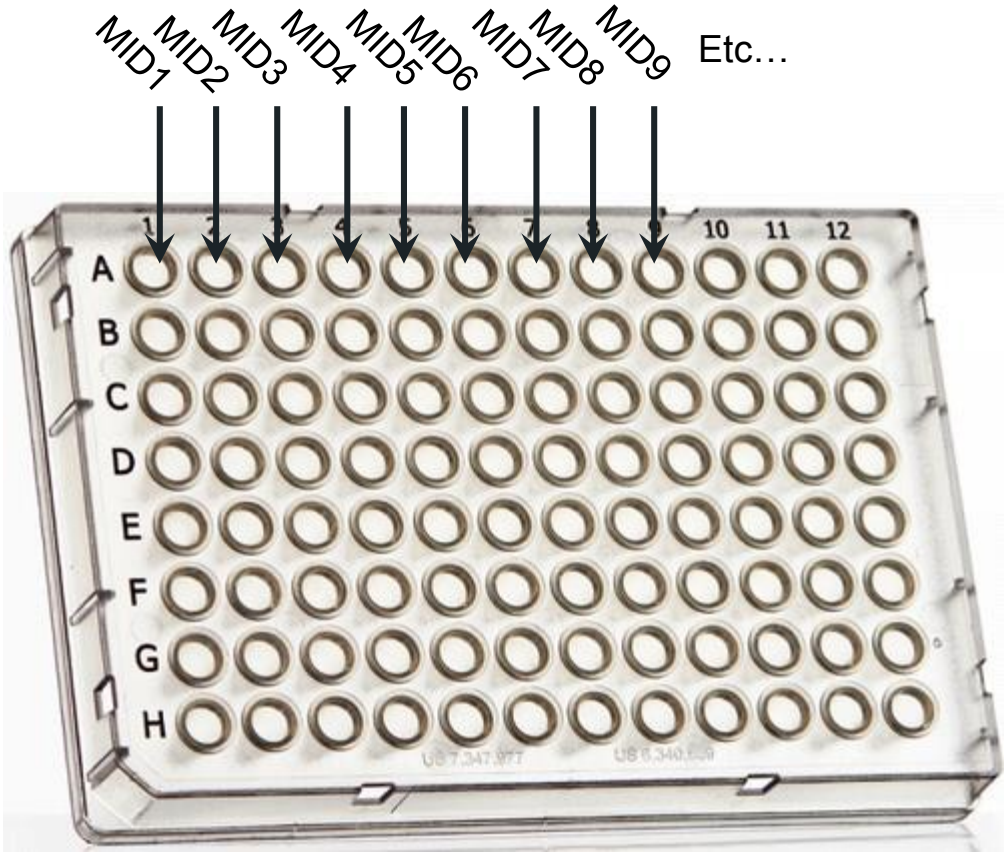
MID-Tagging

- Unique sequence fragment added before the primers for each sample
  - Fragment not usually found in nature



MID-Tagging

- Can add unique tag for each well (samples)

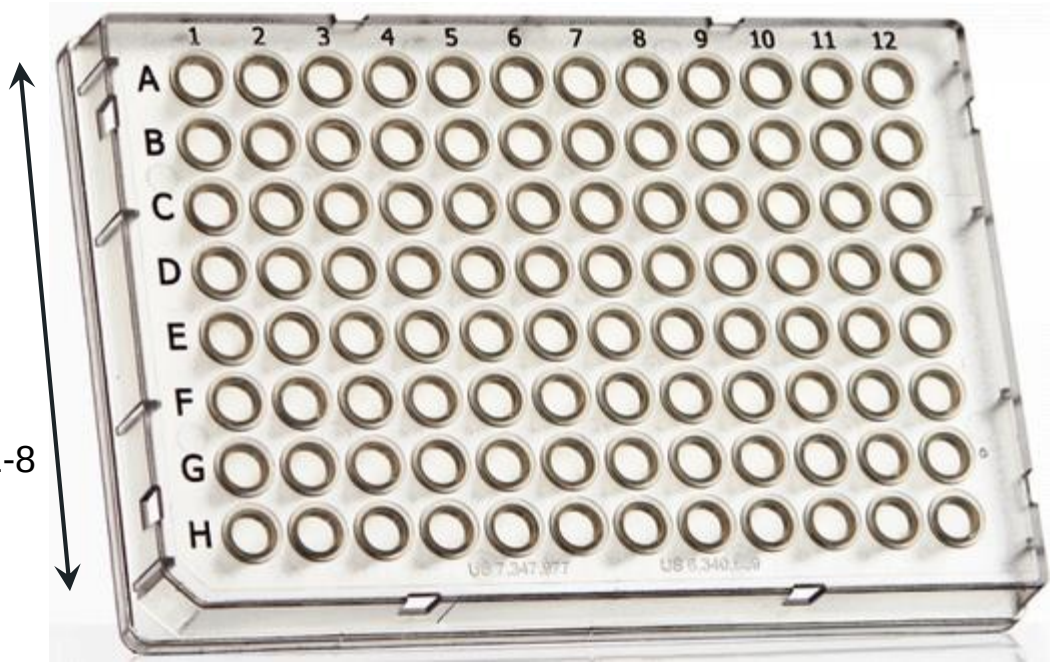


MID-Tagging

Duel MID-Tagging

- Unique tag for rows and one for columns
  - Cost effective
  - Scalable
  - Must be able to read both MID tags!

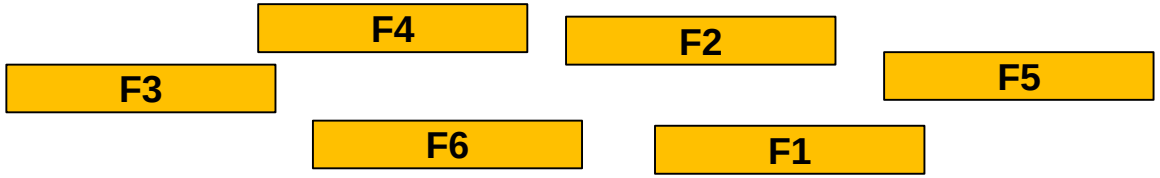
Forward MID 1-12



Reverse MID 1-8

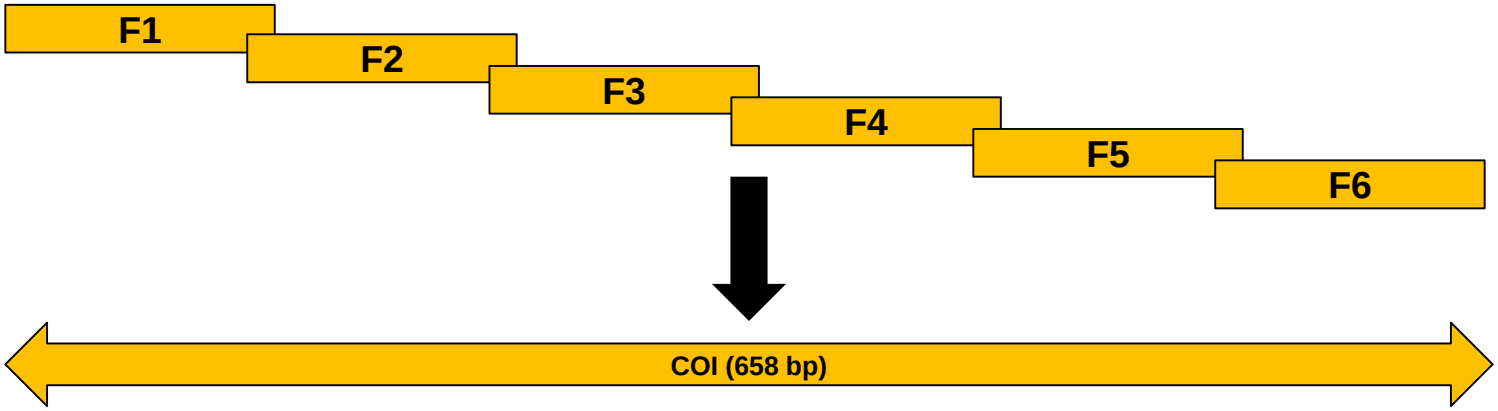
NGS Data Assembly and Analysis

1. Align to reference



OR

2. De novo assembly



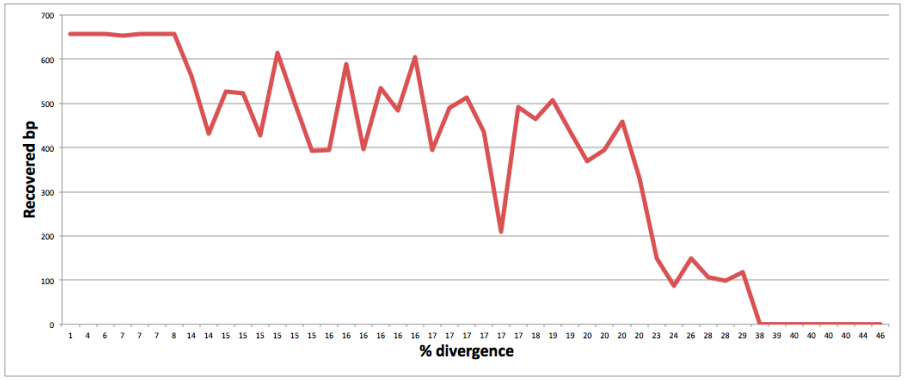
NGS Data Assembly and Analysis

Problems with Reference-Based Assembly

Dalopius tristis (Coleoptera) 

Identification	Percent Divergent	Notes	Recoverd bp
Dalopius tristis	0.6%	Same species	658
Dalopius asellus	3.8%	Same genus	658
Dalopius marginatus	6.2%	Same genus	658
Dalopius vagus	7%	Same genus	654
Dalopius naomii	7%	Same genus	657
Dalopius asellus	7%	Same genus	658
Dalopius pallidus	8%	Same genus	658
Agriotes avulsus	14%	Same family	561
Agriotes sordidus	14%	Same family	432
Agriotes obscurus	15%	Same family	527
Agriotes proximus	15%	Same family	523
Agriotes lineatus	15%	Same family	427
Agriotes acutus	15%	Same family	615
Agriotes brevis	15%	Same family	501
Agriotes tardus	15%	Same family	392
Agriotes limosus	16%	Same family	395
Agriotes ustulatus	16%	Same family	589
Agriotes quebecensis	16%	Same family	396
Agriotes pilosellus	16%	Same family	534
Agriotes stabilis	16%	Same family	484
Agriotes pubescens	16%	Same family	604
Agriotes acuminatus	17%	Same family	395
Agriotes mancus	17%	Same family	489
Agriotes apicalis	17%	Same family	513
Agriotes insanus	17%	Same family	435
Agriotes gallicus	17%	Same family	210
Agriotes sputator	17%	Same family	492
Agriotes pallidulus	18%	Same family	465
Agriotes fucosus	19%	Same family	508
Agriotes collaris	19%	Same family	438
Podeonius acuticornis	19.8%	Same family	370
Agriotes oblongicollis	20%	Same family	395

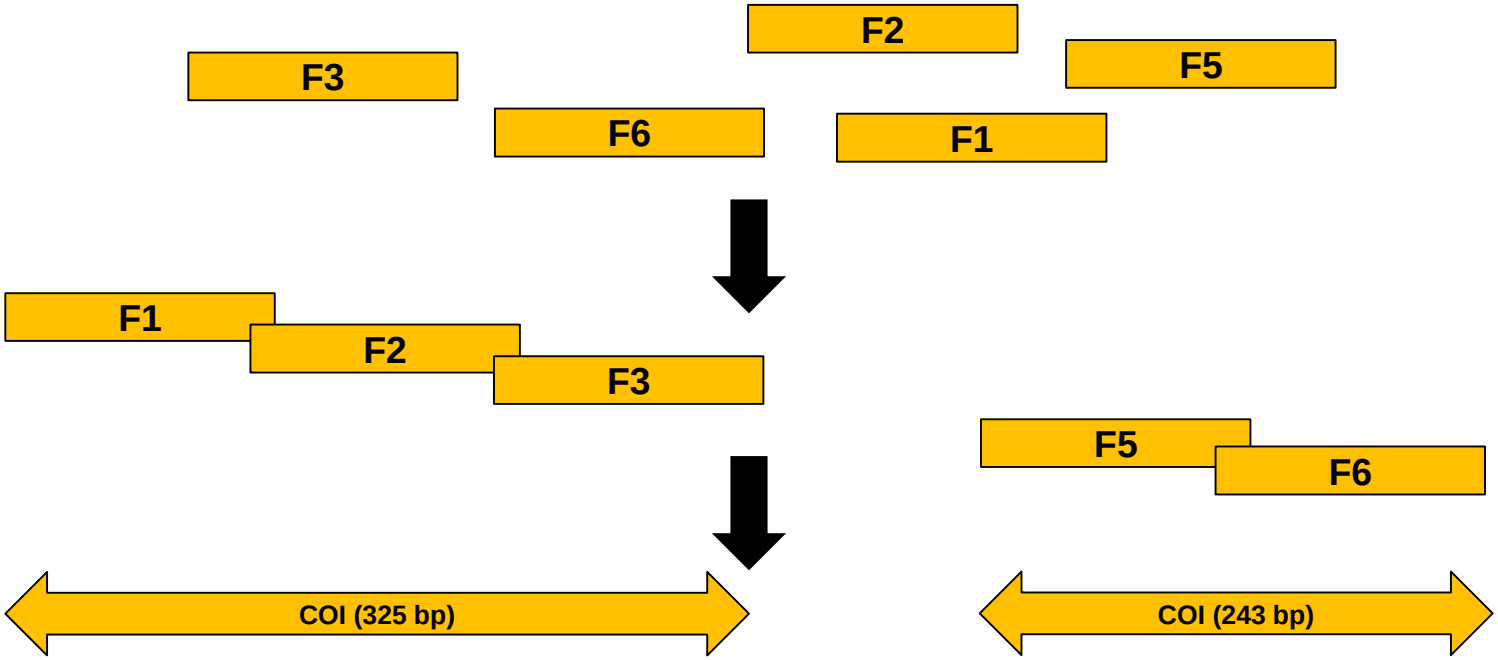
Identification	Percent Divergent	Notes	Recoverd bp
Drosophila melanogaster	20.1%	Fly	459
Gryllus campestris	20.3%	Dragonfly	330
Danaus plexippus	23.4%	Monarch butterfly	150
Mulsanteus arizonensis	24.2%	Same family	87
Tettigonia viridissima	25.7%	Cricket	149
Acanthosoma haemorrhoidale	27.6%	Shield bug	106
Xyleborinus saxeseni	28.2%	Same order	98
Homarus americanus	29.1%	Lobster	119
Rana sylvatica	38.3%	Frog	0
Apis mellifera	38.7%	Honey bee	0
Opistophthalmus macer	39.6%	Scorpion	0
Castor canadensis	39.8%	Beaver	0
Oncorhynchus mykiss	40.1%	Rainbow trout	0
Larus delawarensis	40.1%	Sea gull	0
Cyanea capillata	43.7%	Lion's mane jellyfish	0
Thamnophis sirtalis	46.3%	Garter snake	0



NGS Data Assembly and Analysis

Problems with de novo Assembly

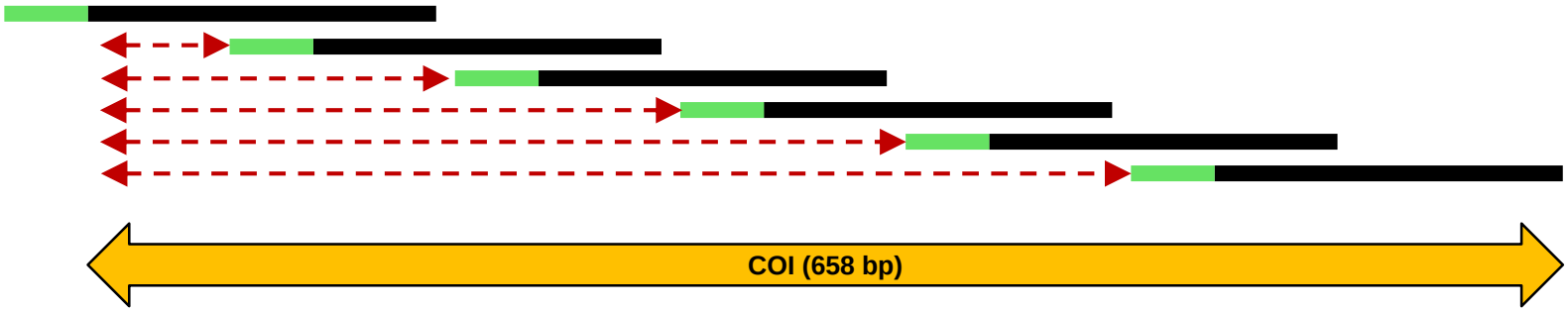
- If a fragment is not recovered → Obtained 2 short seq



NGS Data Assembly and Analysis

Primer Guided de novo Assembly

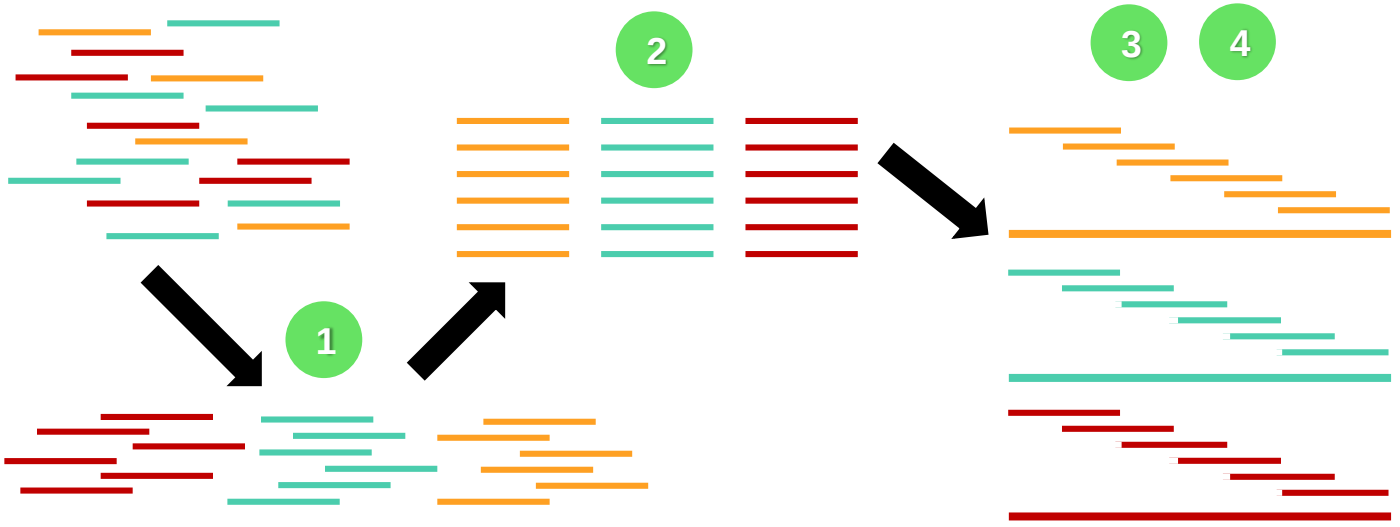
- Looking for the bp location where the primer starts to assemble the different fragments together



NGS Data Assembly and Analysis

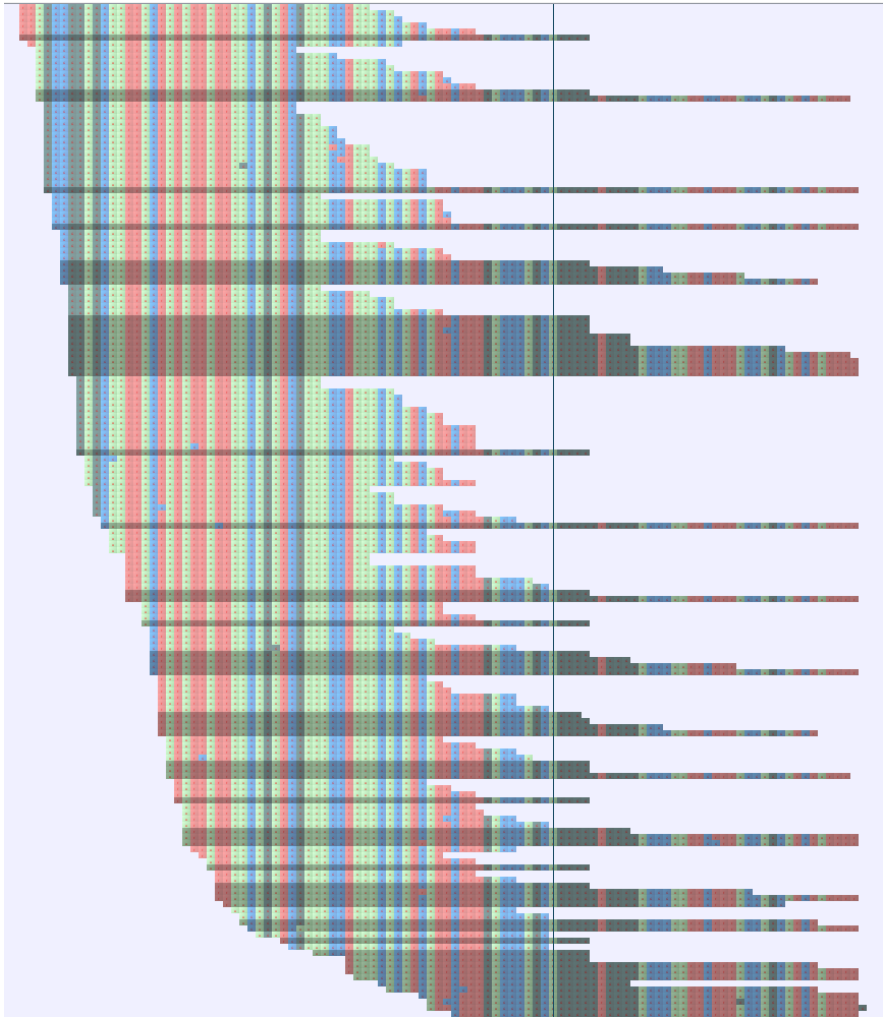
Primer Guided de novo Assembly

1. Assign reads to samples
2. Assign reads to a fragment based on primers
3. Insert N's in front of reads to force into alignment
4. Take majority consensus of entire assemblage



NGS Data Assembly and Analysis

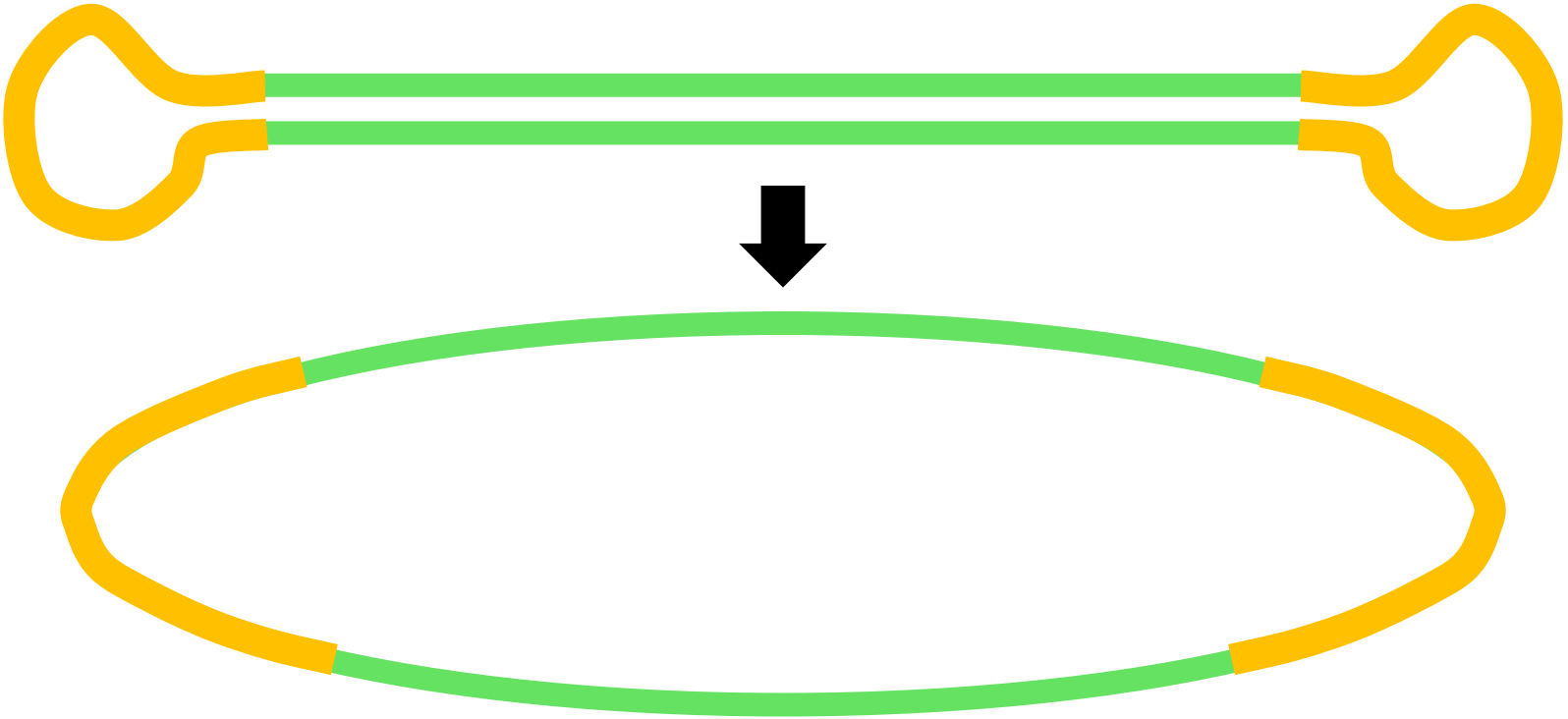
- **Primers are often not visible in reads produced by second generation platforms:**
  - Unidirectional sequencing
  - Sequencing errors
  - Quality trimming



NGS Platforms

Single Molecule Real Time Sequencing

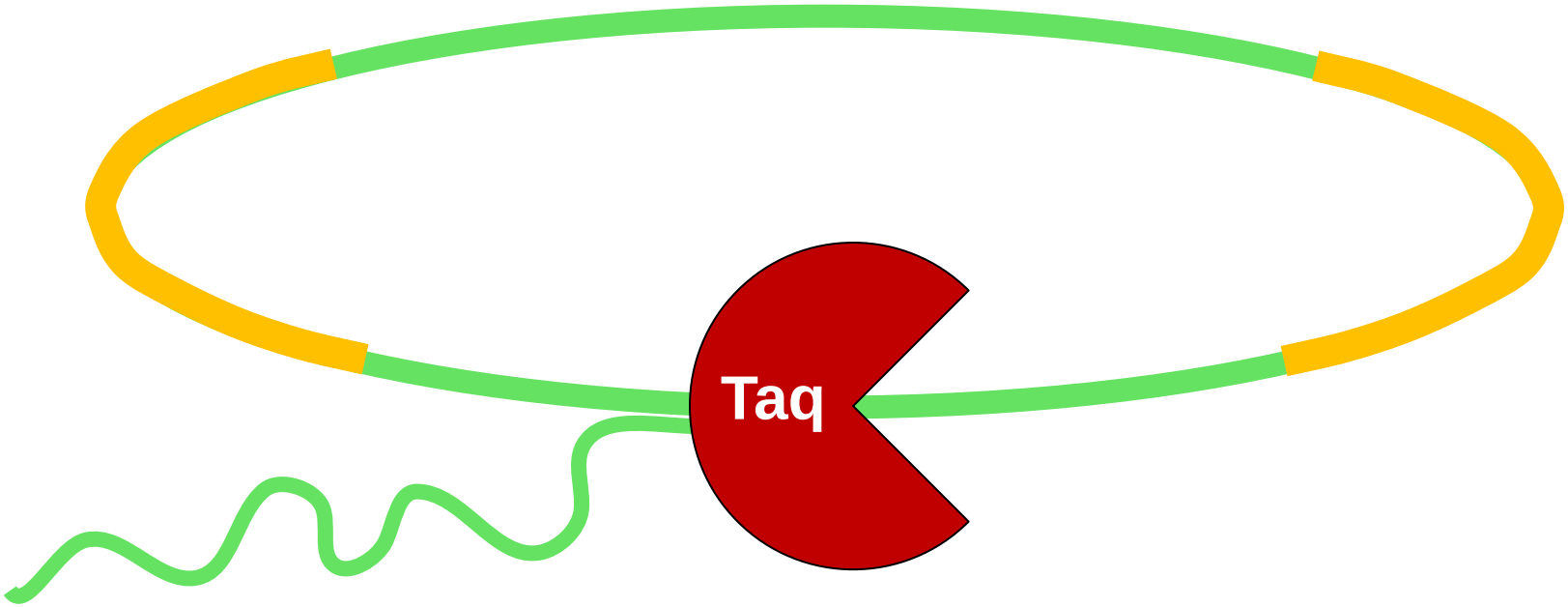
- Addition of SMRT bell adapters at each end of the DNA fragment to turn it into a circular form



NGS Platforms

SMRT Sequencing

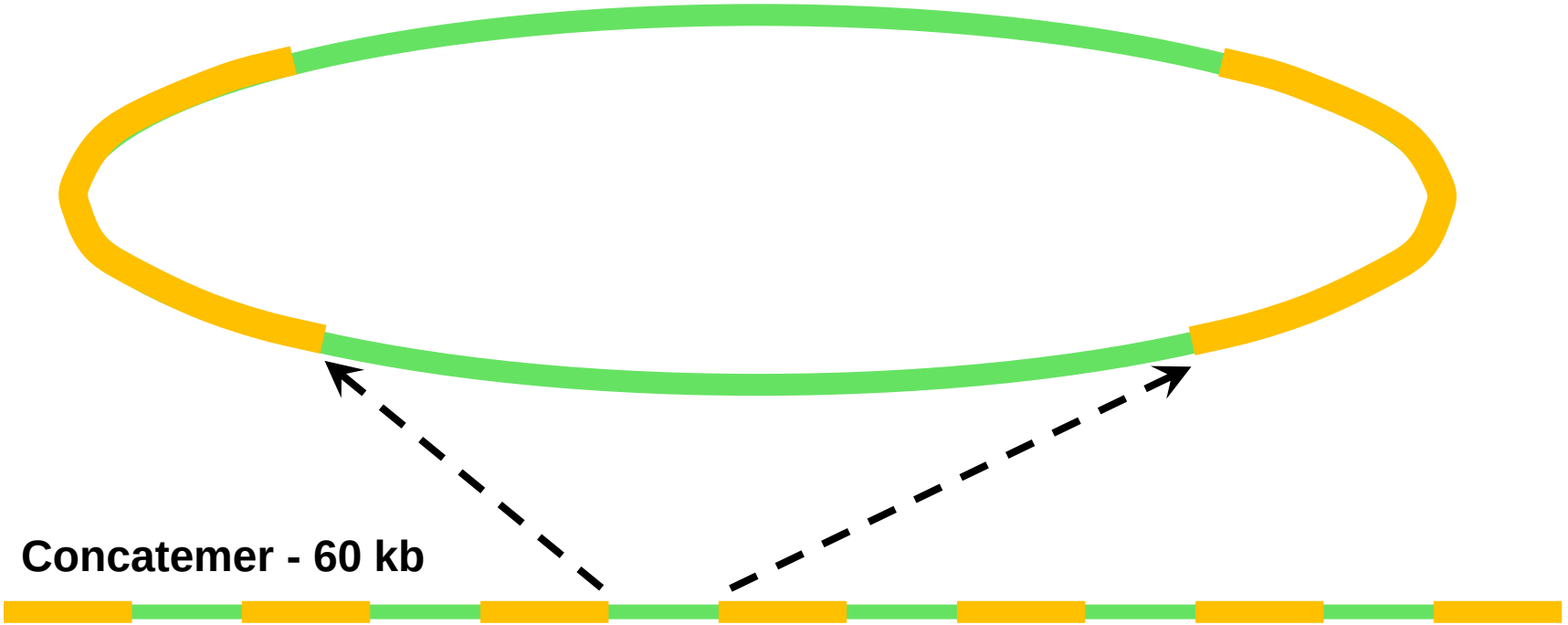
- Multiple passes of DNA polymerase



NGS Platforms

SMRT Sequencing

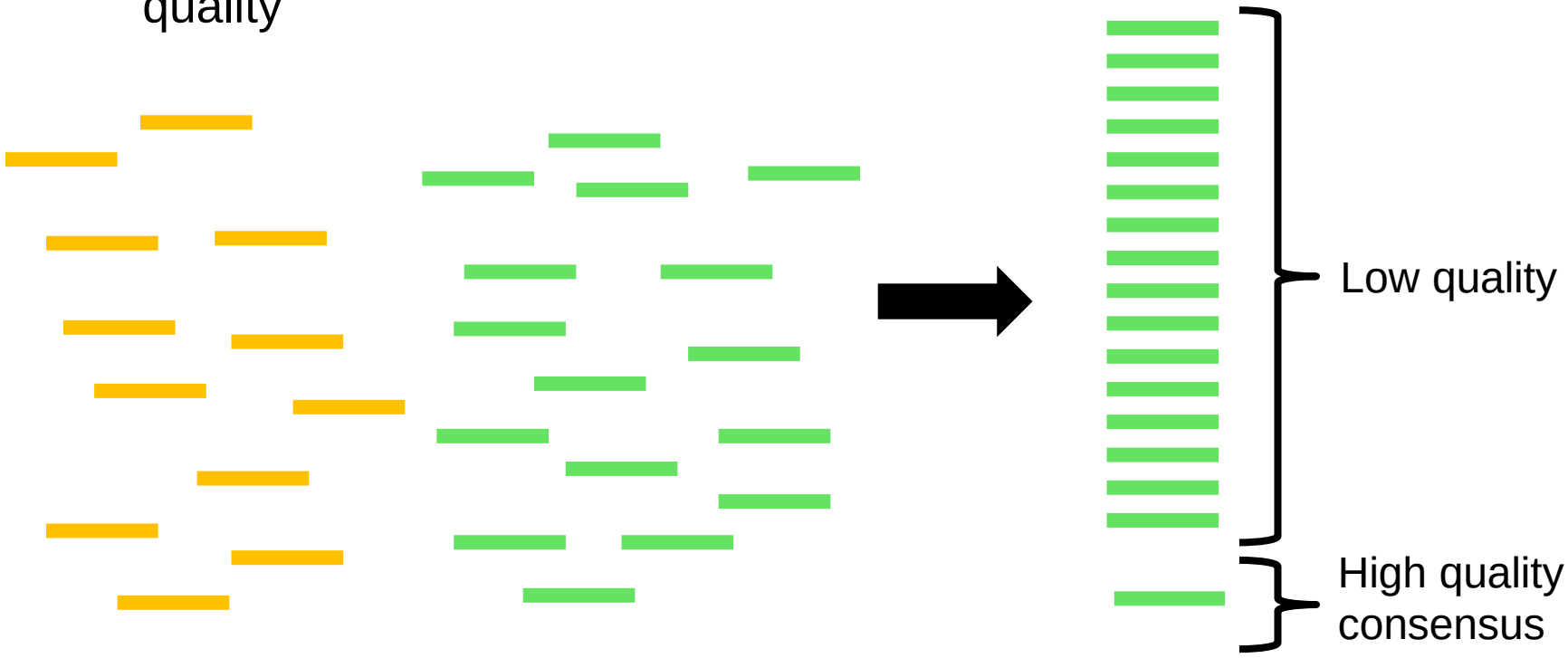
- Results in several short DNA fragments of low quality for the same section



NGS Platforms

SMRT Sequencing

- After removing all the SMR bell adapters:
  - Create a consensus to obtain the final DNA sequence of high quality

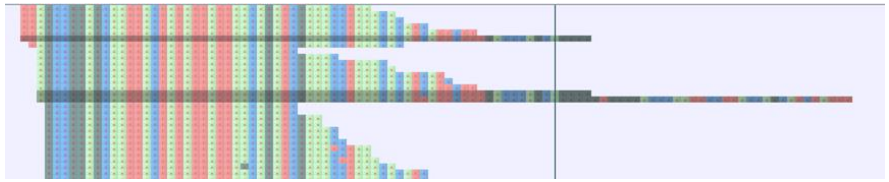


NGS Platforms

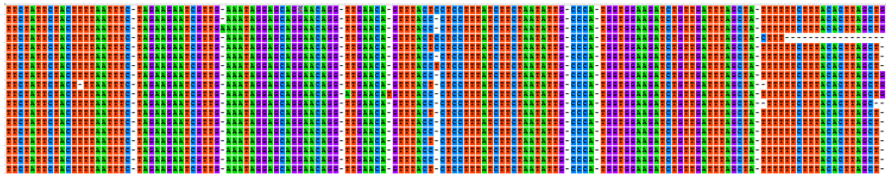
Advantages of SMRT Sequencing

- High quality, full-length reads
  - More confidence in low coverage areas
  - Reference free “de novo” assembly
  - Can use MID-Tags at each end of amplicon
    - Increase throughput at almost no cost (asymmetrical tagging)
    - Can de-multiplex using either end of read (symmetrical tagging)

ION TORRENT

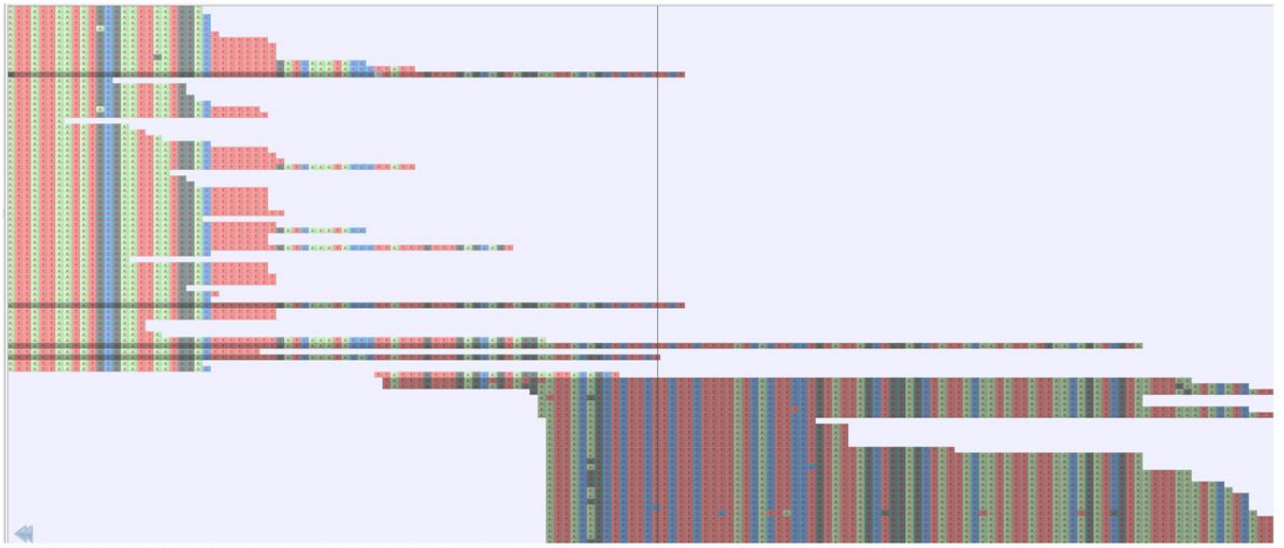


SMRT

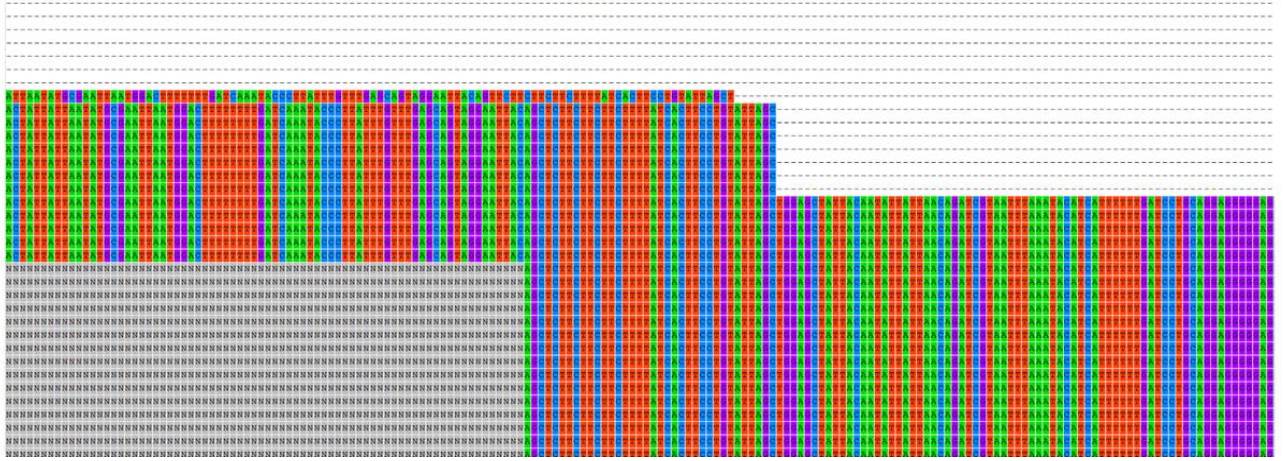


NGS Platforms

Ion Torrent with reference sequence



PacBio without reference sequence

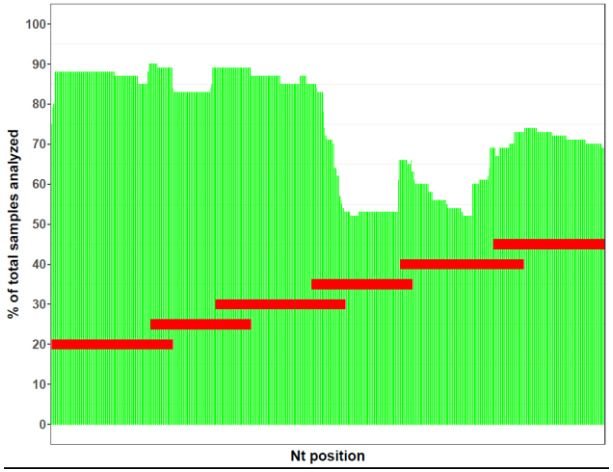


NGS Platforms

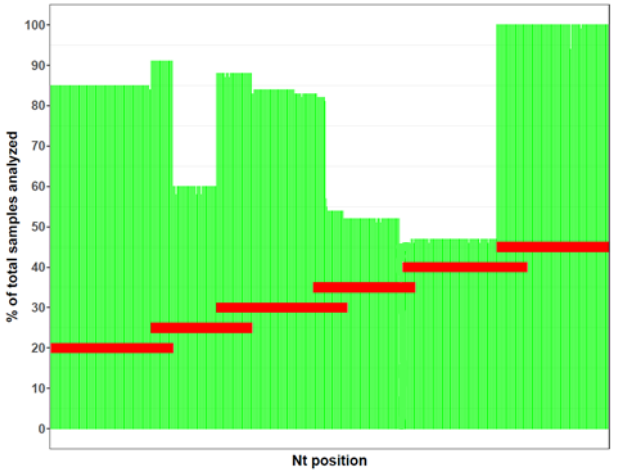
Disadvantages of SMRT Sequencing

- Lower throughput
  - Effects of amplification bias will be more pronounced

Ion Torrent PGM - 5M reads



PacBio Sequel - 250K reads

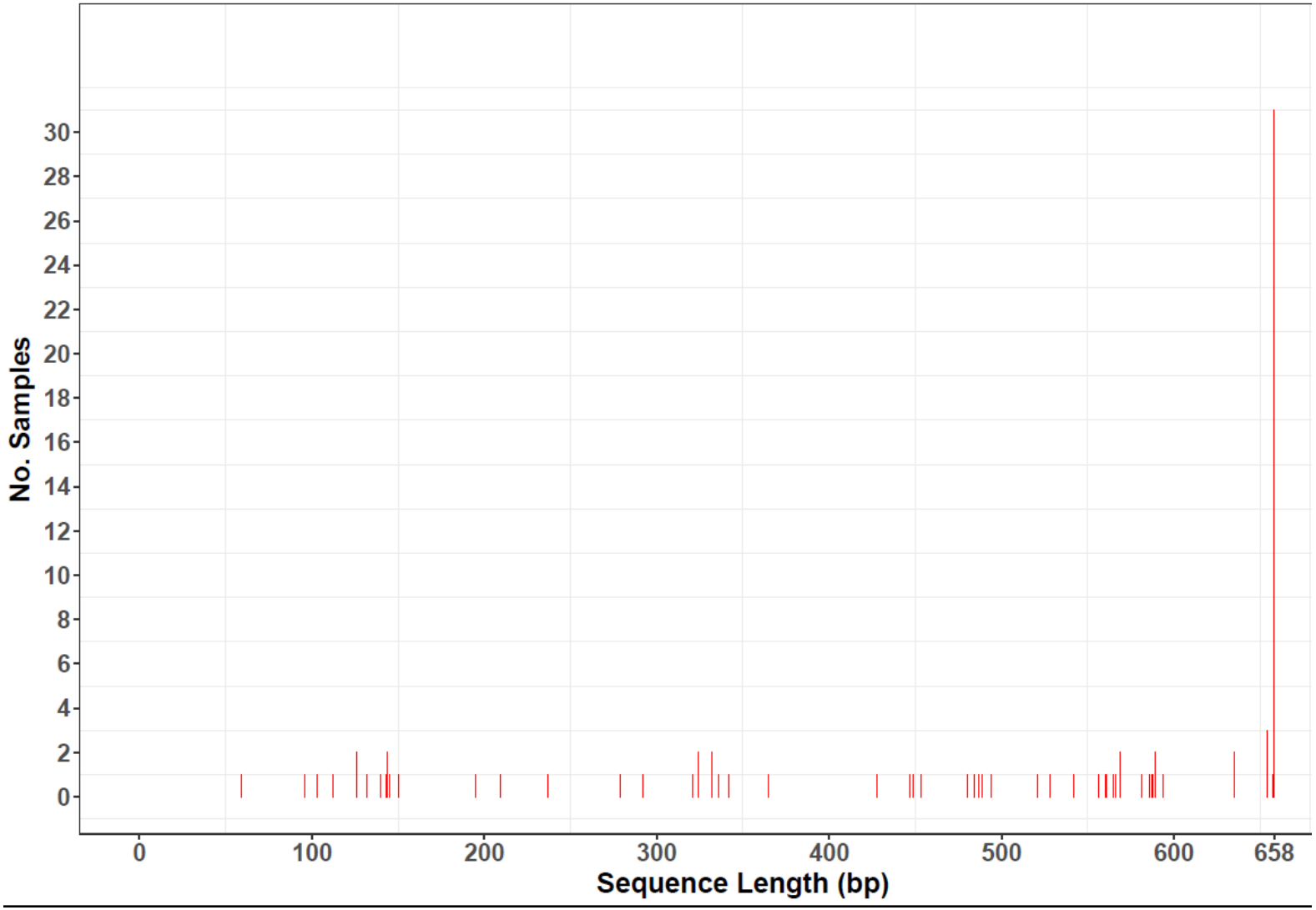


NGS Platforms

Disadvantages of SMRT Sequencing

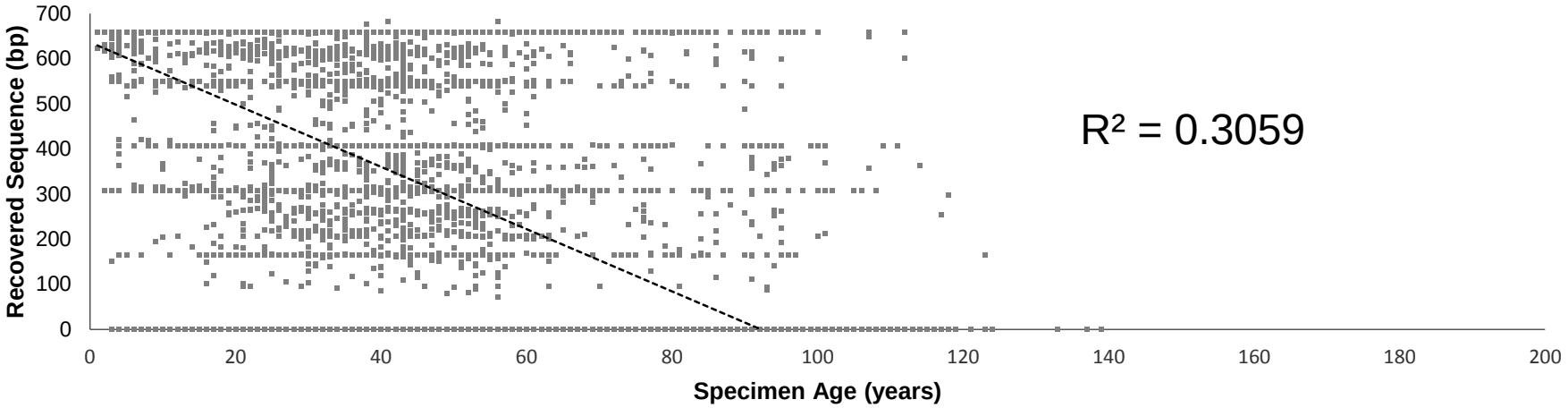
- Need to invent custom “de novo” assembly software
- De novo assembly is not smart
  - No alignments
  - Will create chimeric sequences if input data is not clean

NGS Platforms

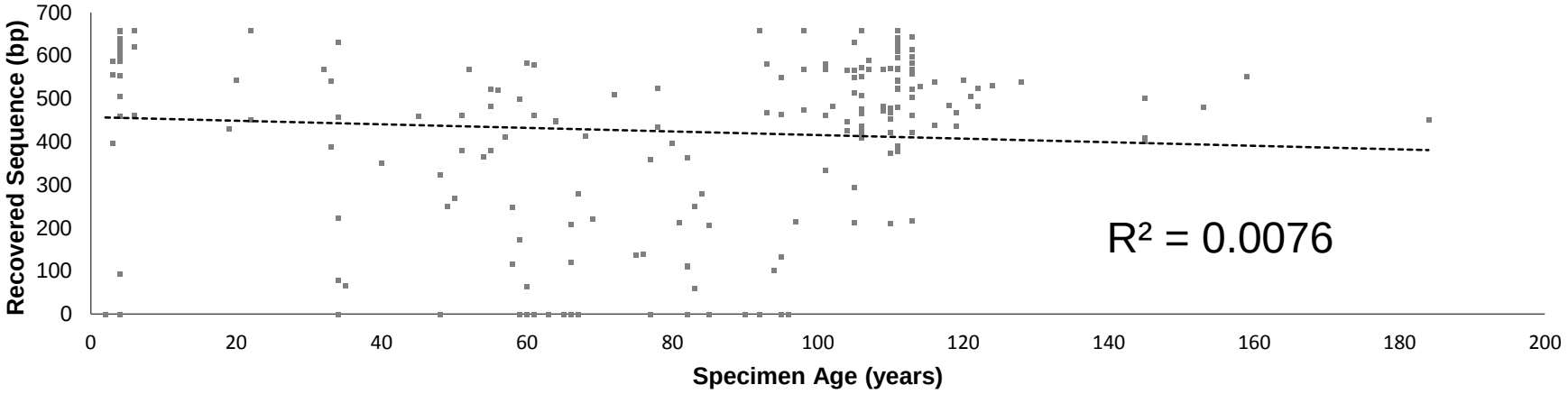


NGS Platforms

Sanger-based sequencing



NGS-based method



NGS Platforms

Taxon	Sanger recovery (%)	NGS-based recovery (%)
Moths (old)	7	87
Beetles (old)	13	67
Spiders (ethanol)	7	95
Spiders (formalin)	0	86
Reptiles & amphibians (formalin)	1	22
Mammals (formalin)	0	24

Summary

- **Full-length barcodes can be recovered from musuem specimens even when Sanger fails**
- **<u>Advantages</u>:**
  - **DNA damage due to age and/or preservation method can be circumvented with this method**
  - **Currently works across major insect and arachnid orders**
  - **Primer can be customized for any taxa**
    - **Mammals, fish, birds**
    - **Marine inverterates**
  - **Data analysis can be highly automated**
- **<u>Disadvantages</u>**
  - **Risk of chimeric sequences – sequences need to be validated**
  - **Throughput is mediocre – currently 95 samples per sequencing reaction but expected to increase with improved sequencing efficiency**

Resources

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 487–497 doi: 10.1111/1755-0998.12474

DNA barcodes from century-old type specimens using next-generation sequencing

SEAN W. J. PROSSER,* JEREMY R. DEWAARD,* SCOTT E. MILLER† and PAUL D. N. HEBERT*
**Biodiversity Institute of Ontario, University of Guelph, Guelph, ON, Canada, †National Museum of Natural History, Smithsonian Institution, Washington, DC, USA*

Sean Prosser - sprosser@uoguelph.ca