Genotyping By Sequencing (GBS) Method Overview

RJ Elshire, JC Glaubitz, Q Sun, JV Harriman
ES Buckler, and SE Mitchell

http://www.maizegenetics.net/

Topics Presented

- Background/Goals
- GBS lab protocol
- Illumina sequencing review
- GBS adapter system
- How GBS differs from RAD
- Modifying GBS for different species
- Workflow in our lab

Background

Genotyping by sequencing (GBS) in any large genome species requires reduction of genome complexity.

I. Target enrichment

- **Long range PCR of specific genes or genomic subsets**
- **Molecular inversion probes**
- **Sequence capture approaches hybridization-based (microarrays)**

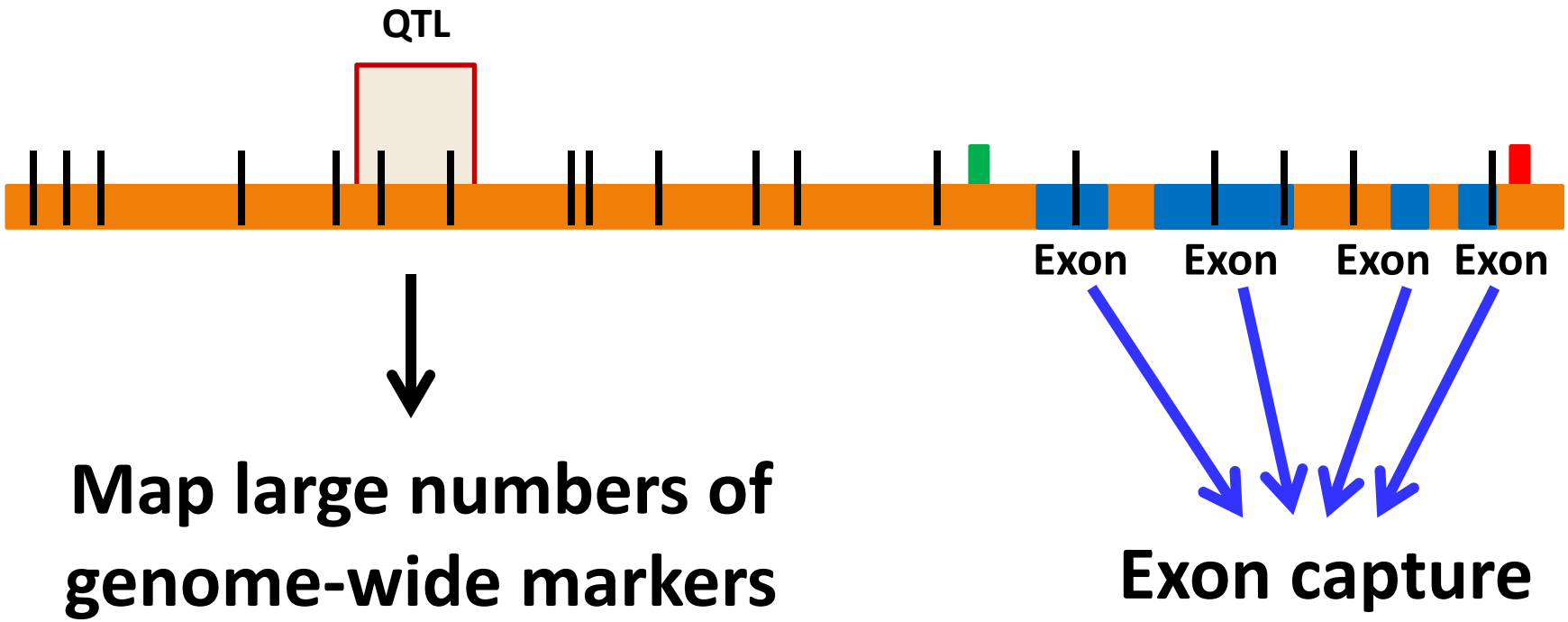
II. Restriction Enzymes (REs)

Technically less challenging

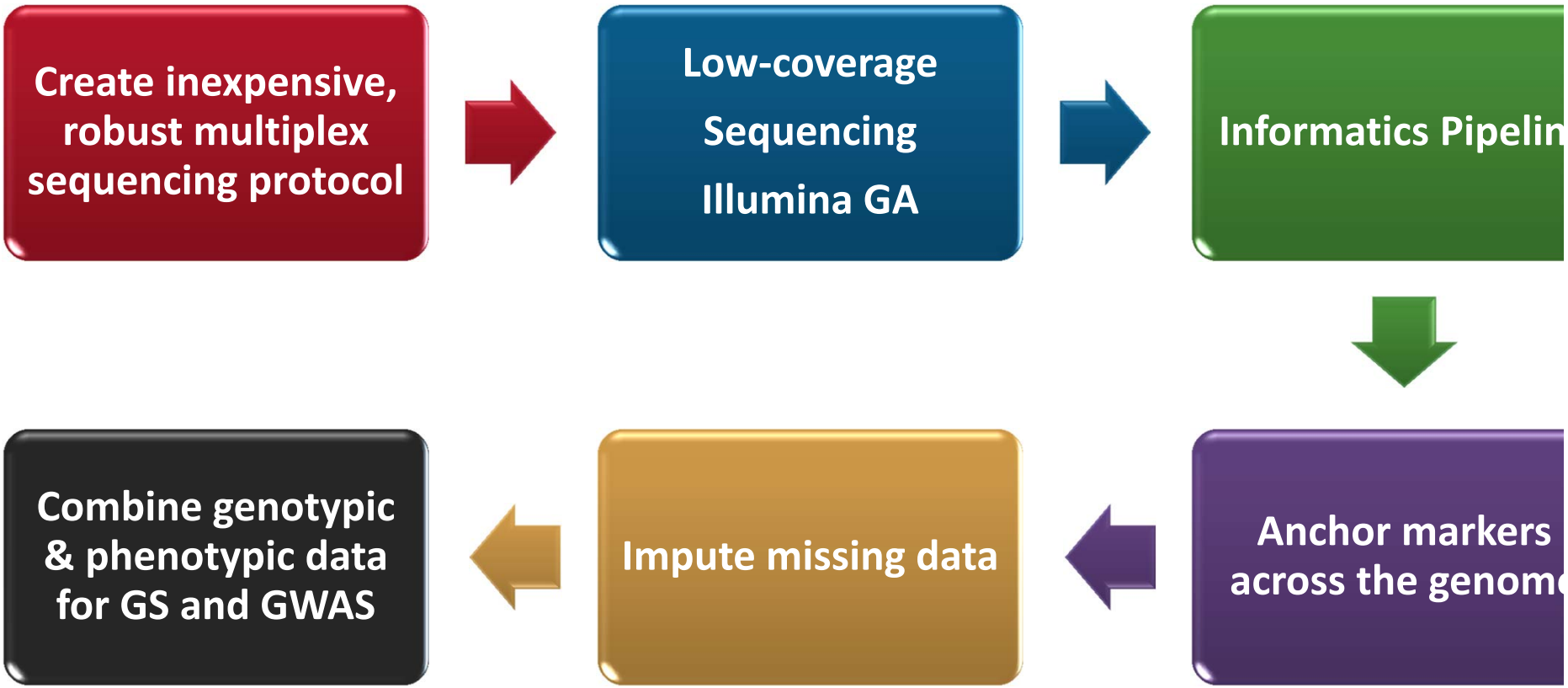
- **Methylation sensitive REs filter out repetitive genomic fraction**

QTL are often located in non-coding regions

Vgt1, *Tb*, *B* regulatory regions 60-150kb from gene



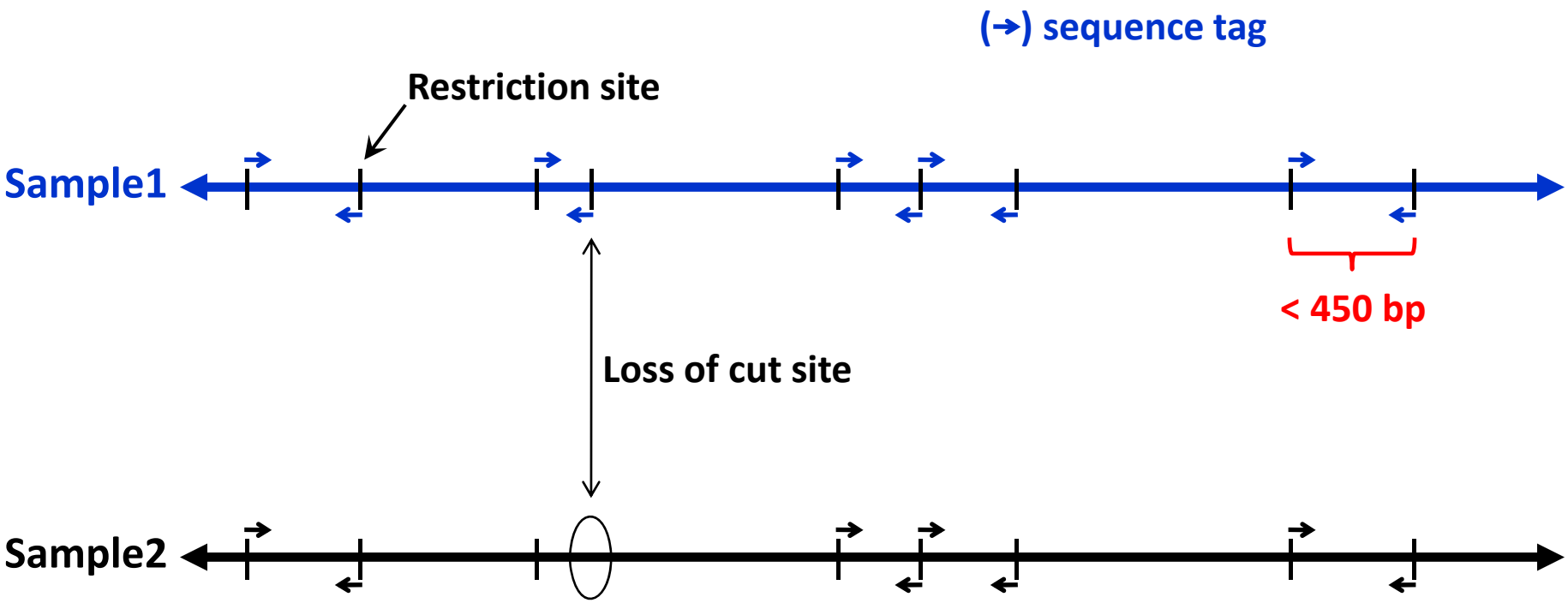
Our goal is to create a public genotyping/informatics platform based on next-generation sequencing



Open Source

- **Method available for anyone to use / modify.**
- **Analysis pipeline details and code are public.**
- **Promote dataset compatibility.**
- **Method published in *PLoS ONE* to promote accessibility.**
- **Genotype calls publically available.**

Overview of Genotyping by Sequencing (GBS)



- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both SNPs and presence/absence markers

GBS is a simple, highly multiplexed system for constructing libraries for next-gen sequencing

- Reduced sample handling
- Few PCR & purification steps
- No DNA size fractionation
- Efficient barcoding system
- Simultaneous marker discovery & genotyping
- Scales very well

GBS 96-plex Protocol

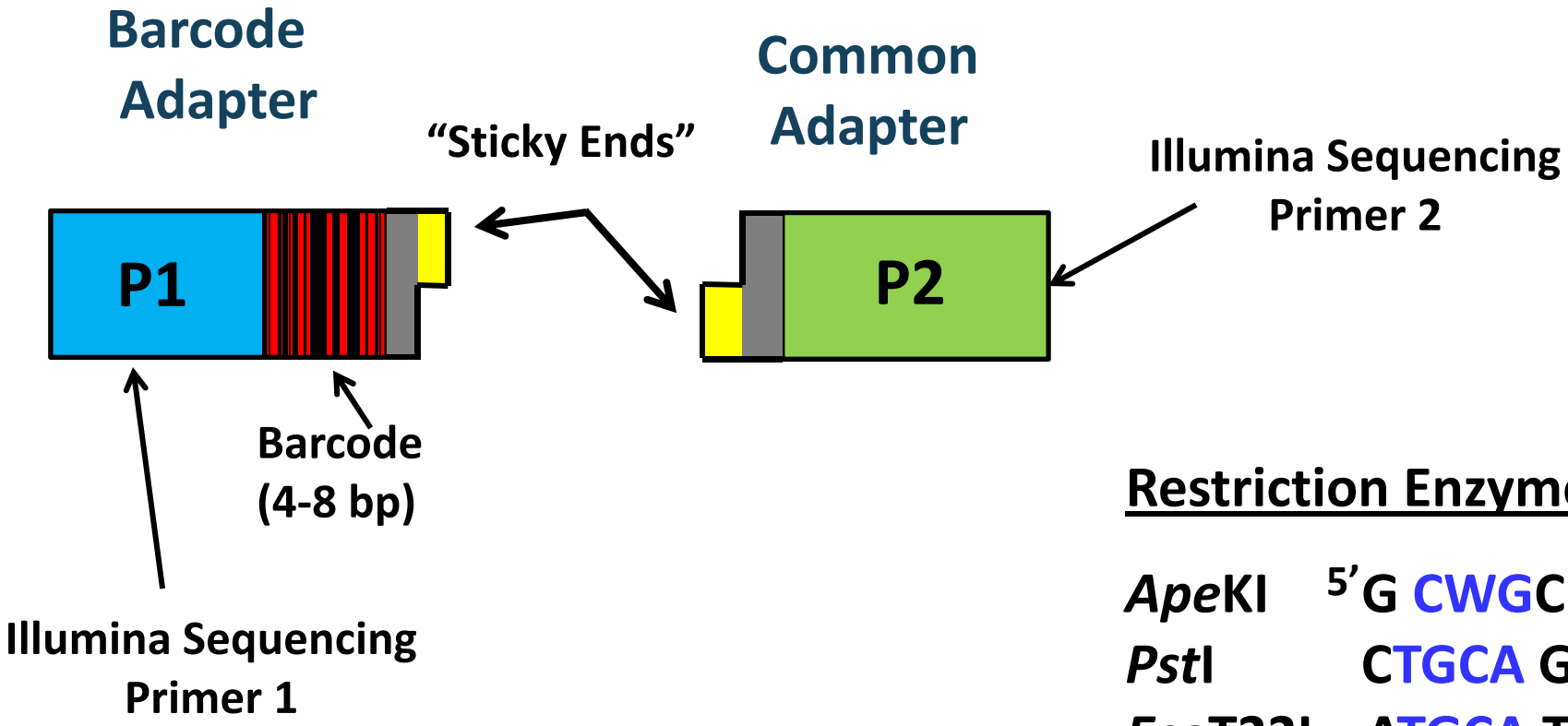
(http://www.maizegenetics.net/gbs-overview)

Now at 384-plex

1. Plate DNA & adapter pair

2. Digest DNA with RE
3. Ligate adapters

GBS Adapters and Enzymes

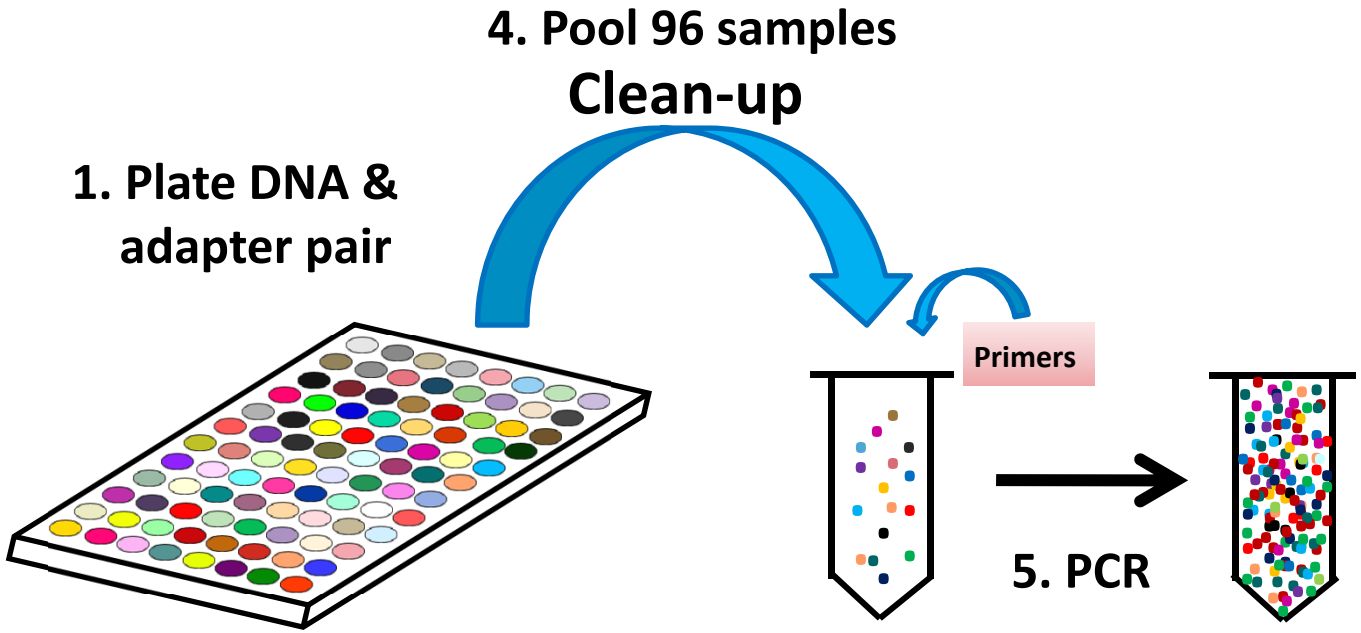


Restriction Enzymes

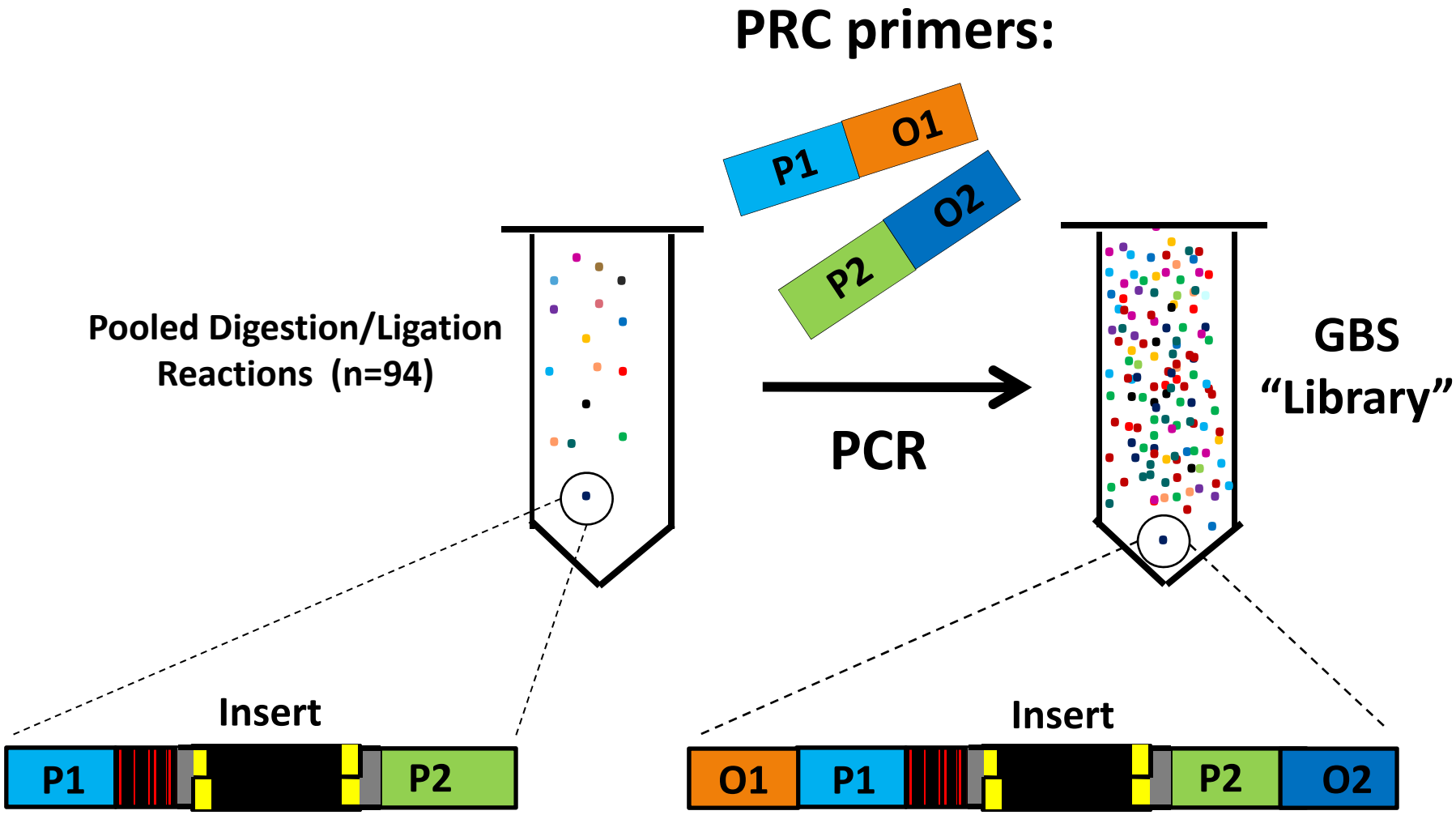
*Ape*KI	5'G CWGC3'
*Pst*I	CTGCA G
*Eco*T22I	ATGCA T

GBS 96-plex Protocol

(http://www.maizegenetics.net/gbs-overview)

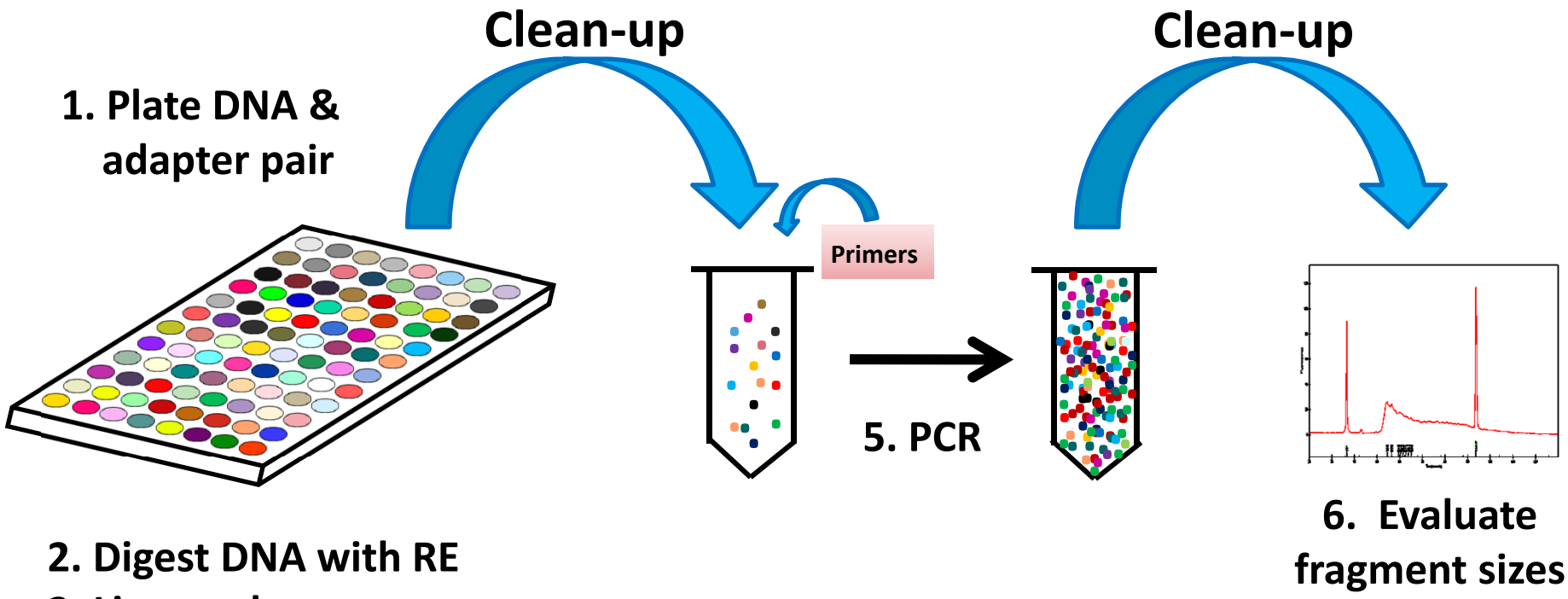


2. Digest DNA with RE
3. Ligate adapters


PRC primers:
P1
O1
P2
O2
Pooled Digestion/Ligation
Reactions (n=94)
PCR
GBS
“Library”
Insert
P1
P2
Insert
O1
P1
P2
O2

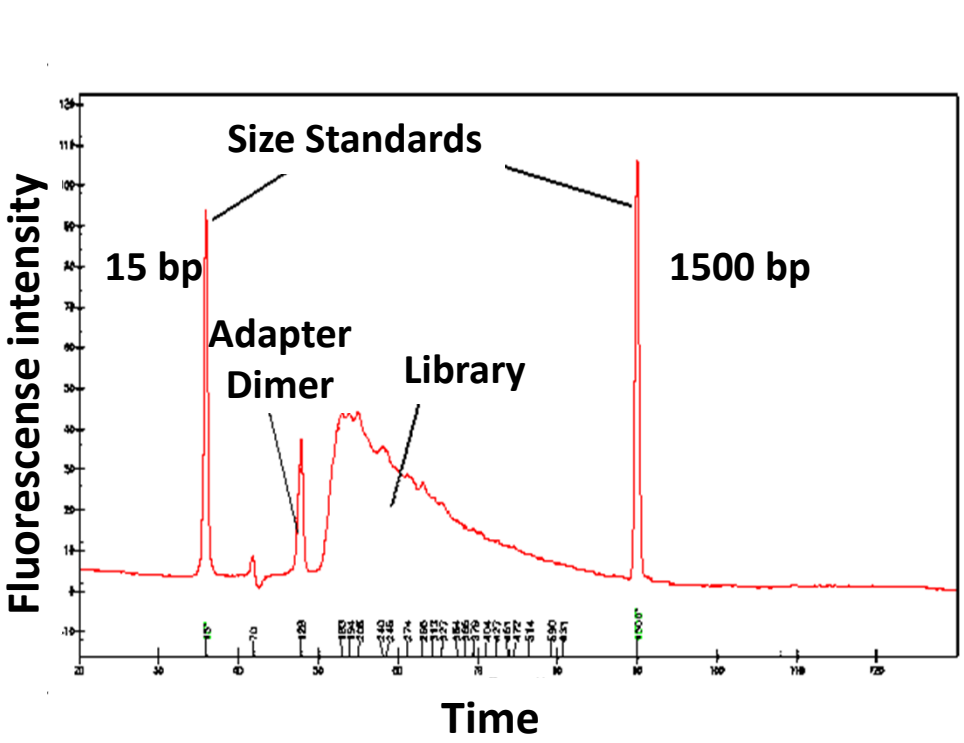
GBS 96-plex Protocol

(http://www.maizegenetics.net/gbs-overview)

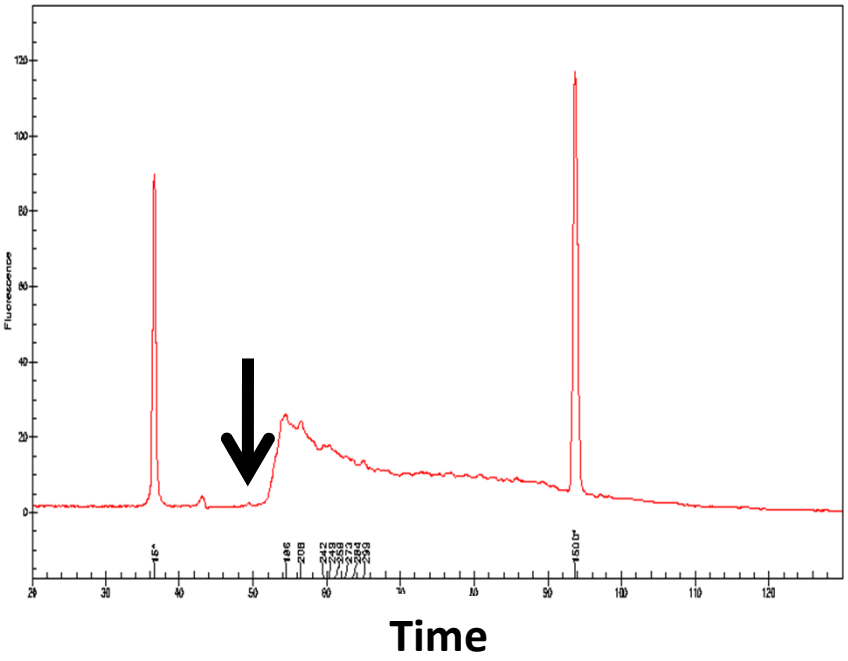


Perform Titration to Minimize Adapter Dimers Before Sequencing

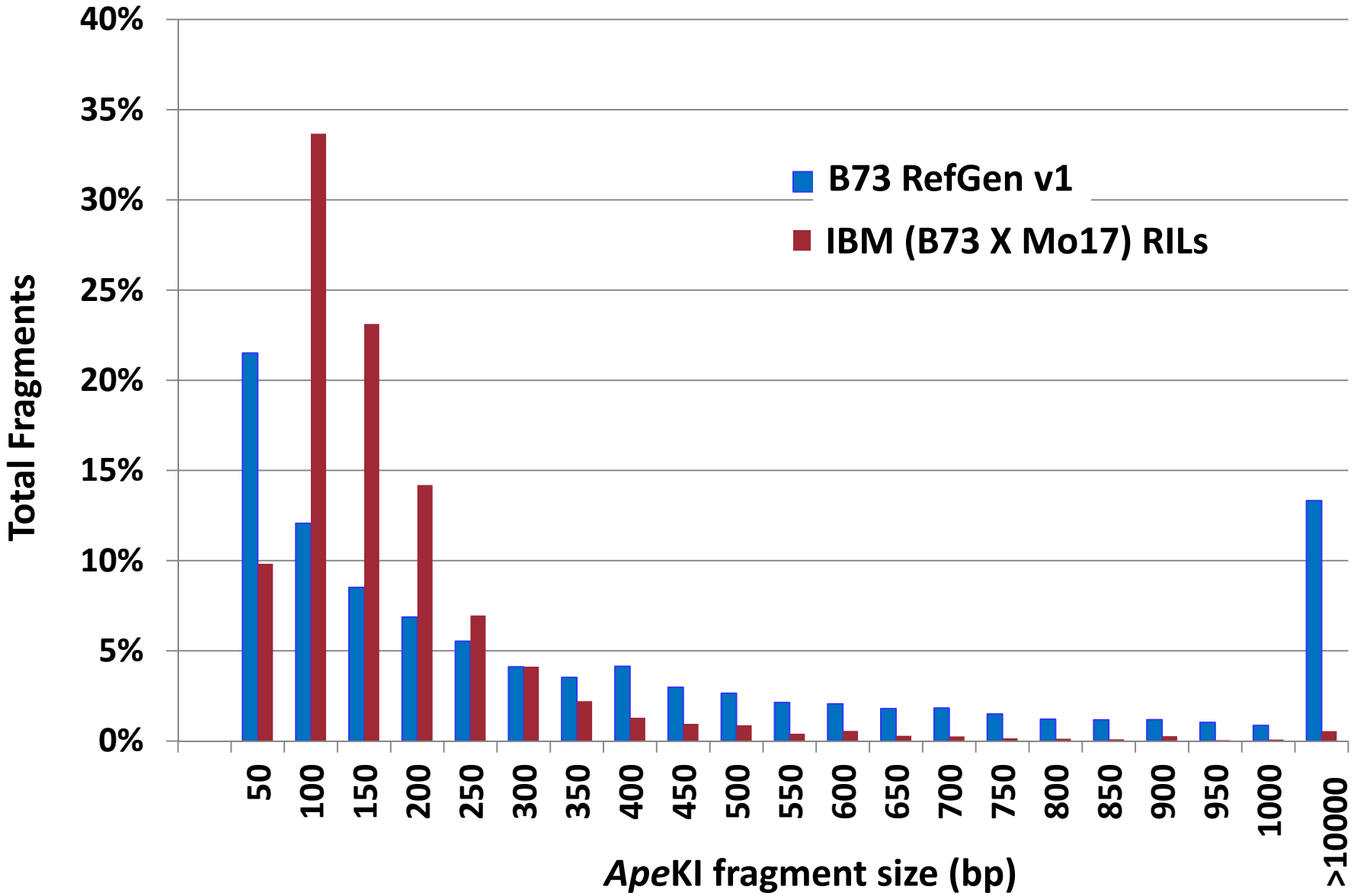
NOTE: Done once with a small number of samples.
Adapter dimers constitute only 0.05% of raw sequence reads



Optimal adapter amount

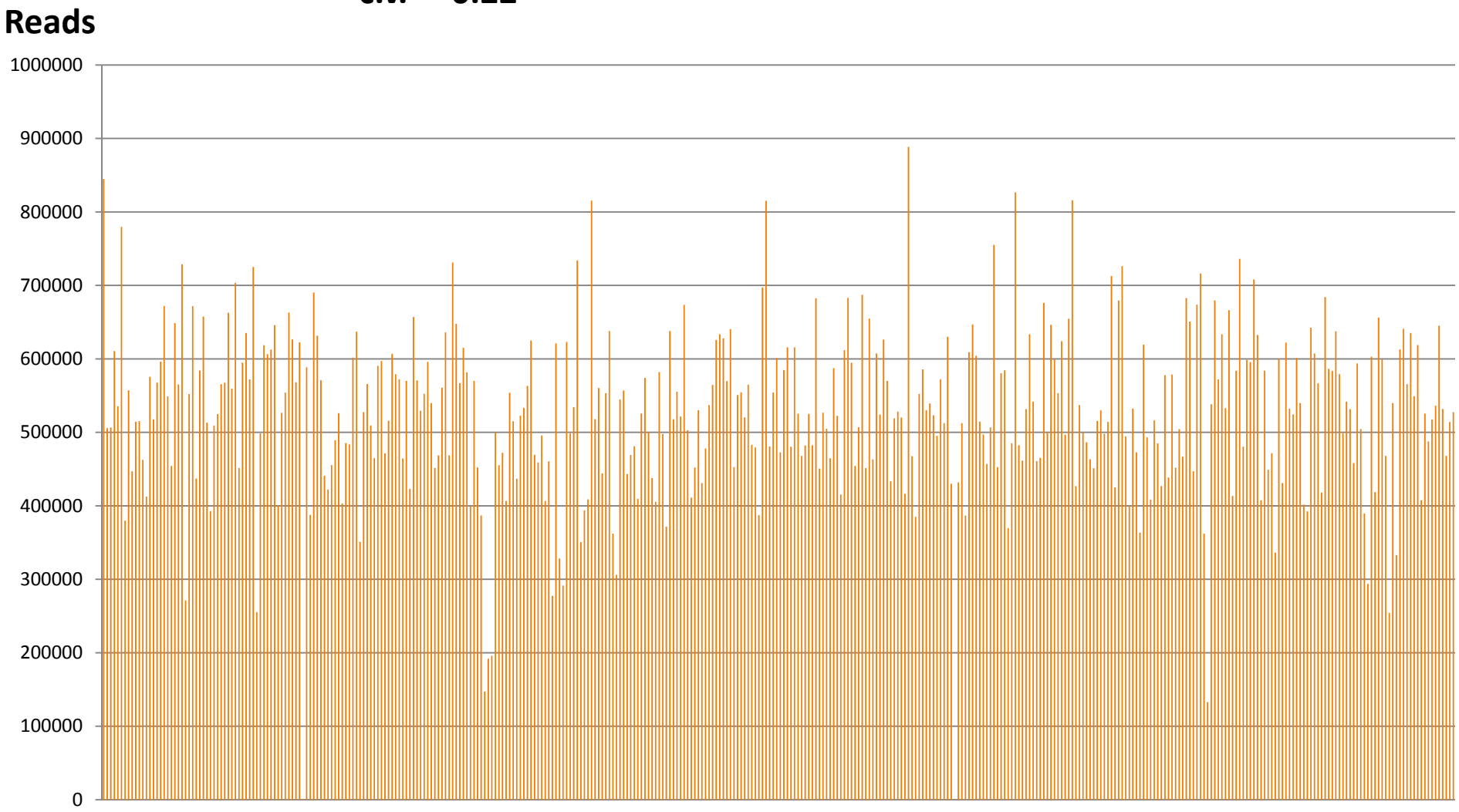


Small Fragments are Enriched in GBS Libraries

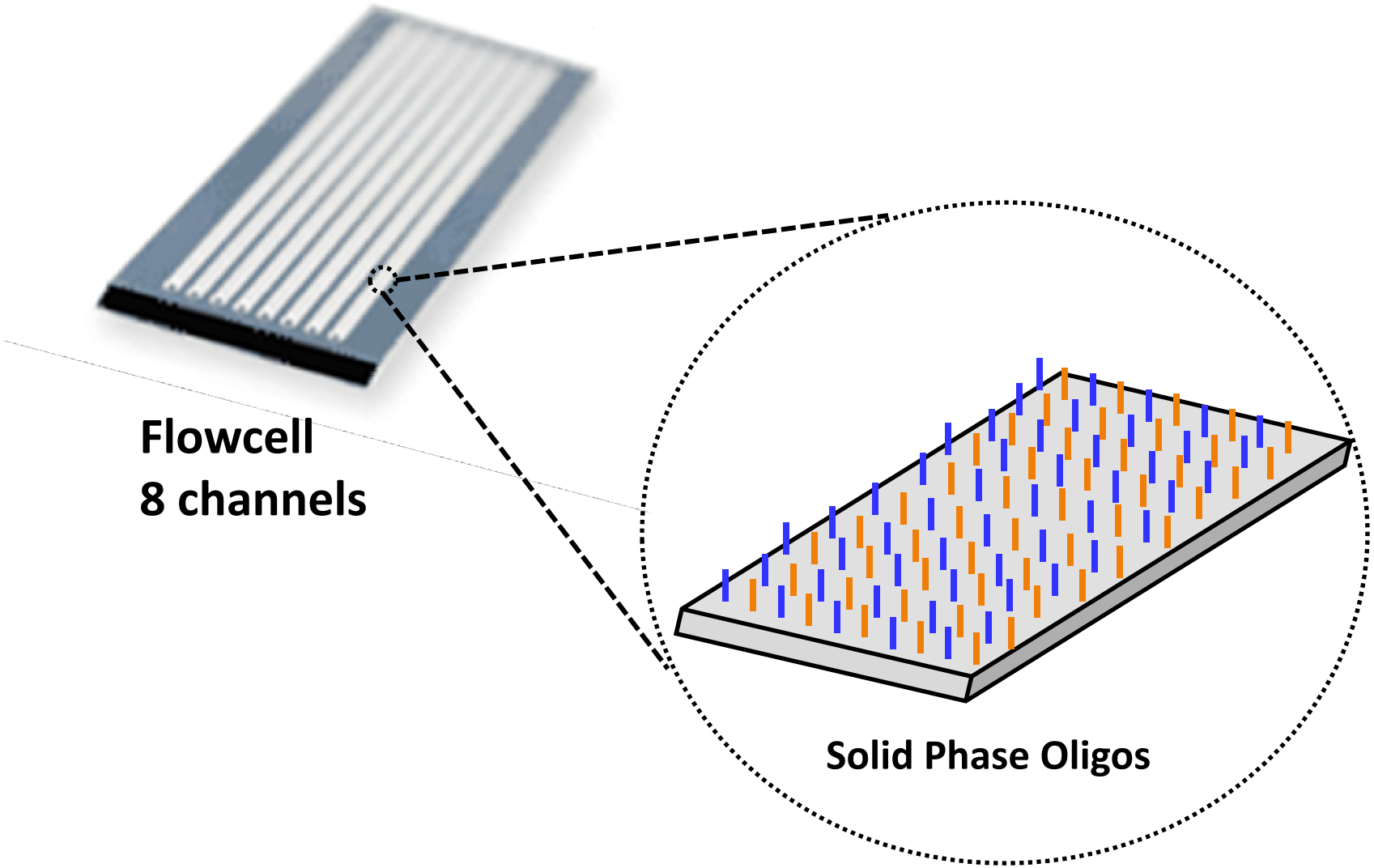


384-plex GBS Results for Maize

Mean read count per line = 528,000
c.v. = 0.22



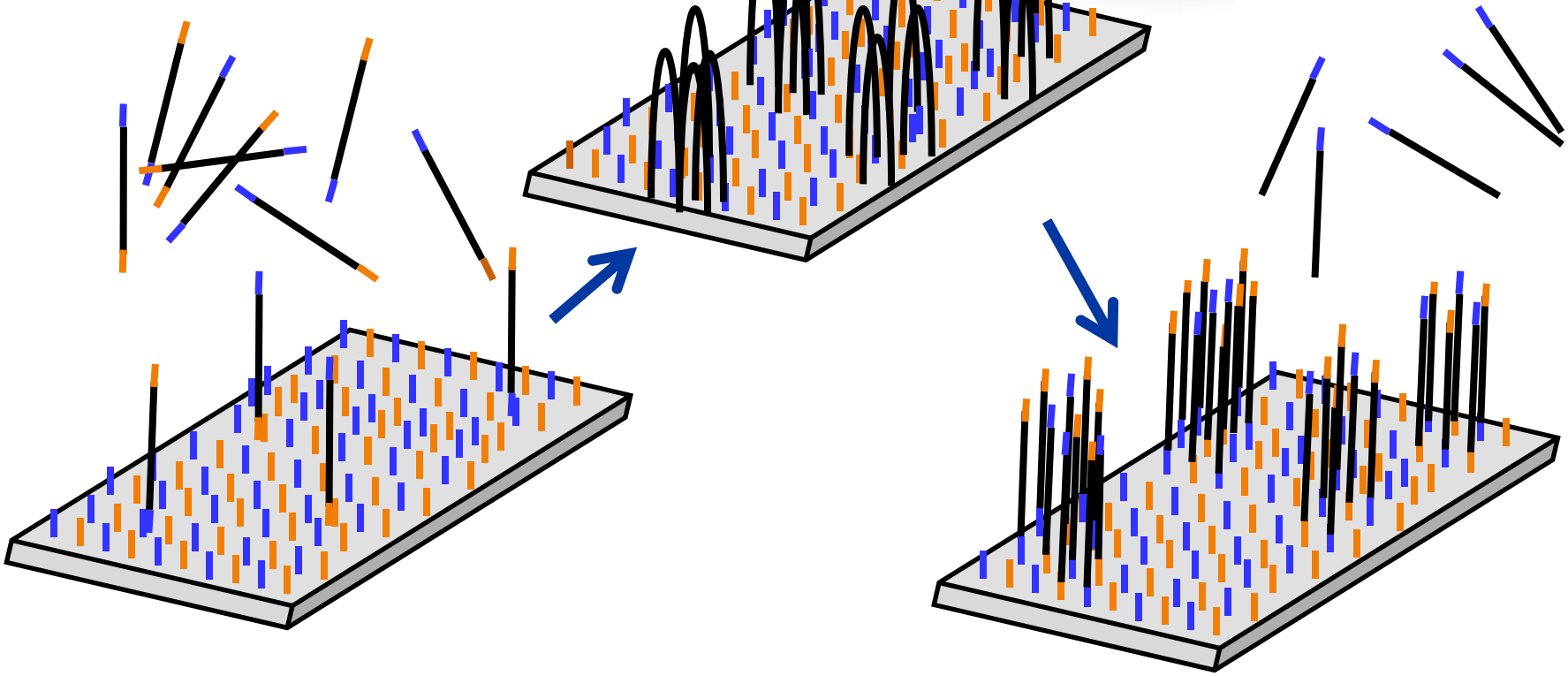
Illumina Sequencing by Synthesis Review



Cluster Formation Amplifies Sequencing Signal

Bridge Amplification

Denatured “Library”



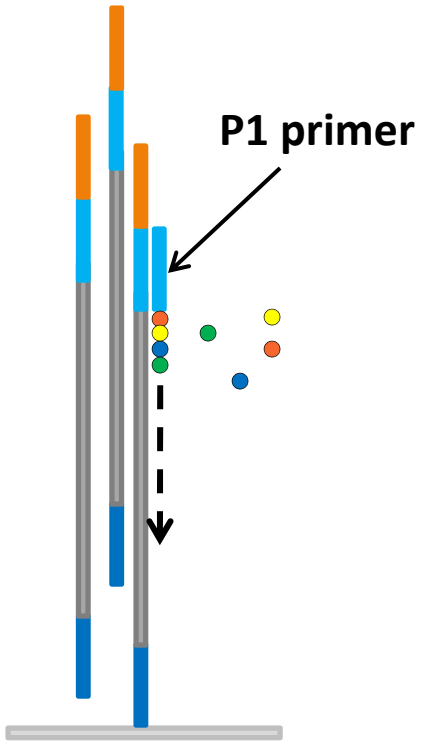
Flow cell with bound oligos

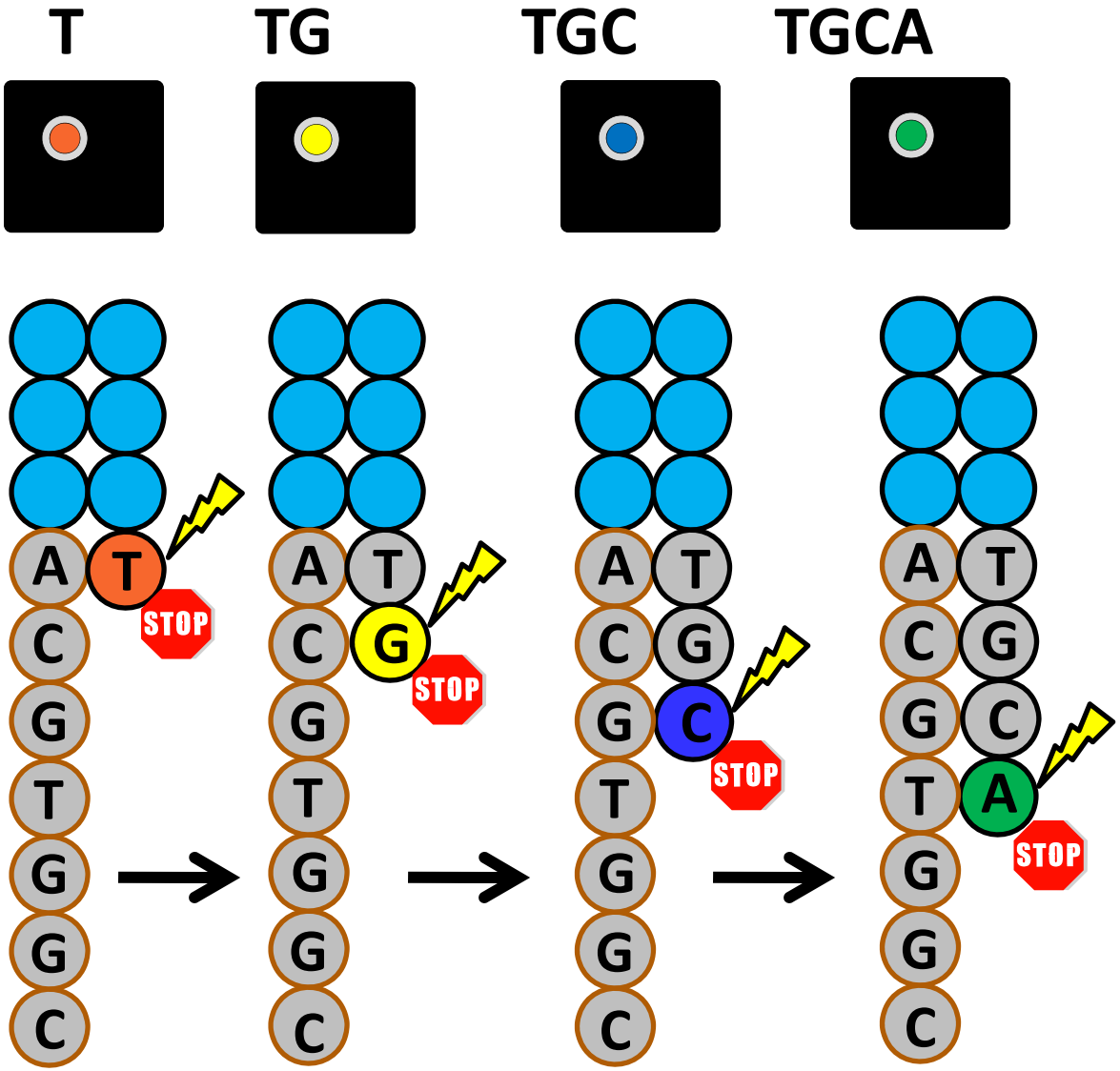
Linearization

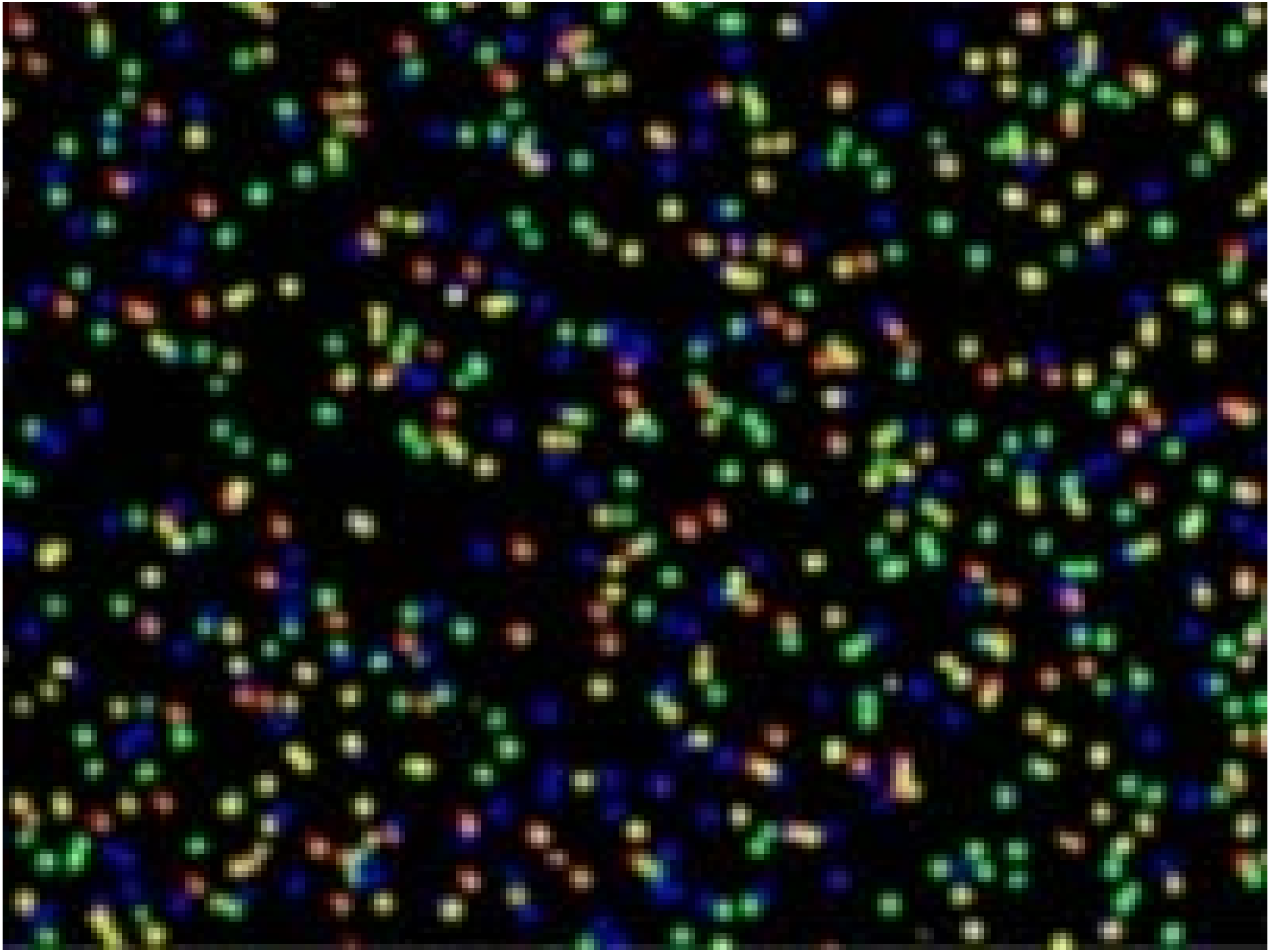
Sequencing by Synthesis

HiSeq 2000

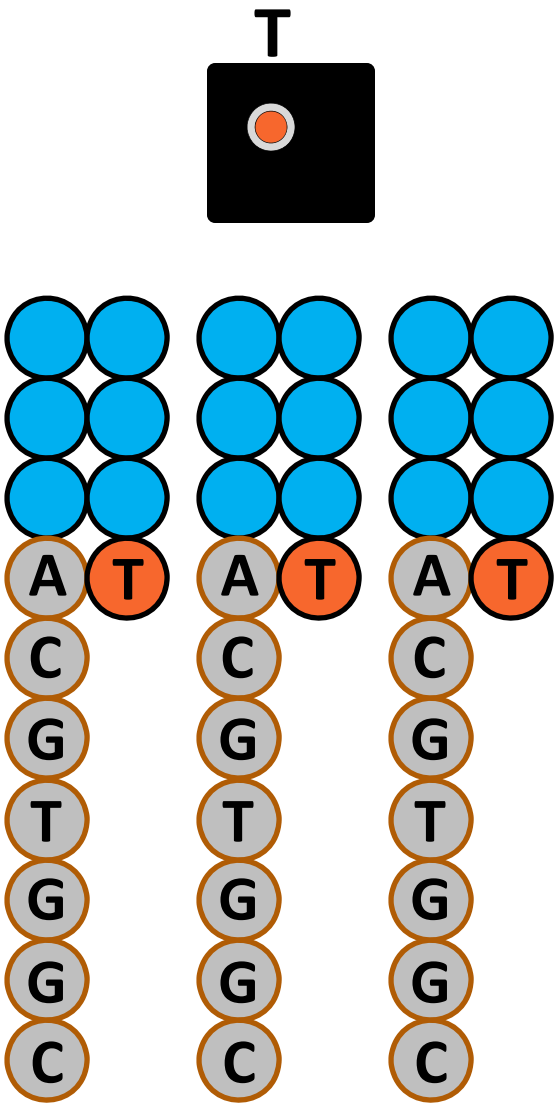
Flowcell



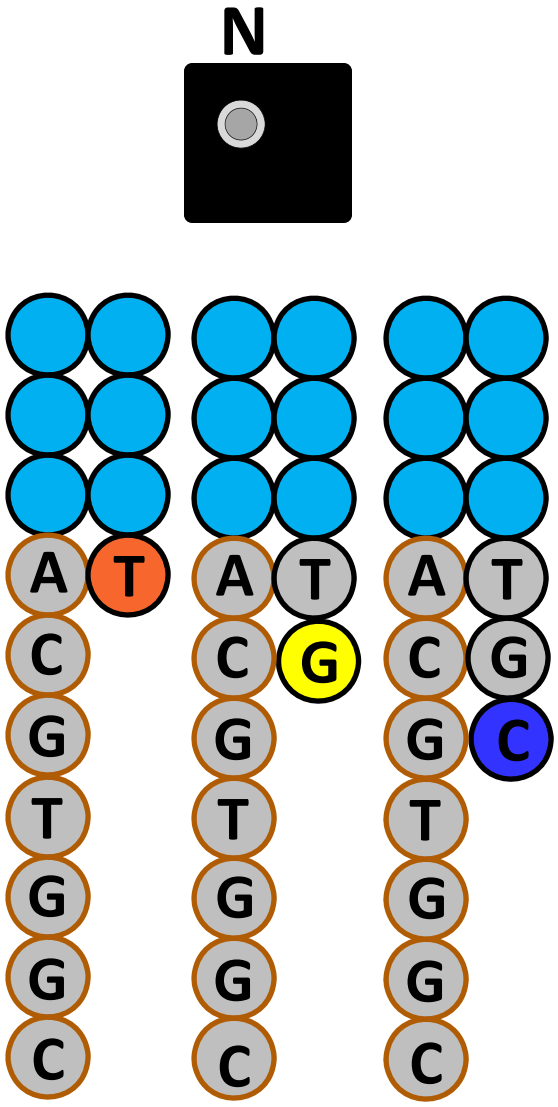




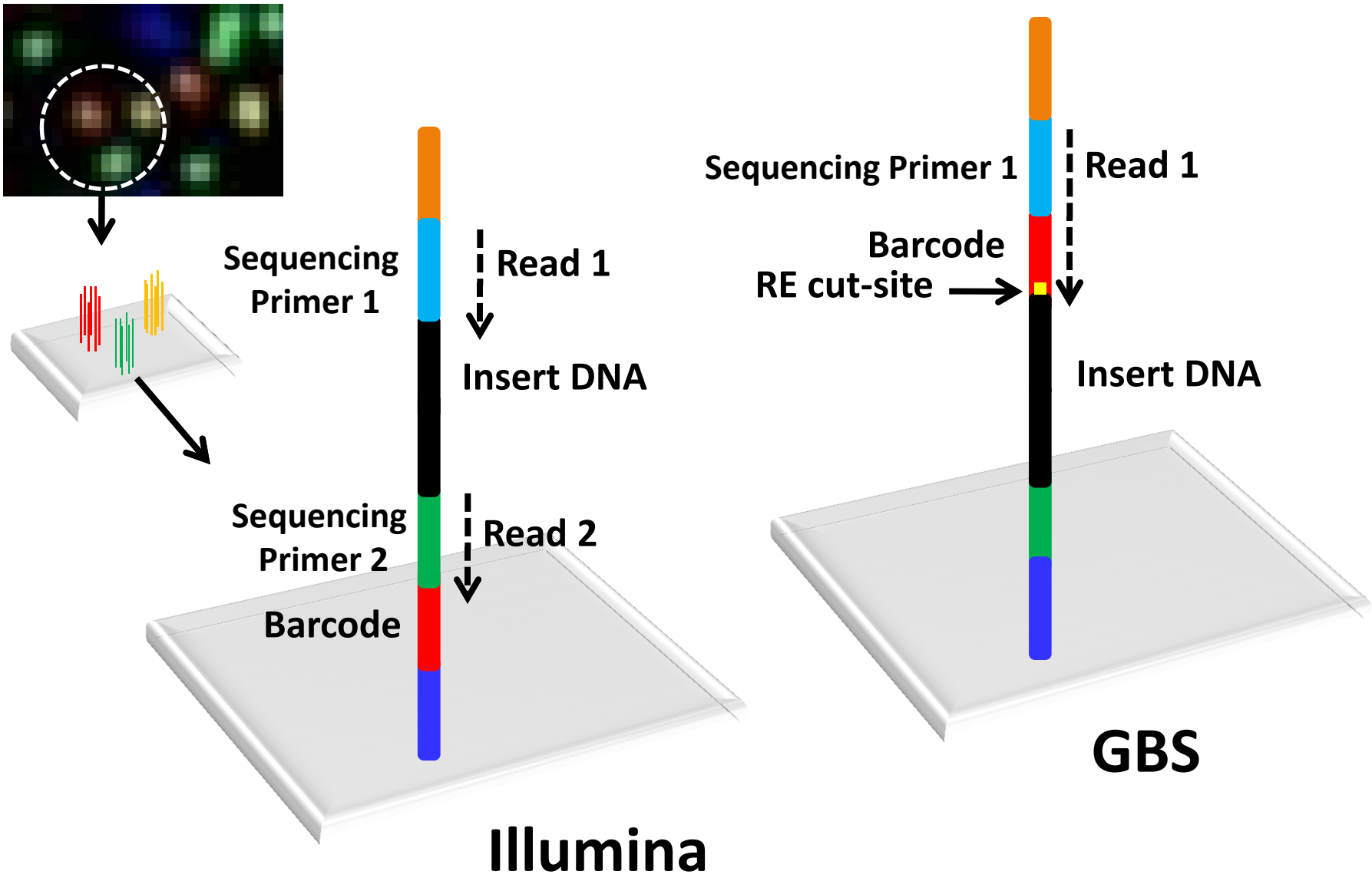
“In Phase”



“Out of Phase”



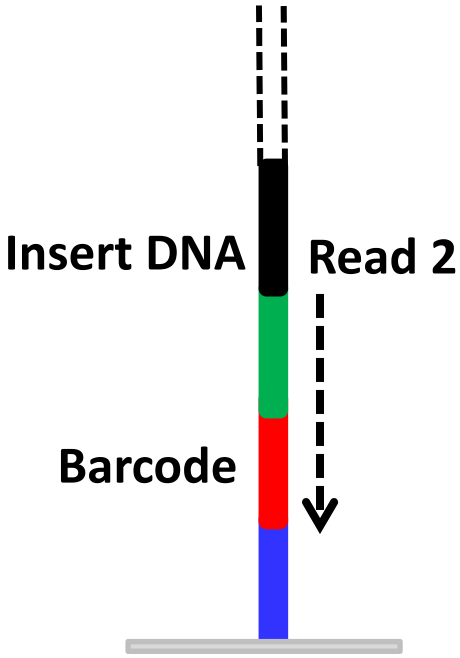
GBS captures barcode and insert DNA sequence in single read



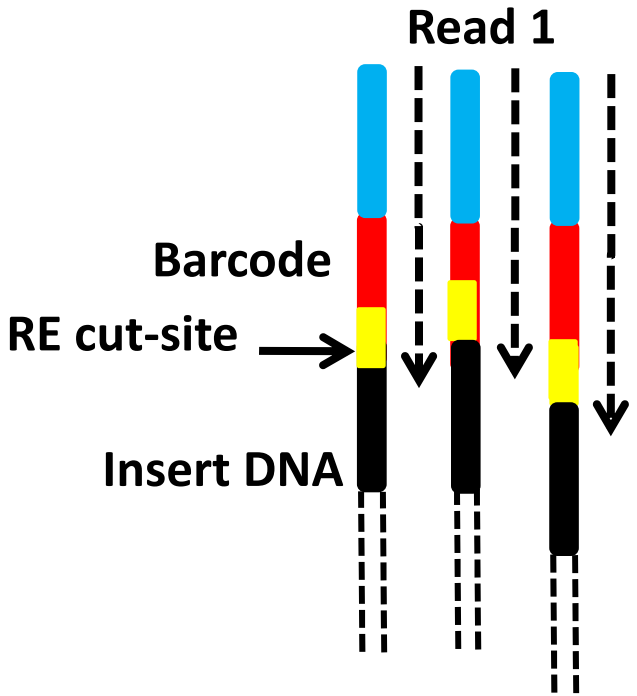
Variable Length GBS Barcodes Solves Sequence Phasing Issues

- First 12 nt used to calculate phasing.
- Algorithm assumes random nt distribution.
- Incorrect phasing causes incorrect base calls.

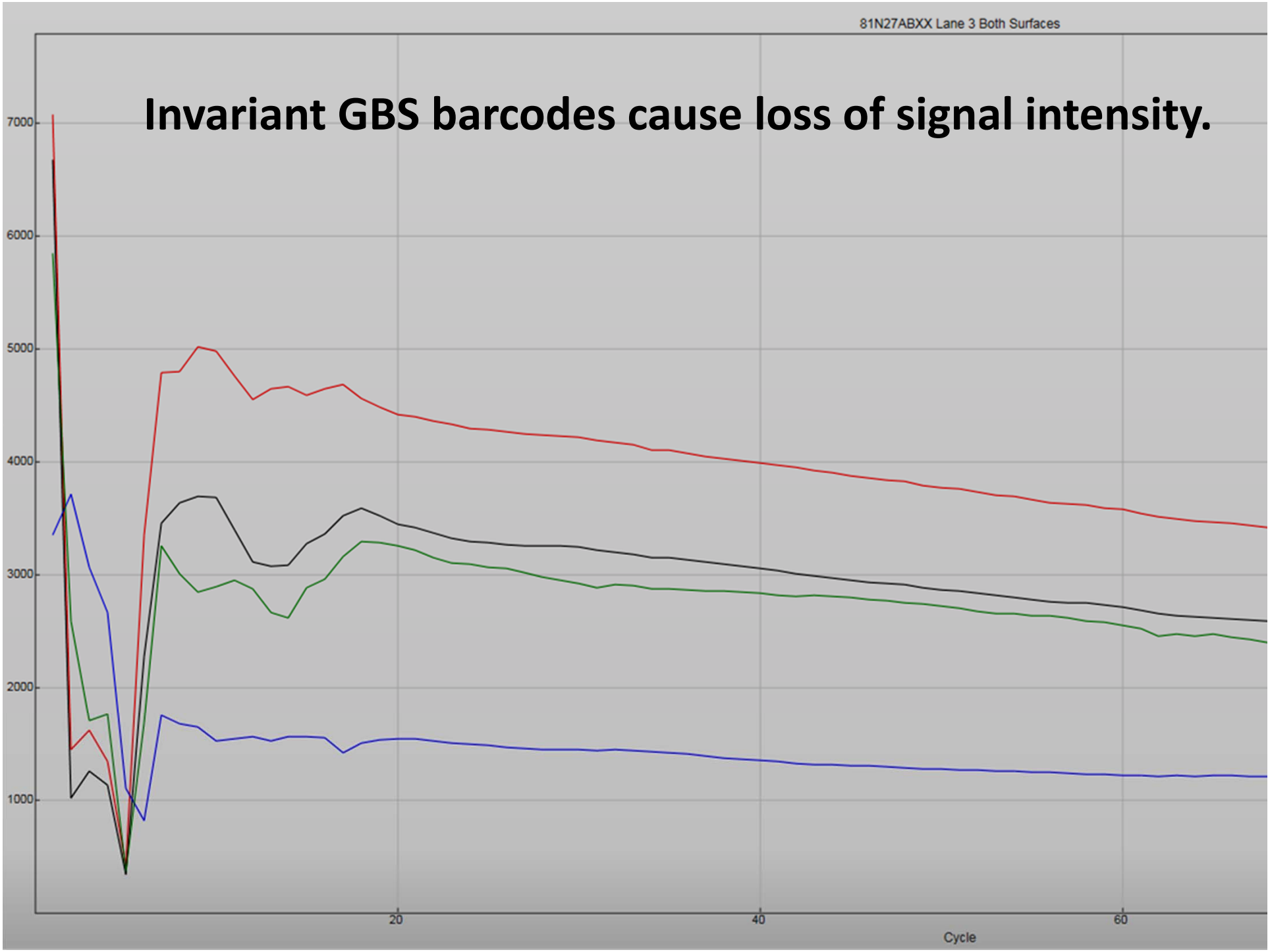
- Good design and modulating the RE cut site position with variable length barcodes produces even nt distribution.



Illumina



GBS

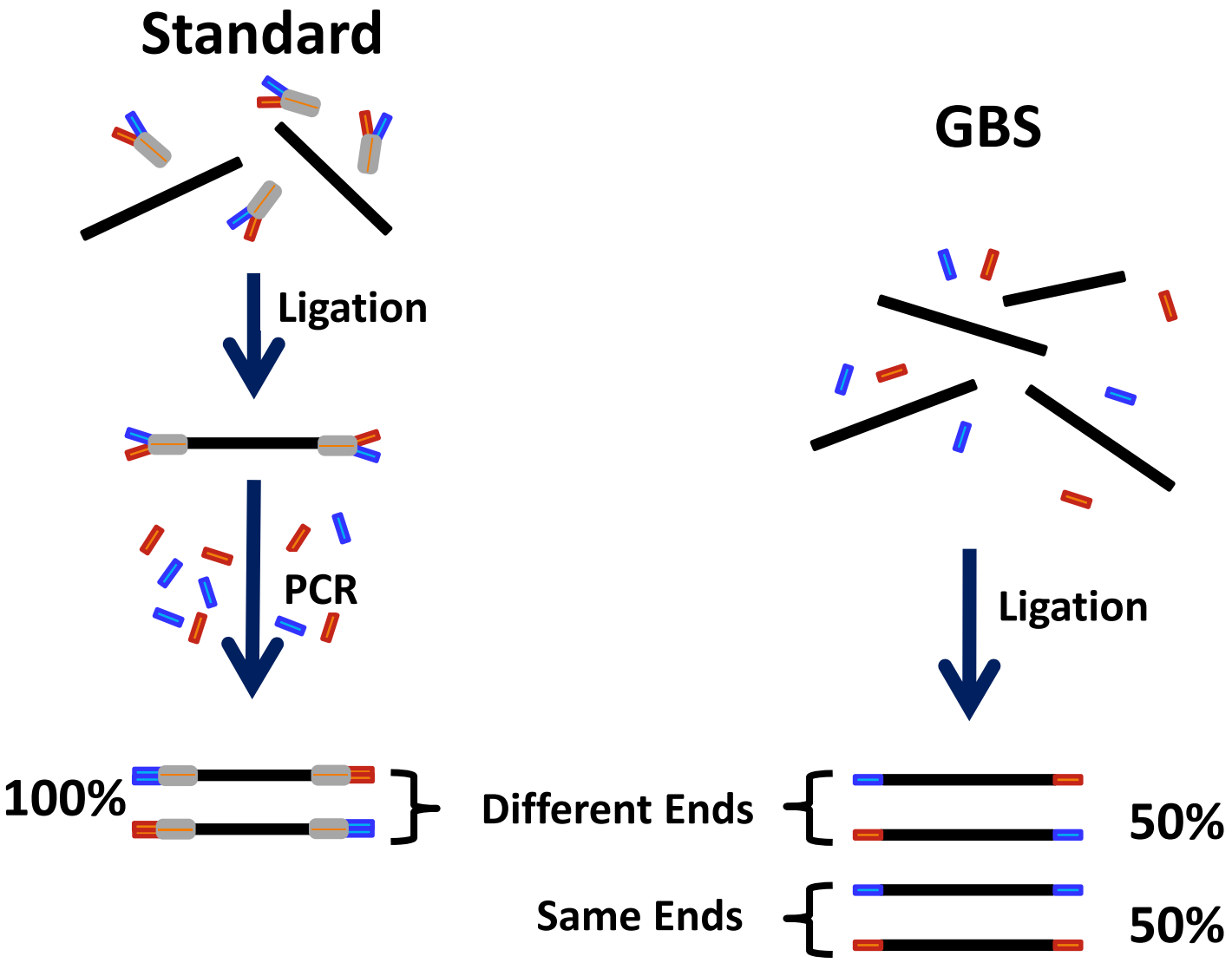

81N27ABXX Lane 3 Both Surfaces
Invariant GBS barcodes cause loss of signal intensity.
7000
6000
5000
4000
3000
2000
1000
20
40
60
Cycle


81N27ABXX Lane 1 Both Surfaces
Successful GBS sequencing run.
All Bases Intensity
8000
7000
6000
5000
4000
3000
2000
1000
20
40
60
Cycle

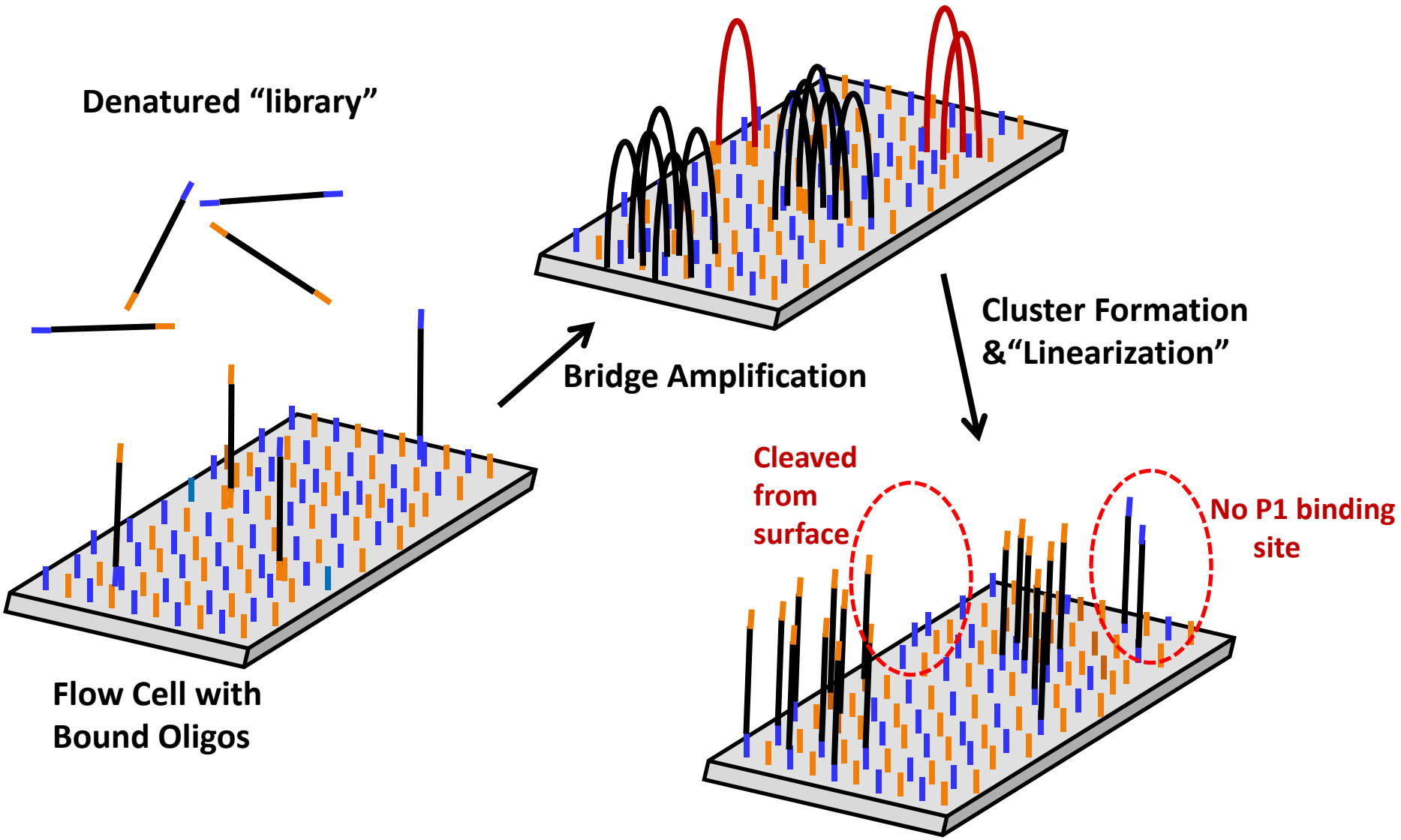
Most significant GBS technical issue?

DNA quality

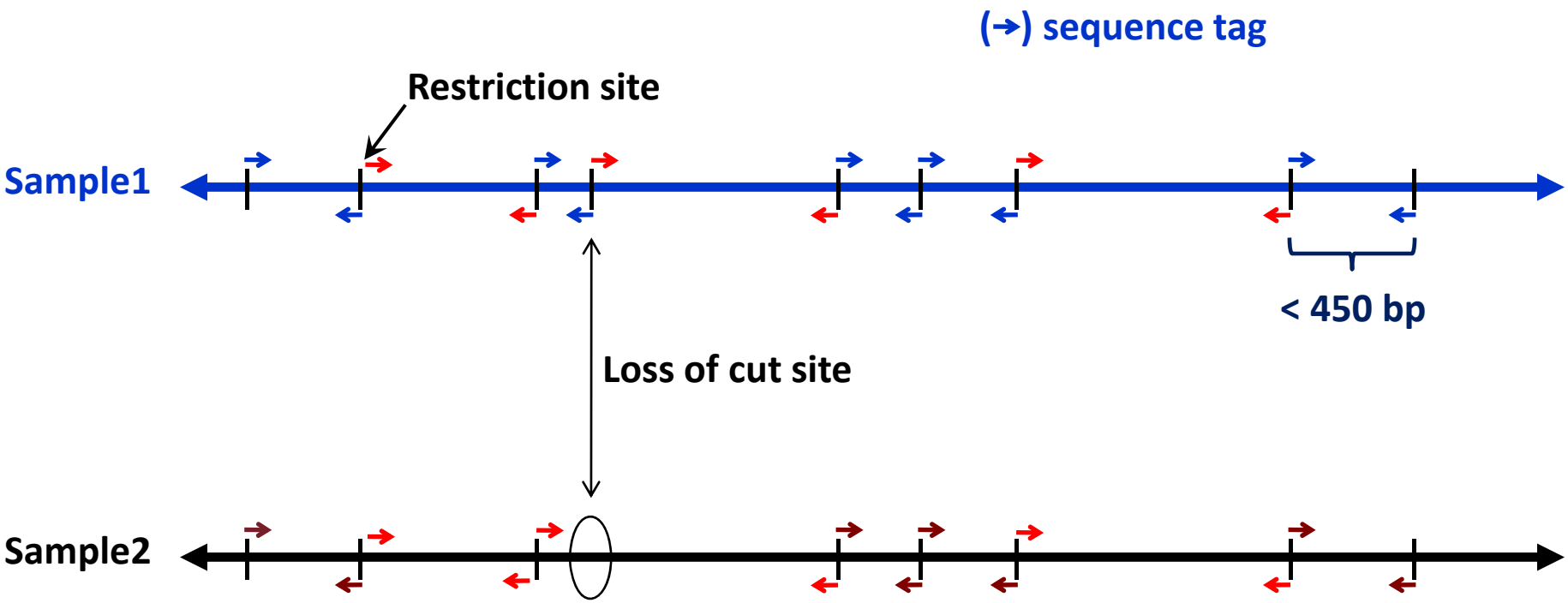
GBS does not use standard “Y” adapters



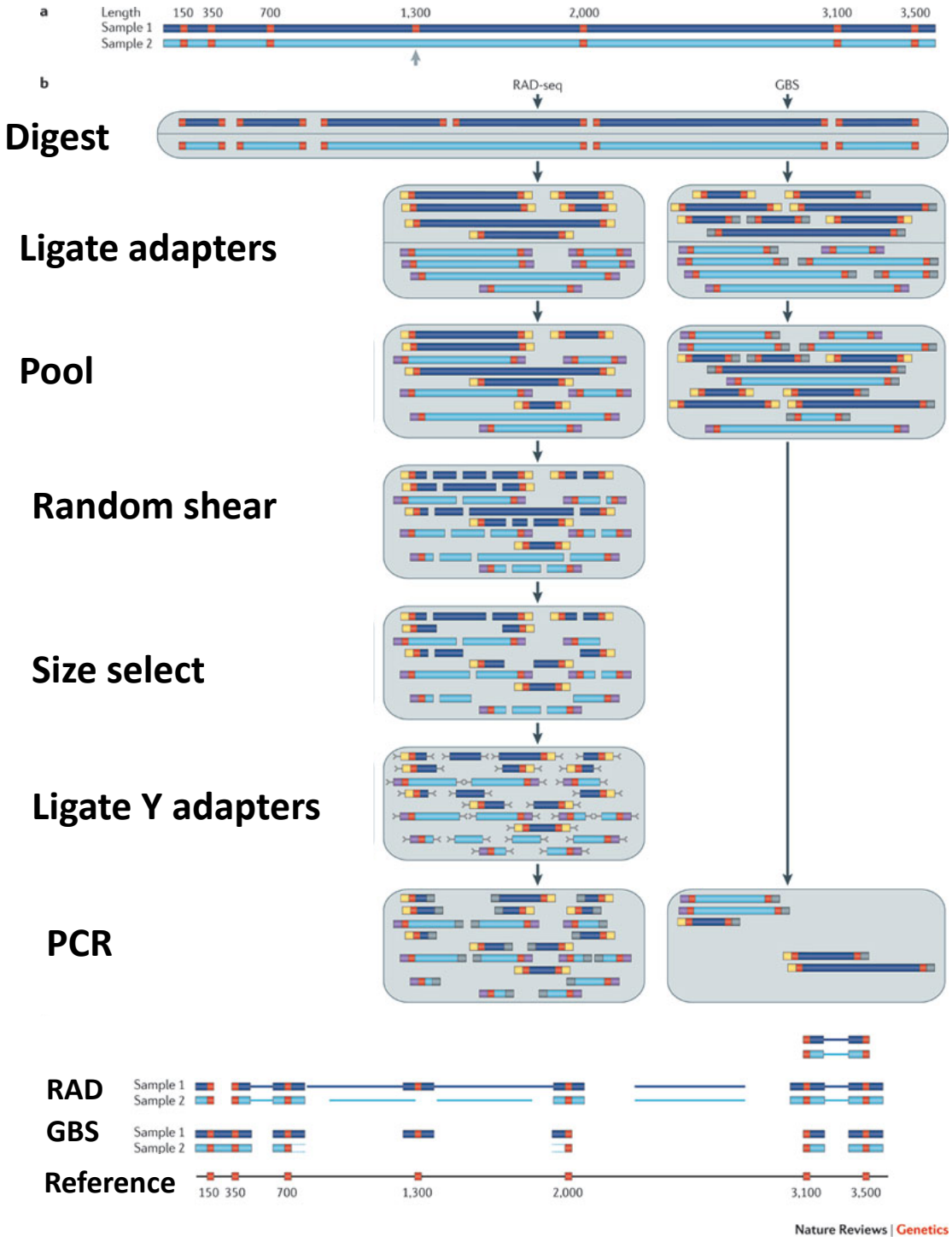
Same-ended Fragments Do Not Form Clusters



GBS vs. RAD



- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both SNPs and presence/absence markers



Davey et al. 2011

Modifying GBS

Considerations for using GBS with new species and / or different enzymes.

Why Modify the GBS Protocol?

- More markers
- Fewer markers (deeper sequence coverage per locus)
- Increase multiplexing
- More genome appropriate (avoid more repetitive DNA classes)
- Other novel applications (i.e., bisulfite sequencing).

Genome Sampling Strategies Vary by Species

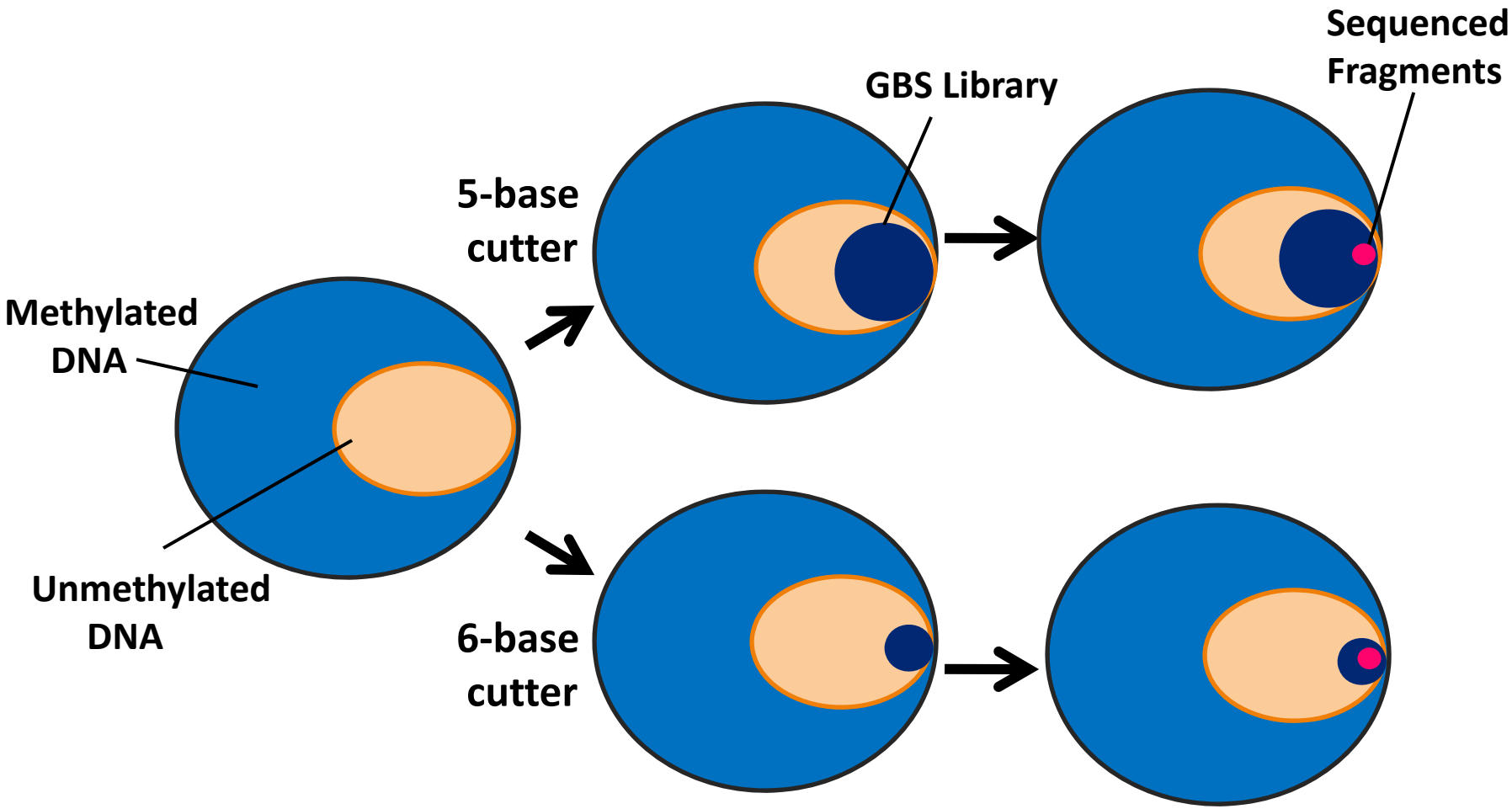
Dependent on Factors that Affect Diversity:

- **Mating System**
  (Outcrosser, inbreeder, clonal?)
- **Ploidy**
  (Haploid, diploid, auto- or allopolyploid?)
- **Geographical Distribution**
  (Island population, cosmopolitan?)

Other Factors

- **Genome size**
  - The size of the genome has some bearing on the size of the fragment pool.
  - Amount of repetitive DNA directly correlated with genome size.
- **Genome composition**
  - The composition of the genome can affect the frequency and distribution of the cut sites.

Sampling large genomes with methylation-sensitive restriction enzymes.

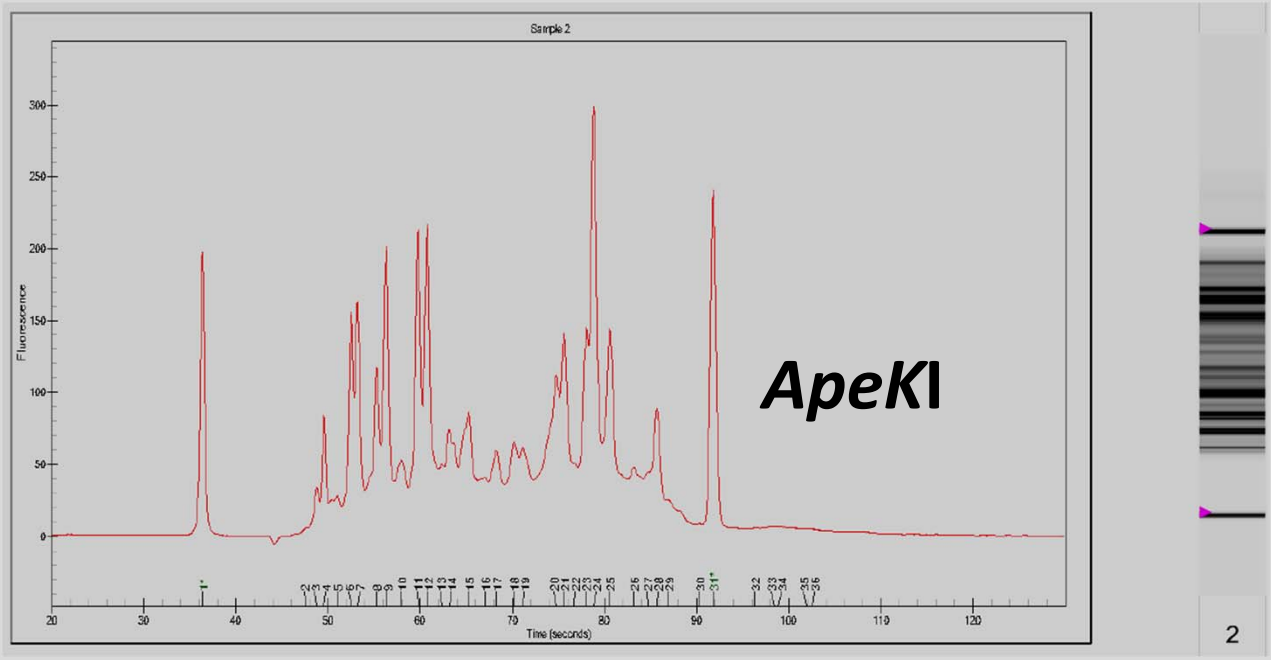


Optimizing GBS in New Species

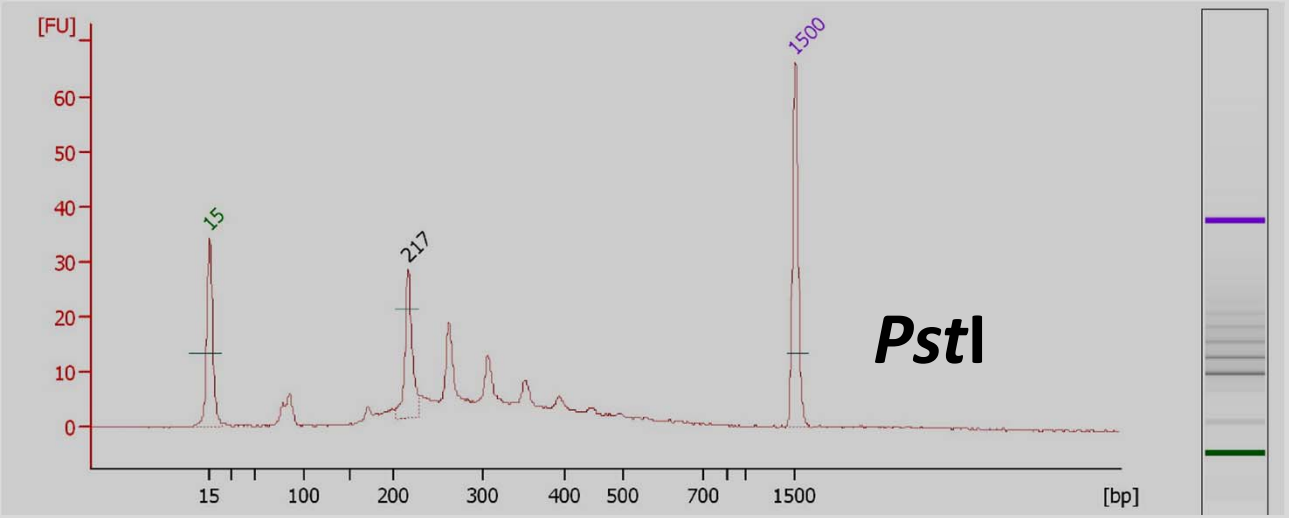
Choosing Appropriate Restriction Enzymes

“Life is like a bowl of chocolates”.

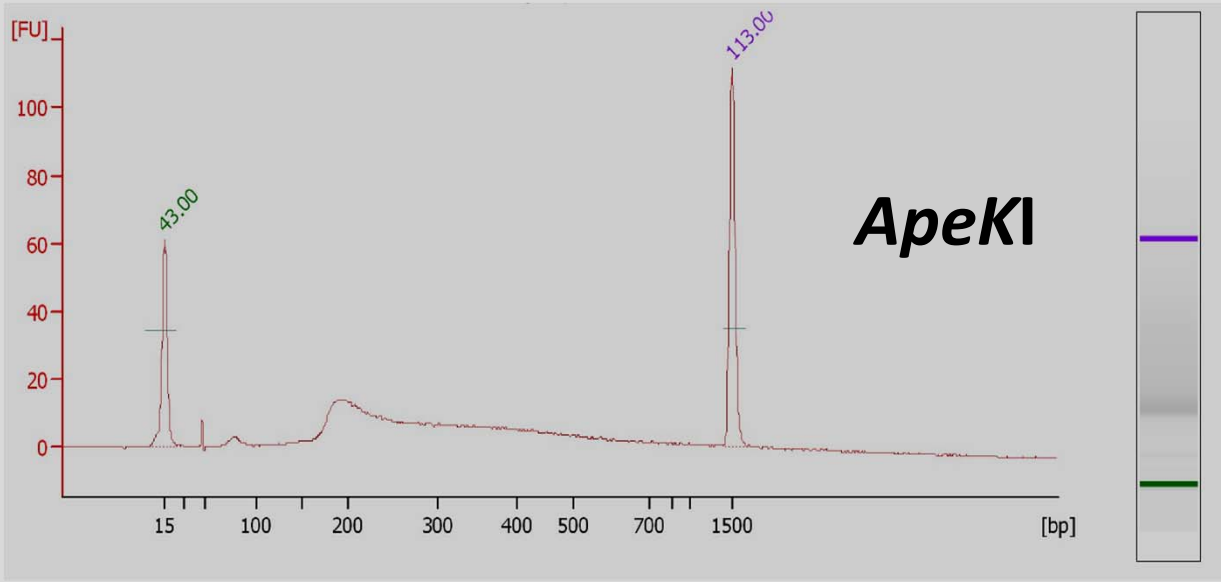
Giant Squid

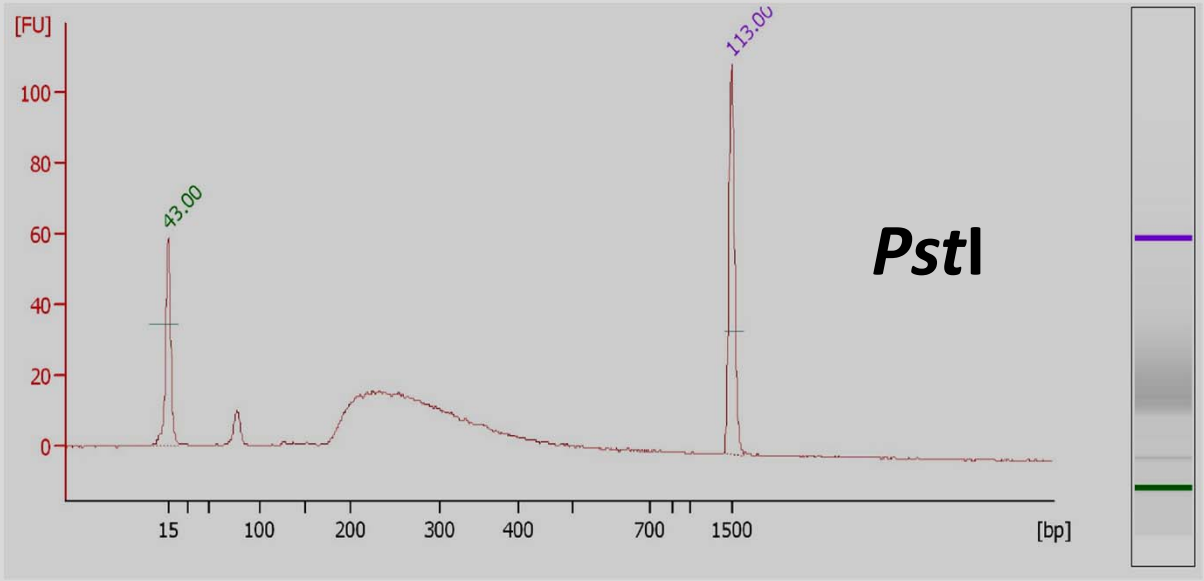


Goose

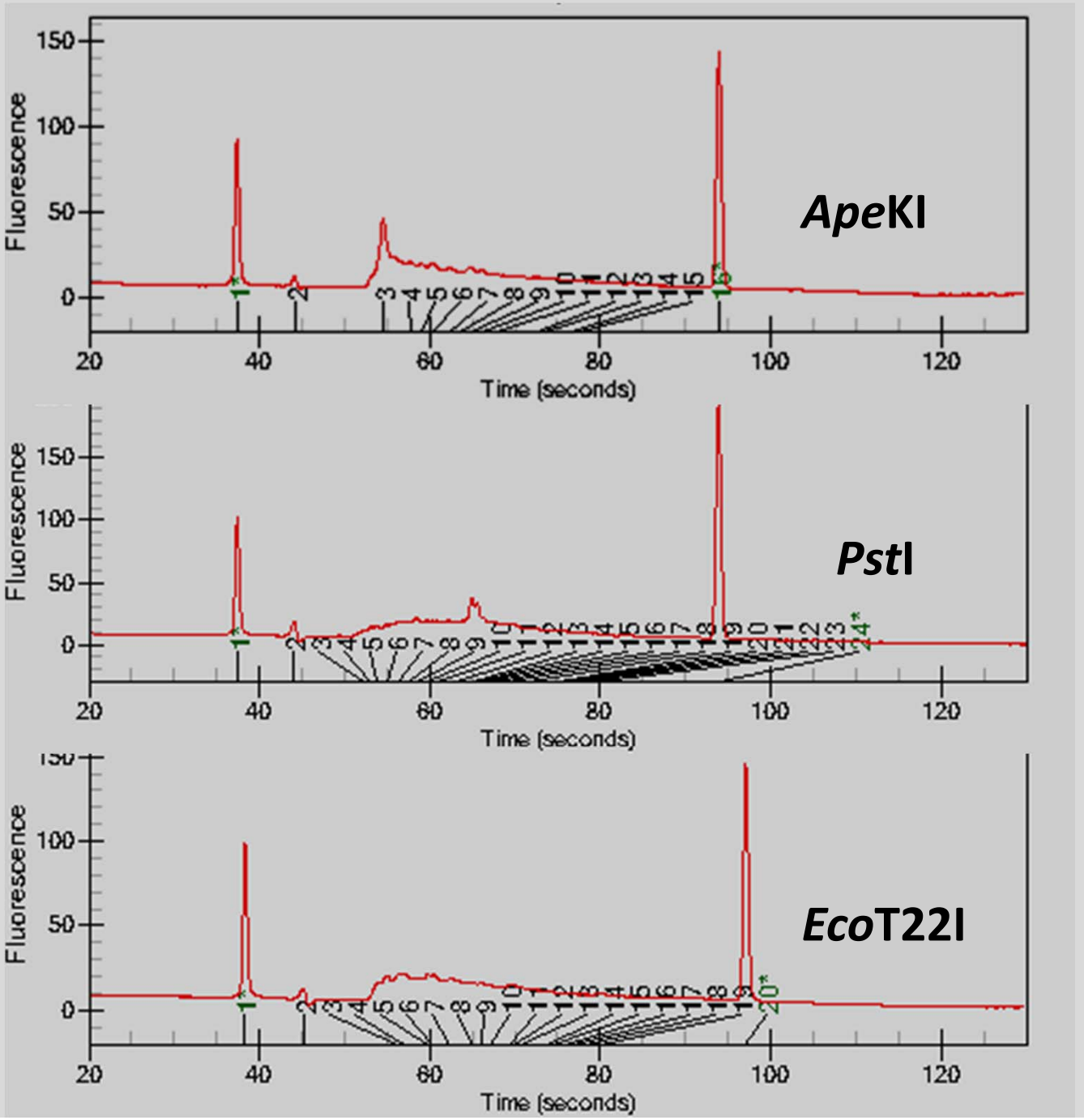


Deer Mouse

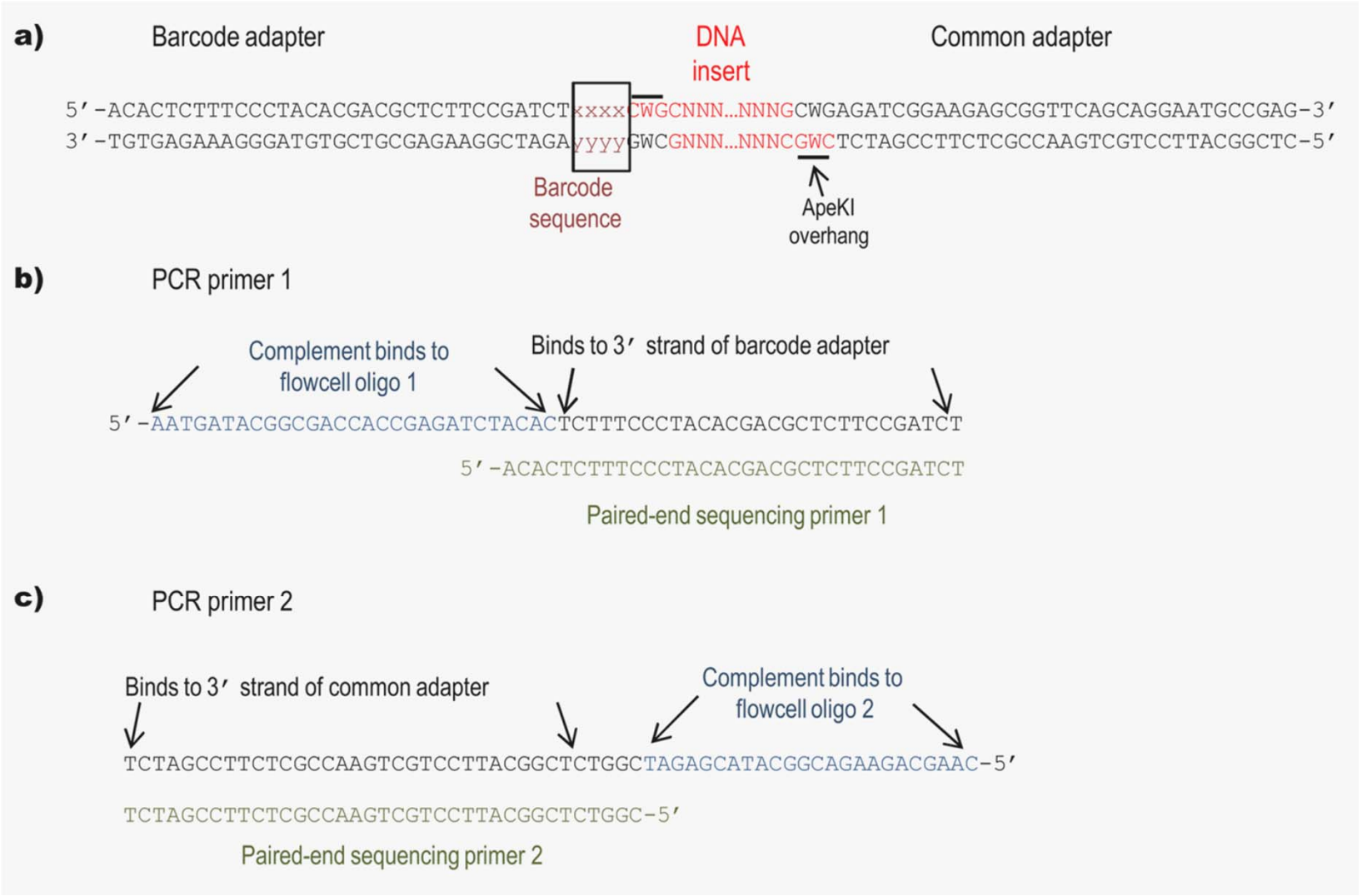

[FU]
100
80
60
40
20
0
43.00
113.00
15
100
200
300
400
500
700
1500
[bp]
*ApeK*I


[FU]
100
80
60
40
20
0
43.00
113.00
15
100
200
300
400
700
1500
[bp]
*Pst*I

Shrub Willow



GBS Adapter Design



Barcode Design Considerations

- **Barcode sets are enzyme specific**
  - **Must not recreate the enzyme recognition site**
  - **Must have complementary overhangs**
- **Sets must be of variable length**
- **Bases must be well balanced at each position**
- **Must different enough from each other to avoid confusion if there is a sequencing error.**
  - **At least 3 bp differences among barcodes.**
- **Must not nest within other barcodes**
- **No mononucleotide runs of 3 or more bases**

http://www.deenabio.com/services/gbs-adapters

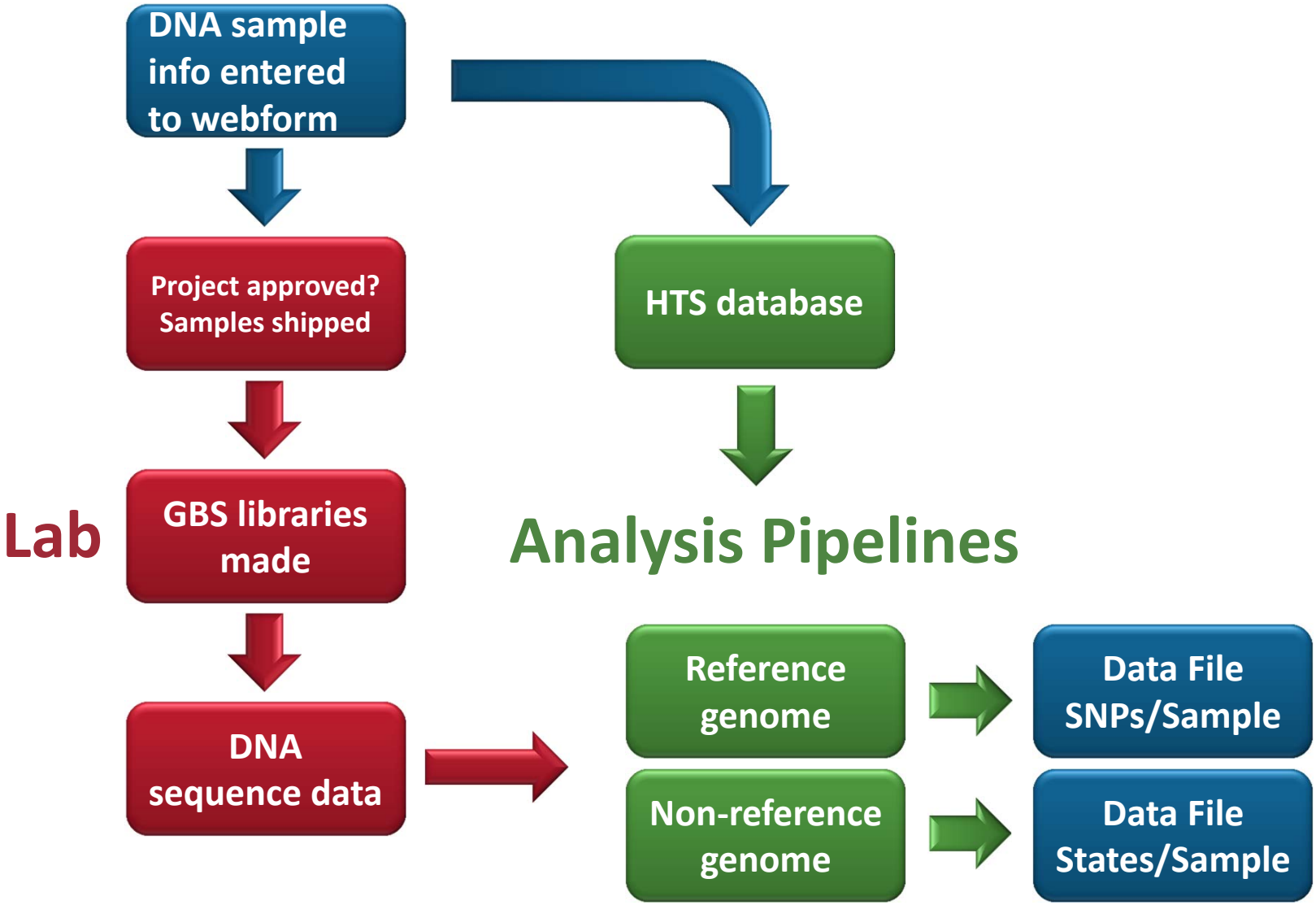
GBS SNP calls in *Sorghum bicolor*

091211.c10.hmp	alleles	chrom	SC283	BR007B	RIL_100	RIL_101	RIL_102	RIL_103	RIL_104	RIL_105	RIL_106	RIL_107	RIL_108	RIL_109	RIL_10	RIL_110	RIL_111	RIL_112	RIL_113	RIL_114	RIL_115
S10_29954	C/A	10	A	N	A	C	A	N	N	C	C	N	N	N	N	N	N	N	N	C	N
S10_29963	A/C	10	C	N	C	A	C	N	N	A	A	N	N	N	N	N	N	N	N	A	N
S10_29965	T/C	10	C	N	C	T	C	N	N	T	T	N	N	N	N	N	N	N	N	T	N
S10_84178	C/G	10	N	N	C	N	G	N	N	N	N	N	N	N	N	N	C	N	N	N	N
S10_84512	T/C	10	T	T	N	N	T	N	T	N	N	N	N	C	N	N	N	N	N	N	N
S10_84513	C/T	10	C	C	N	N	C	N	C	N	N	N	N	T	N	N	N	N	N	N	N
S10_85556	G/A	10	G	A	G	A	G	N	G	G	N	N	G	G	G	R	N	A	N	R	G
S10_85556	G/A	10	G	N	N	A	G	A	G	N	N	G	G	G	N	N	A	N	G	G	G
S10_115300	T/C	10	T	C	T	C	T	C	T	N	N	T	T	N	N	N	C	N	N	T	T
S10_133268	T/C	10	T	C	N	N	N	N	N	T	T	N	N	N	N	N	N	N	N	N	N
S10_158101	G/A	10	G	N	N	A	G	N	N	N	N	N	N	N	G	A	A	N	G	G	N
S10_158101	G/A	10	G	A	N	A	G	A	N	N	N	G	N	G	N	N	N	N	G	G	N
S10_163724	G/A	10	G	A	N	A	N	N	N	G	N	N	G	N	N	N	N	N	G	N	G
S10_163929	T/G	10	G	T	N	T	G	N	G	N	N	G	G	N	G	T	T	T	N	G	N
S10_163931	C/T	10	T	C	N	C	T	N	T	N	N	T	T	N	T	C	C	C	N	T	N
S10_180445	G/A	10	G	A	G	N	G	A	G	G	G	N	G	N	G	A	A	N	G	G	G
S10_209049	C/G	10	N	G	C	S	C	G	C	C	C	N	C	C	N	C	G	G	N	C	N
S10_209231	G/A	10	N	A	G	A	G	A	G	G	N	N	G	G	G	N	A	A	G	G	N
S10_214941	C/T	10	C	N	N	T	N	N	N	N	C	N	C	N	N	N	N	N	N	C	C
S10_234260	T/G	10	T	N	T	N	T	G	T	N	T	T	T	T	T	N	N	G	T	N	T
S10_234260	T/G	10	T	G	T	G	T	G	N	T	T	T	T	T	T	N	G	G	T	T	T
S10_238937	G/A	10	N	A	N	A	N	N	G	G	G	N	G	N	N	N	A	N	N	N	N
S10_252729	G/T	10	G	T	G	T	G	N	N	G	G	G	N	N	N	N	N	T	G	G	G
S10_252730	C/T	10	C	T	C	T	C	N	N	C	C	C	N	N	N	N	N	T	C	C	C
S10_261151	T/C	10	T	C	T	C	T	C	N	N	T	T	N	N	T	N	N	N	T	N	N
S10_268392	C/A	10	N	N	N	N	A	N	A	N	N	N	N	N	N	N	N	N	N	N	N
S10_268393	G/A	10	N	N	N	N	A	N	A	N	N	N	N	N	N	N	N	N	N	N	N
S10_274414	C/T	10	N	N	C	N	N	N	N	C	T	C	N	N	C	T	N	N	C	N	C
S10_281272	T/C	10	T	N	T	N	N	N	N	N	N	N	N	N	N	N	N	N	N	T	T
S10_287523	G/T	10	G	N	G	N	N	N	N	G	G	G	N	N	N	N	N	N	N	N	N
S10_292020	A/T	10	A	T	N	N	A	N	N	A	A	A	N	A	A	N	T	T	A	A	A
S10_292020	A/T	10	A	T	A	T	N	N	N	A	N	A	N	A	N	N	T	T	A	N	A
S10_294189	C/T	10	N	N	C	N	C	N	C	C	C	N	N	N	C	T	N	N	C	N	C
S10_302638	C/T	10	C	T	C	T	C	N	N	C	N	C	C	N	C	T	T	N	C	C	N
S10_312560	A/C	10	A	C	A	N	A	N	A	A	A	A	A	A	A	C	C	N	A	A	A
S10_347848	T/C	10	T	C	T	C	T	C	N	T	T	T	T	N	T	N	C	N	T	T	T

Missing Data Strategies

- Impute
- Technical Options
  - Reduce the multiplexing level
  - Sequence the same library multiple times
- Molecular Options
  - Choose less frequently cutting enzymes

GBS Workflow



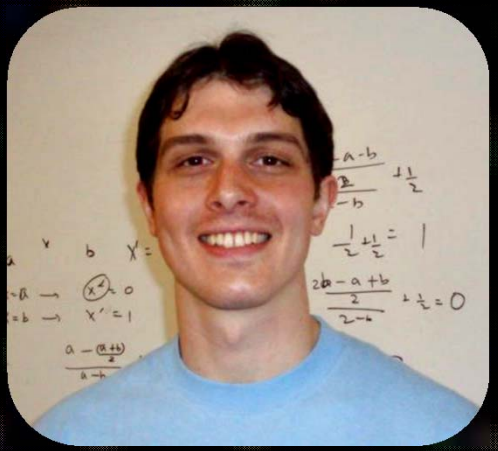
GBS Team

Rob Elshire

Ed Buckler

Sharon Mitchell

Jeff Glaubitz



James Harriman

Qi Sun