Genotyping By Sequencing (GBS) Method Overview

Charlotte B. Acharya
Institute for Genomic Diversity
Cornell University

http://www.igd.cornell.edu/

Cornell University

USDA ARS

NSF

IGD

Topics Presented

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

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)

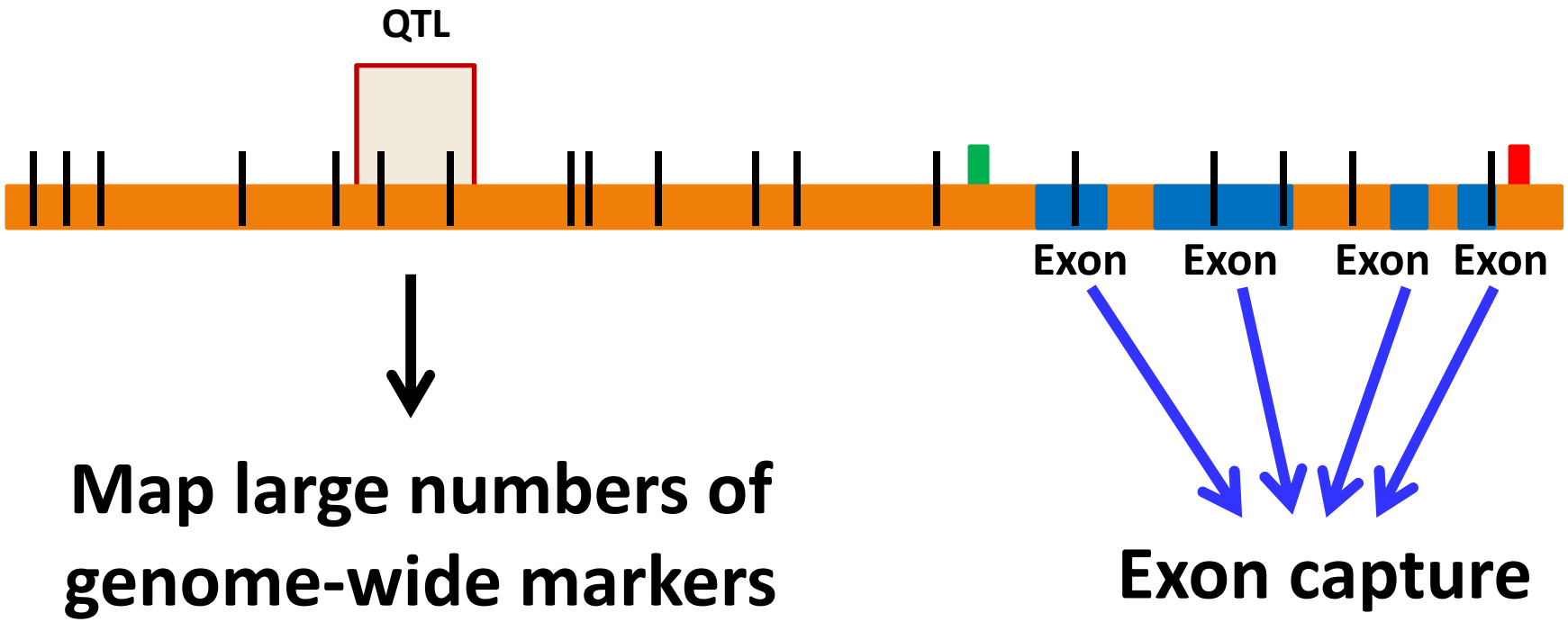
II. Restriction Enzymes (REs)

Technically less challenging

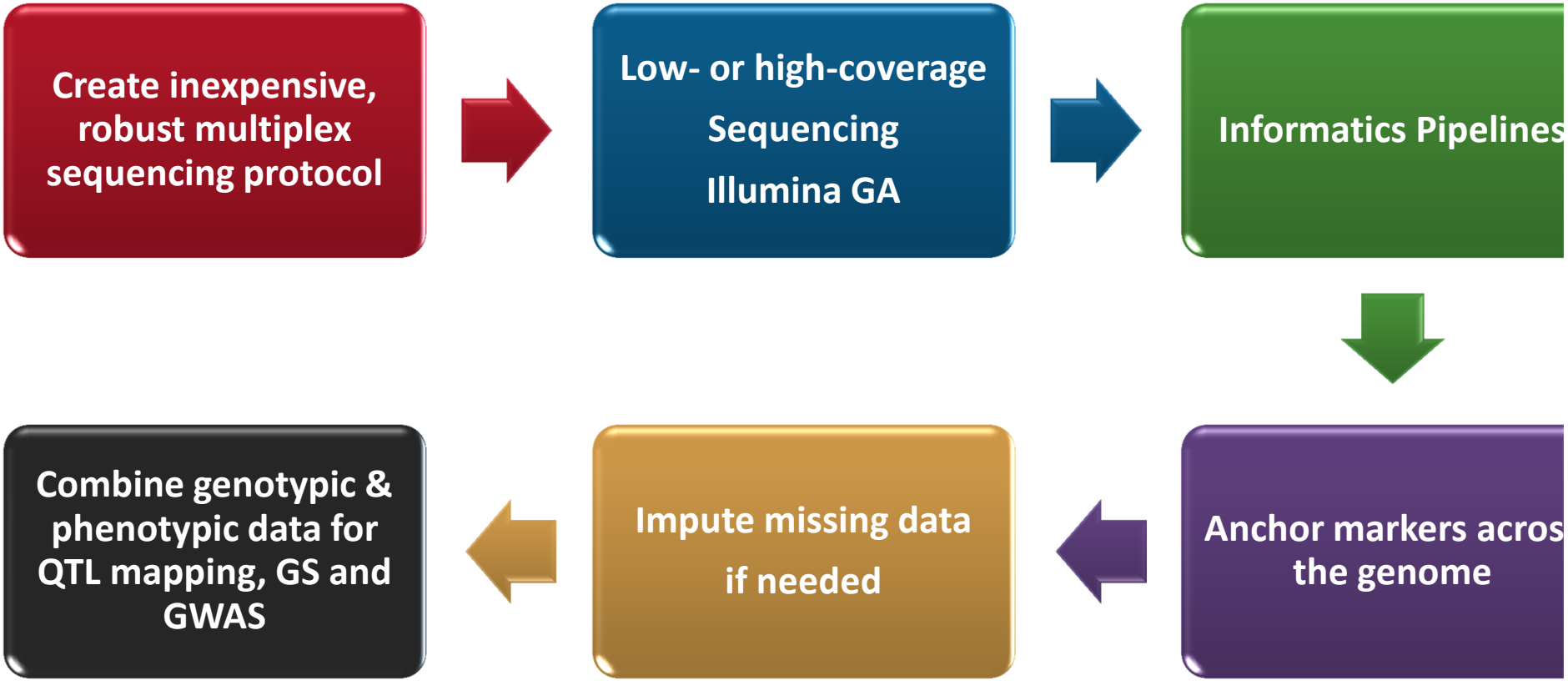
- Methylation sensitive REs filter out repetitive genomic fraction

QTL are often located in non-coding regions

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



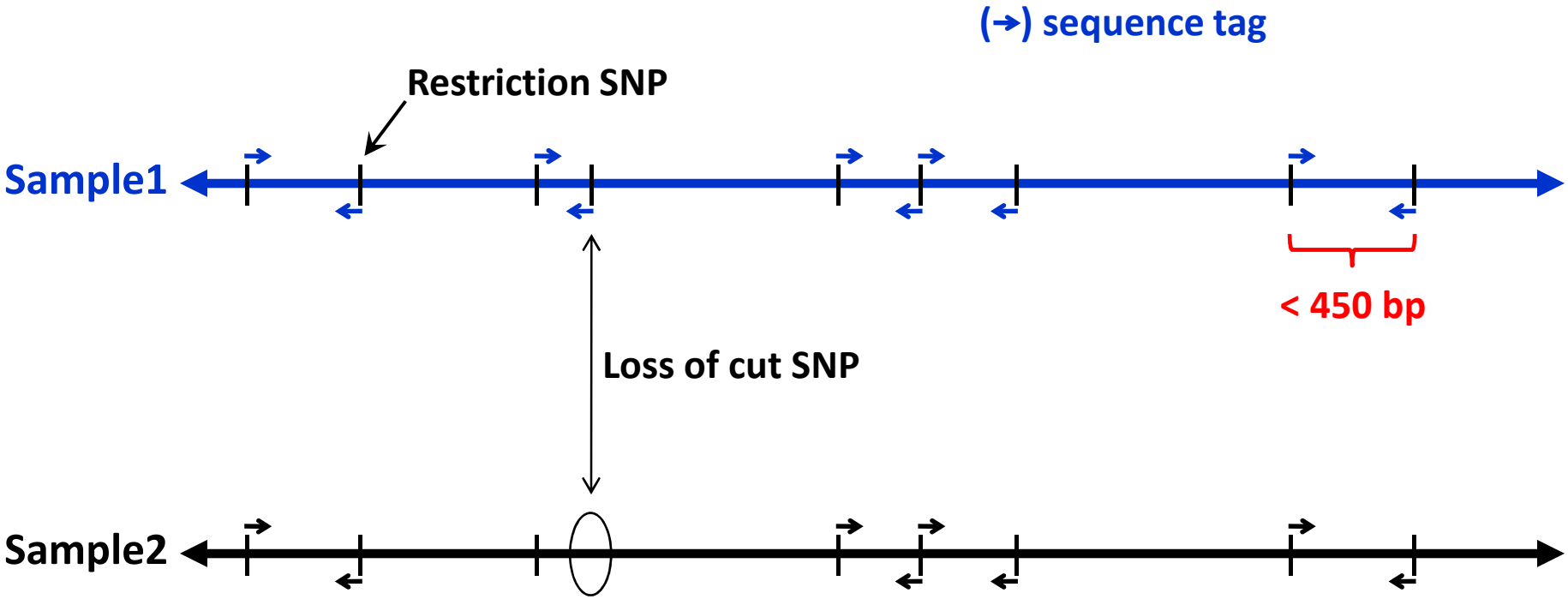
We have created a public genotyping/informatics platform based on next-generation sequencing



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 available for public projects.**

Overview of Genotyping by Sequencing (GBS)



- Focuses NextGen sequencing power to ends of restriction fragments
- Both SNPs and presence/absence markers can be scored
- Small indels are identified but are not scored

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

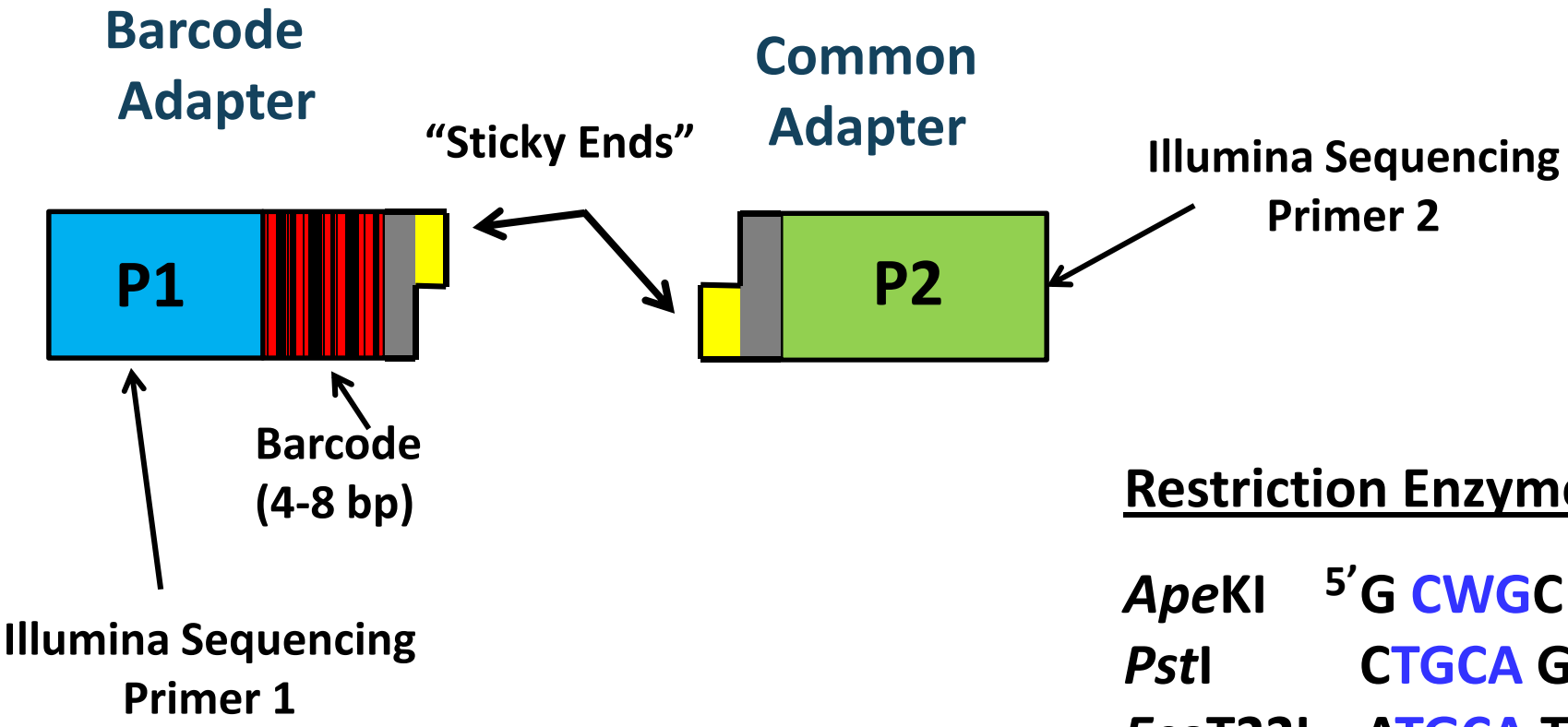
GBS 96- or 384-plex Protocol

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

1. Plate DNA & adapter pair

2. Digest DNA with RE
3. Ligate adapters

GBS Adapters and Enzymes

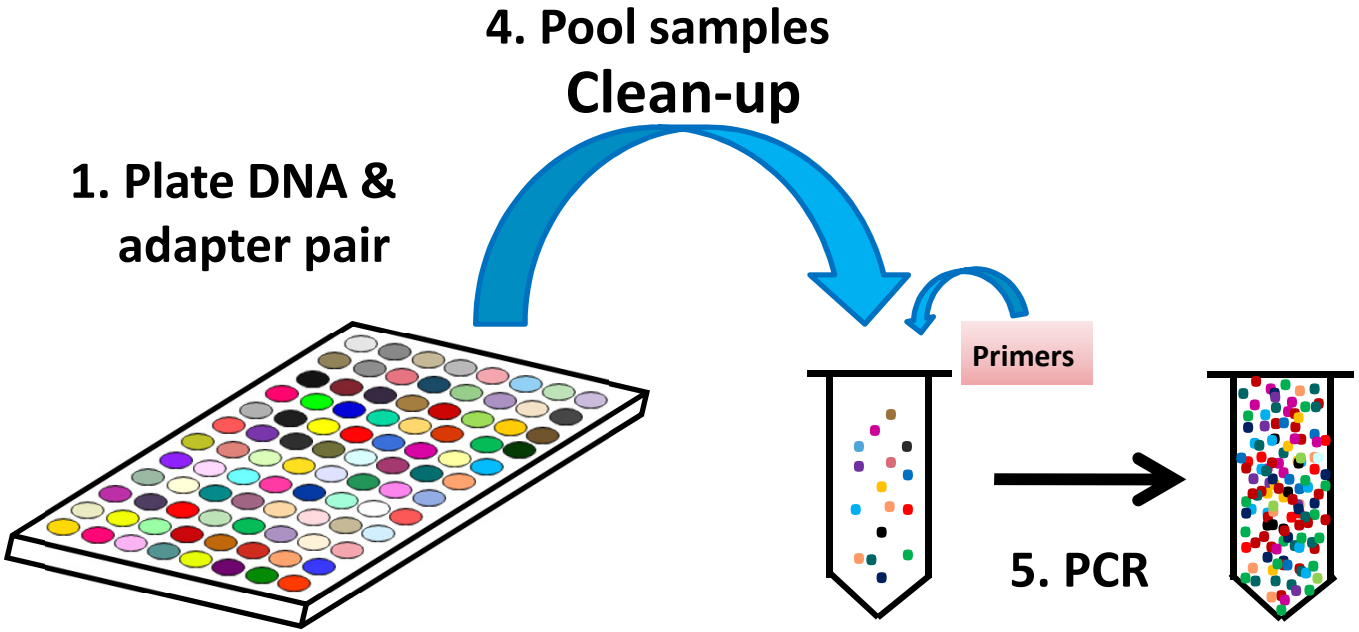


Restriction Enzymes

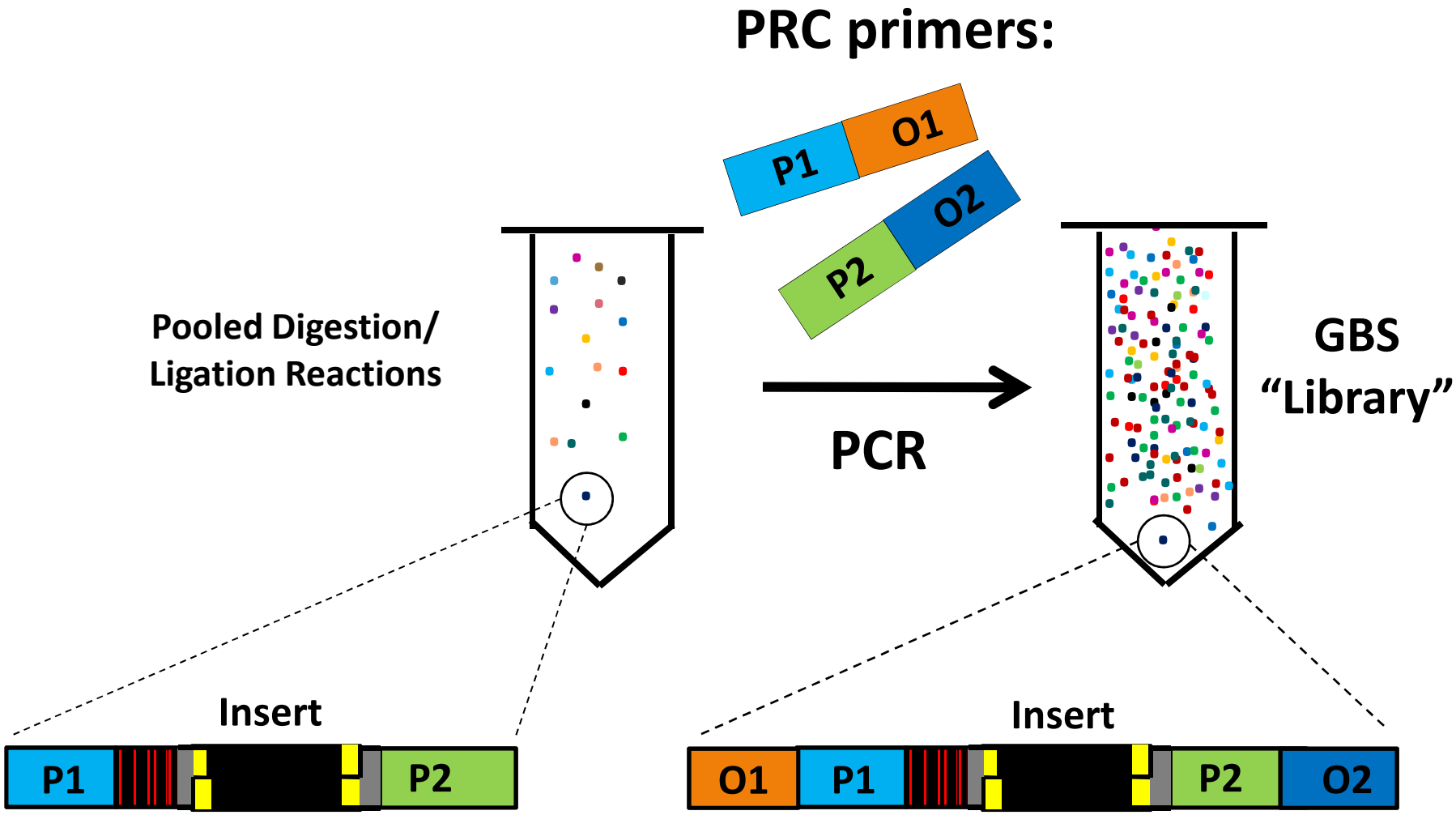
*Ape*KI	5'G CWGC3'
*Pst*I	CTGCA G
*Eco*T22I	ATGCA T

GBS 96- or 384-plex Protocol

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

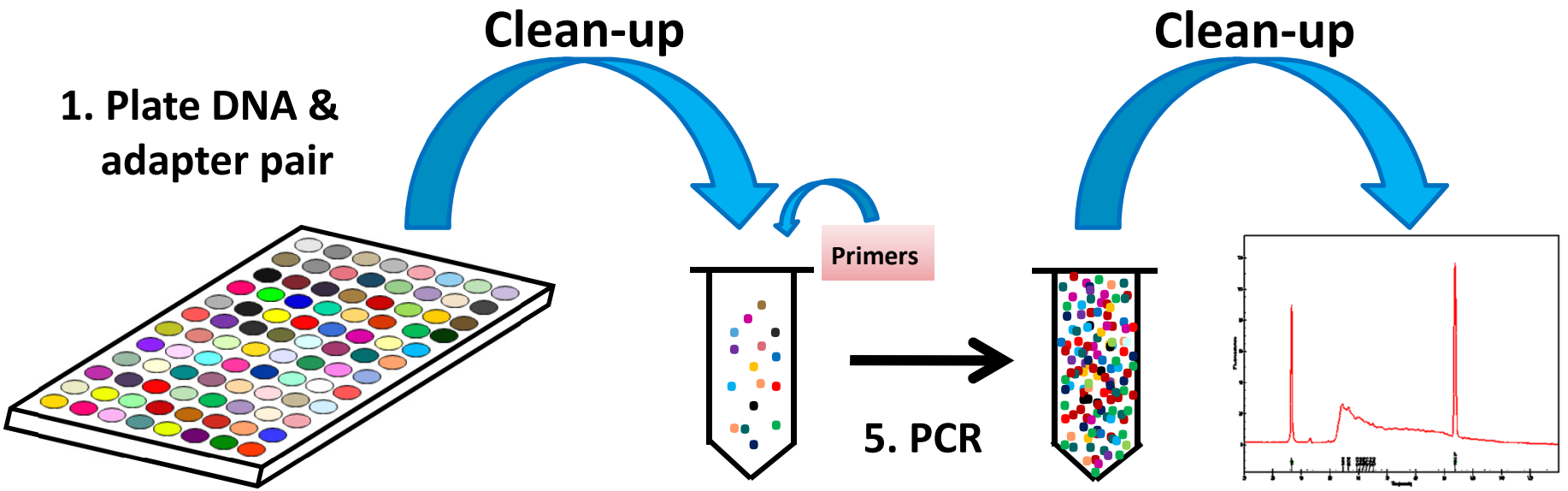


2. Digest DNA with RE
3. Ligate adapters


PRC primers:
P1
O1
P2
O2
Pooled Digestion/
Ligation Reactions
PCR
GBS
“Library”
Insert
P1
P2
Insert
O1
P1
P2
O2

GBS 96- or 384-plex Protocol

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

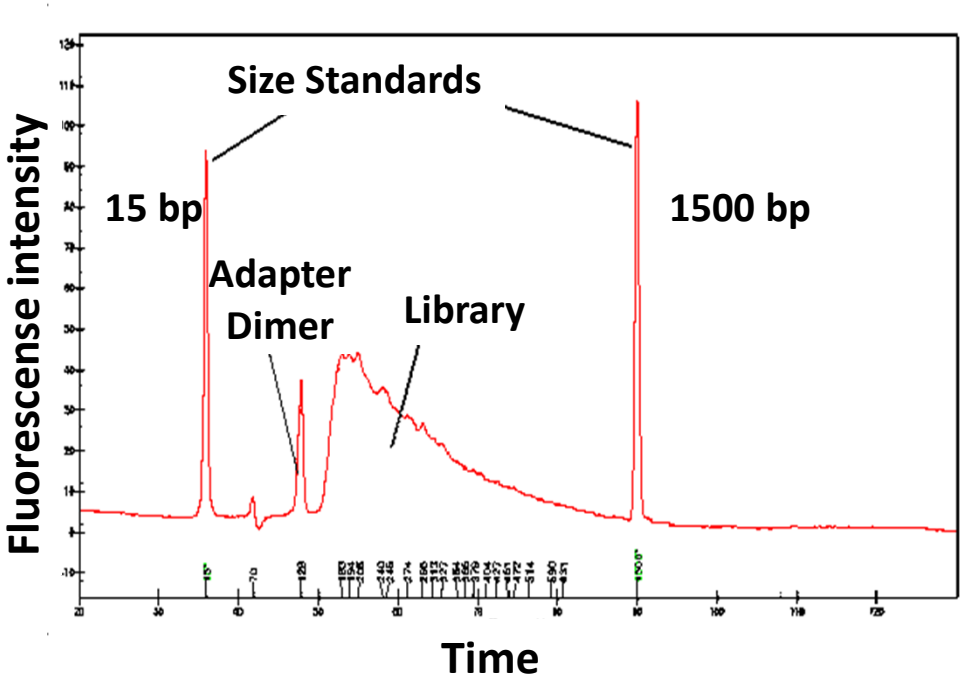


2. Digest DNA with RE
3. Ligate adapters
4. Pool DNA aliquots

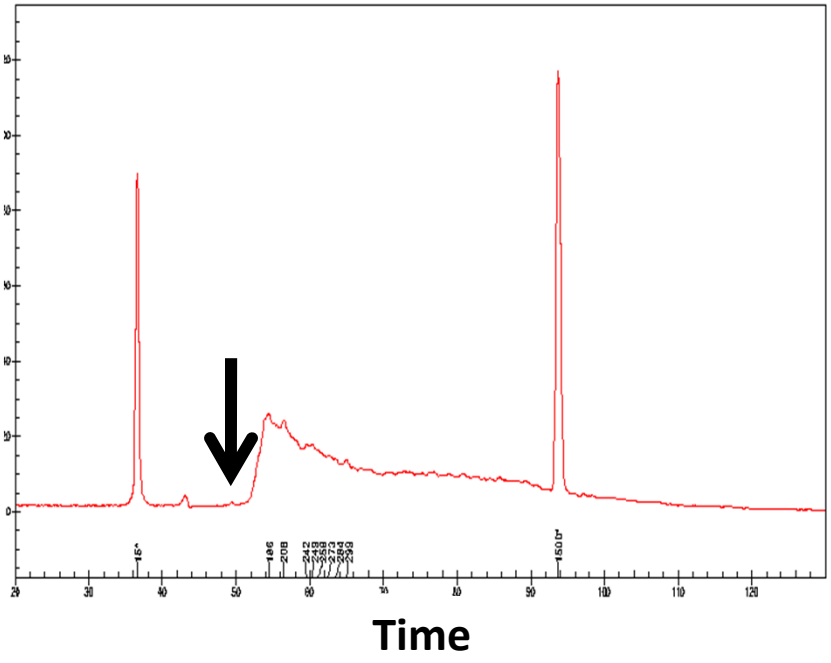
Perform Titration to Minimize Adapter Dimers Before Sequencing

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

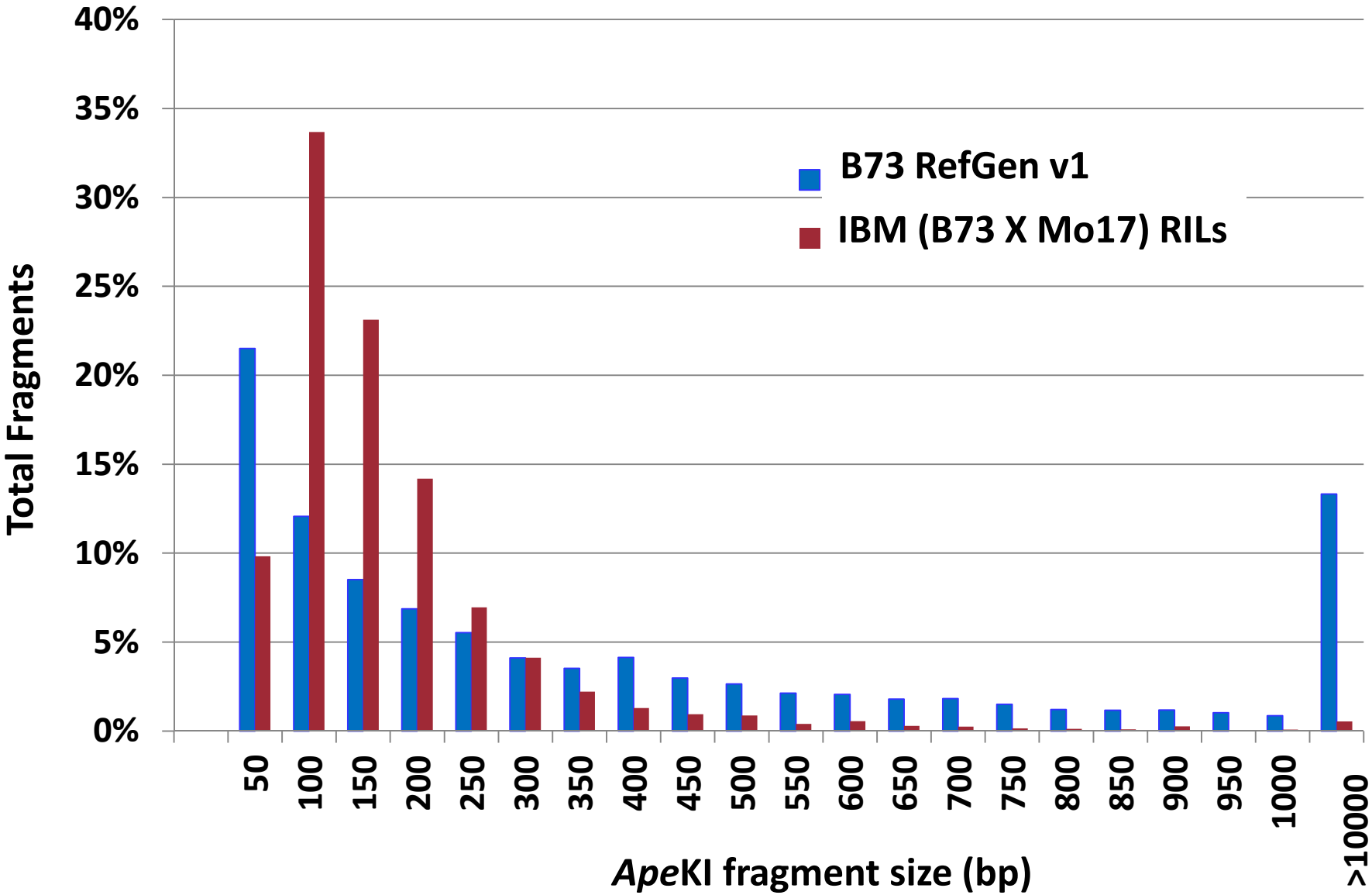
Non-optimized library



Optimal adapter amount

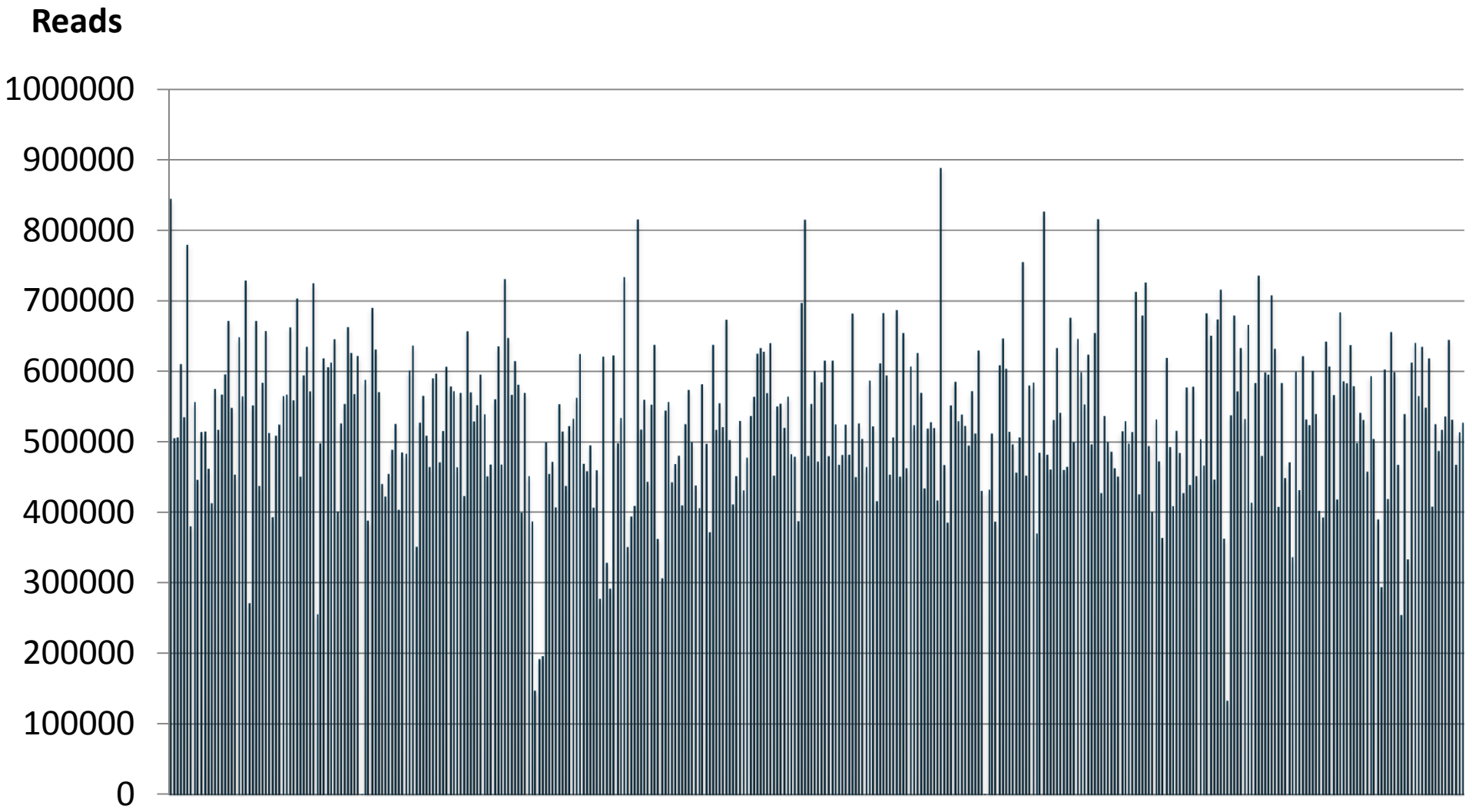


Small Fragments are Enriched in GBS Libraries

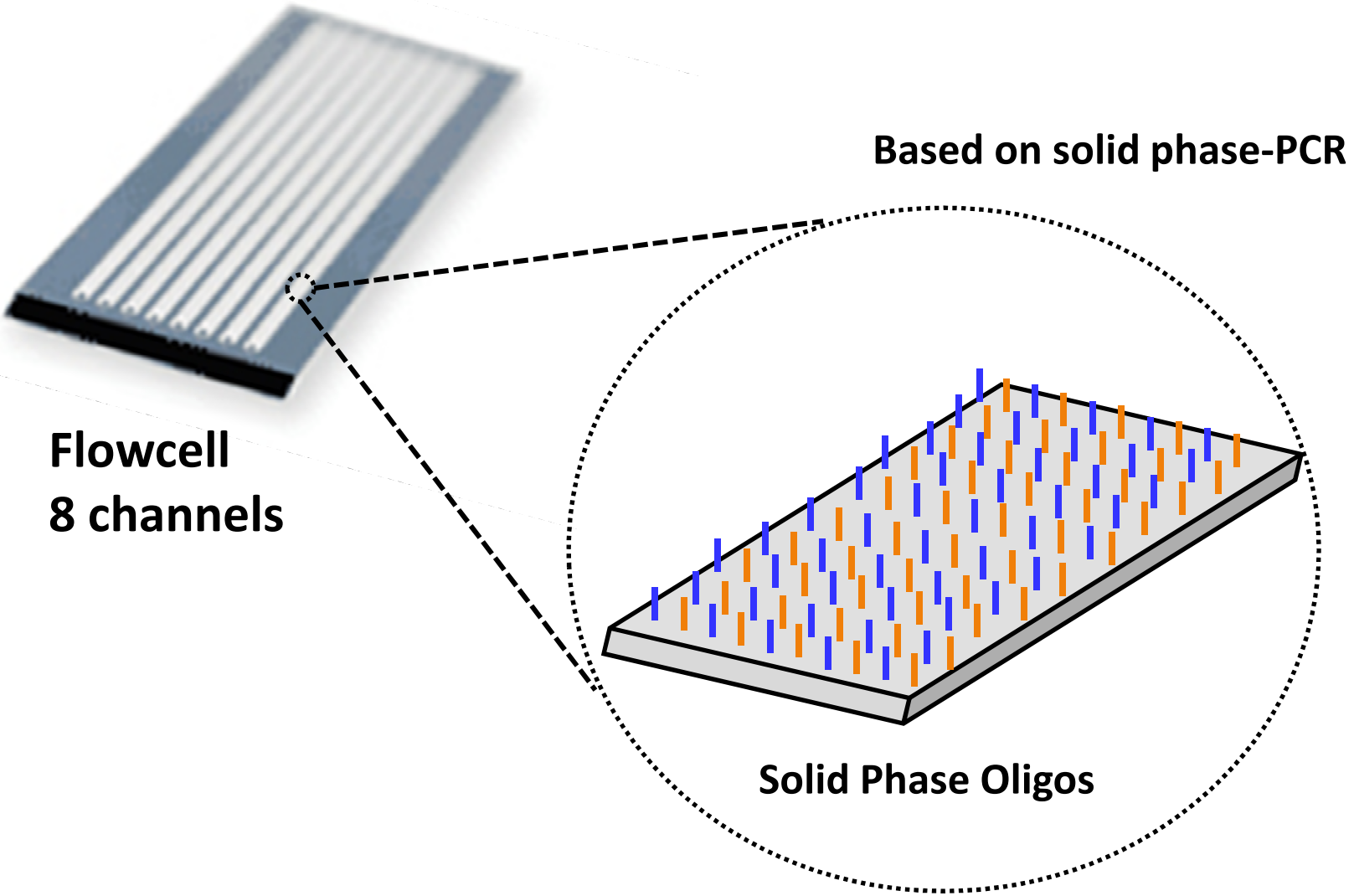


384-plex GBS Results for Maize

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



Illumina Sequencing by Synthesis Review



Cluster Formation Amplifies Sequencing Signal

CBot

Denatured “Library”

Bridge Amplification

PCR

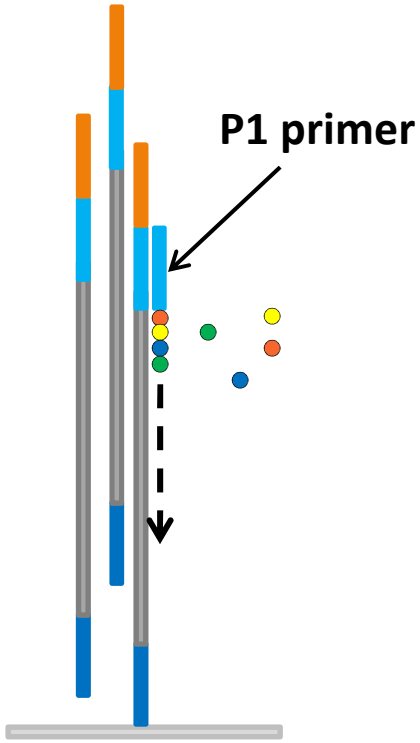
Cleavage

Flow cell with bound oligos

Linearization

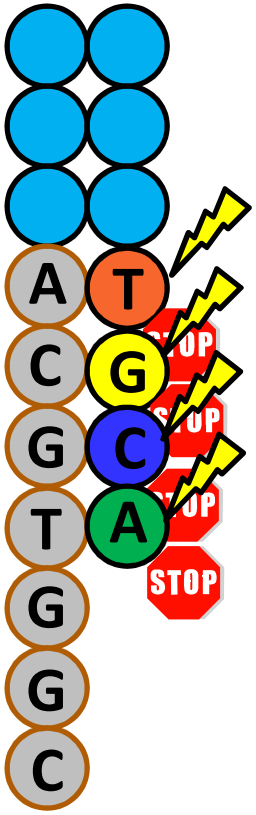
Sequencing by Synthesis

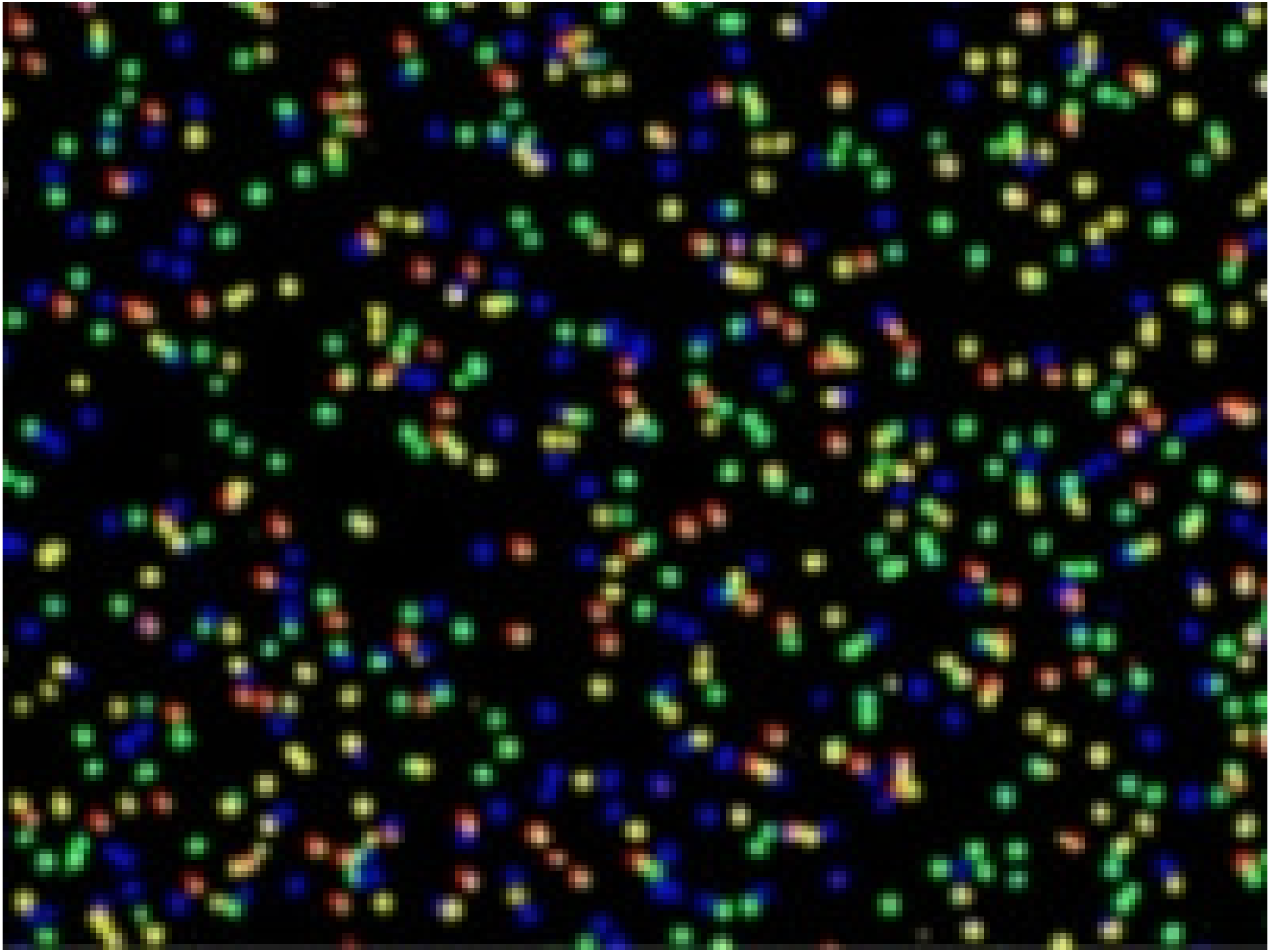
HiSeq 2000



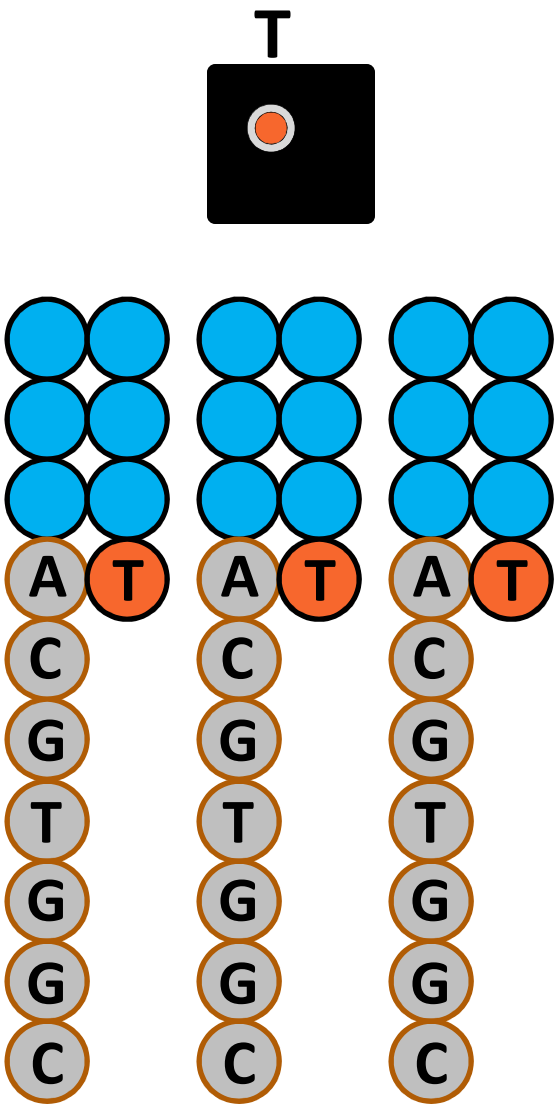
TGCA

Flowcell

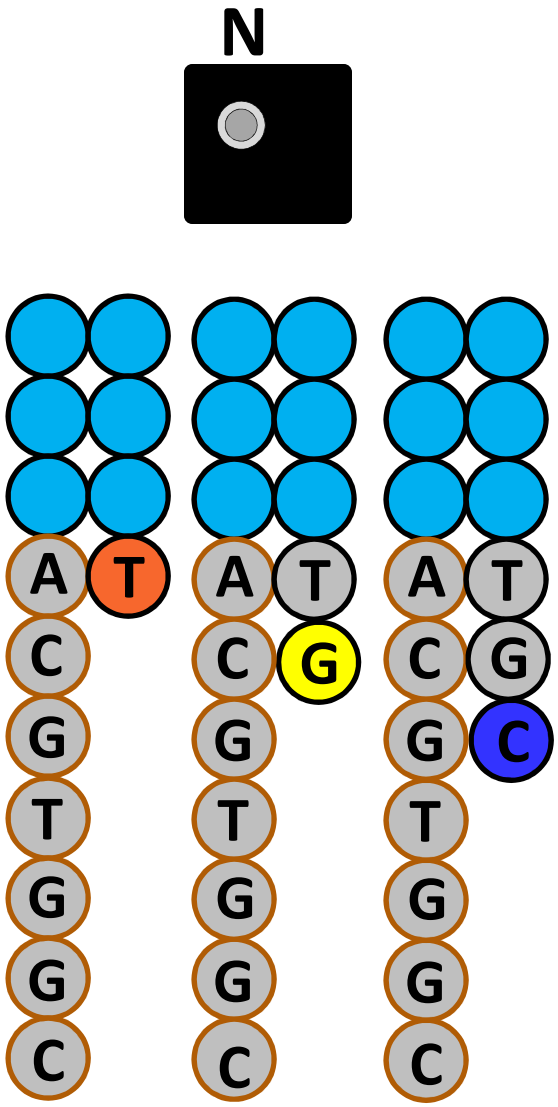




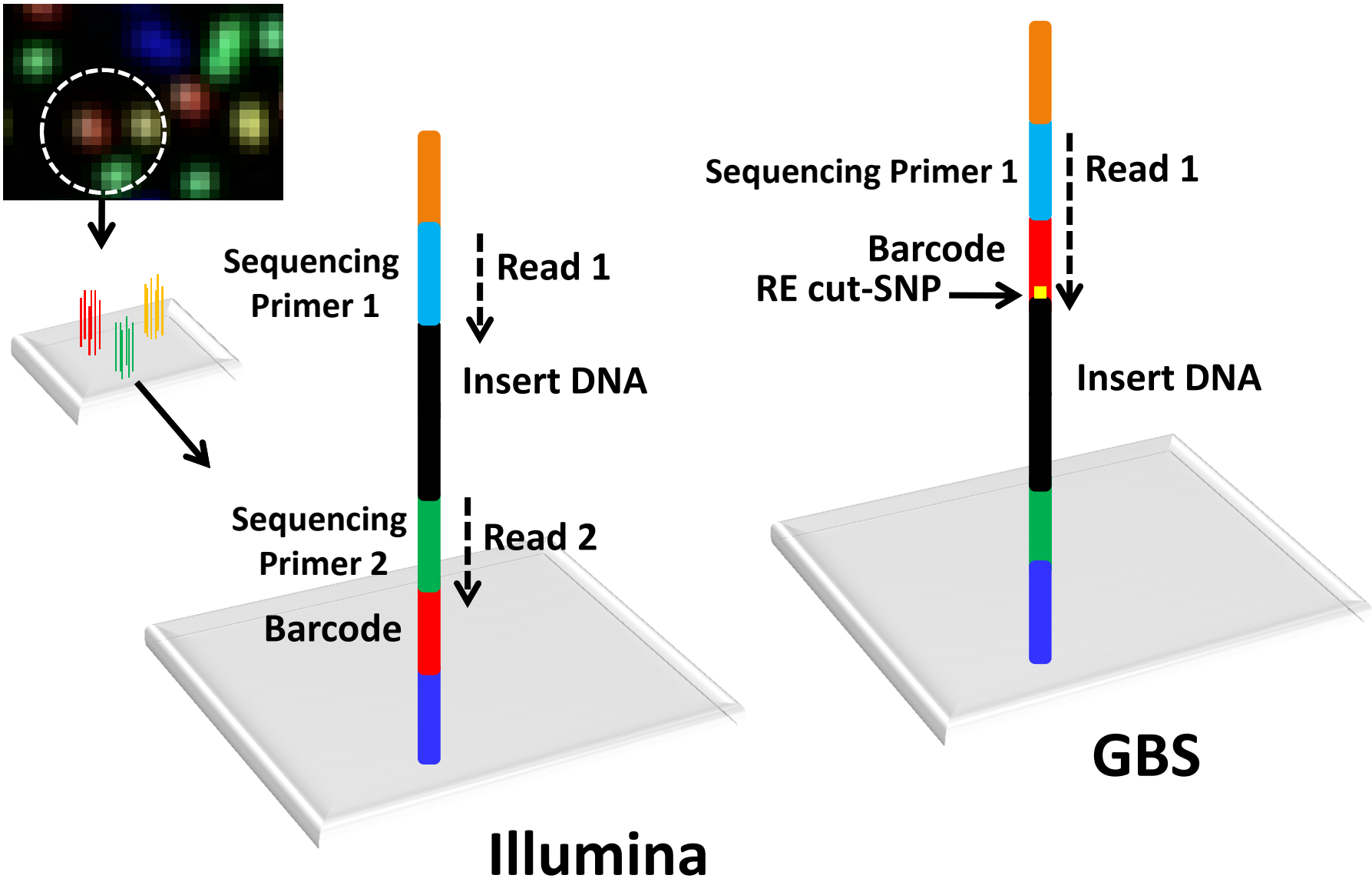
“In Phase”



“Out of Phase”

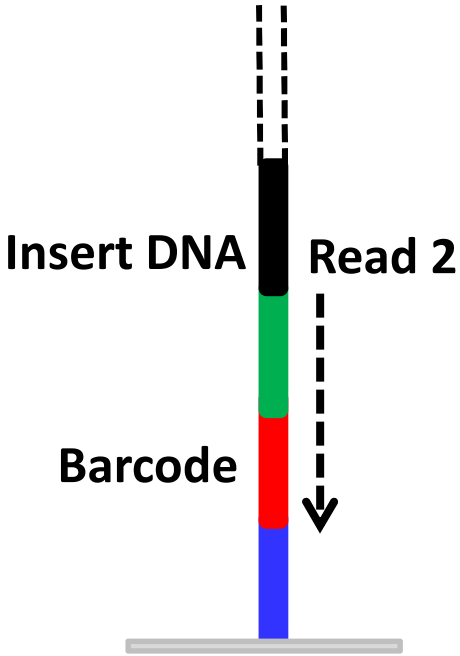


GBS captures barcode and insert DNA sequence in single read



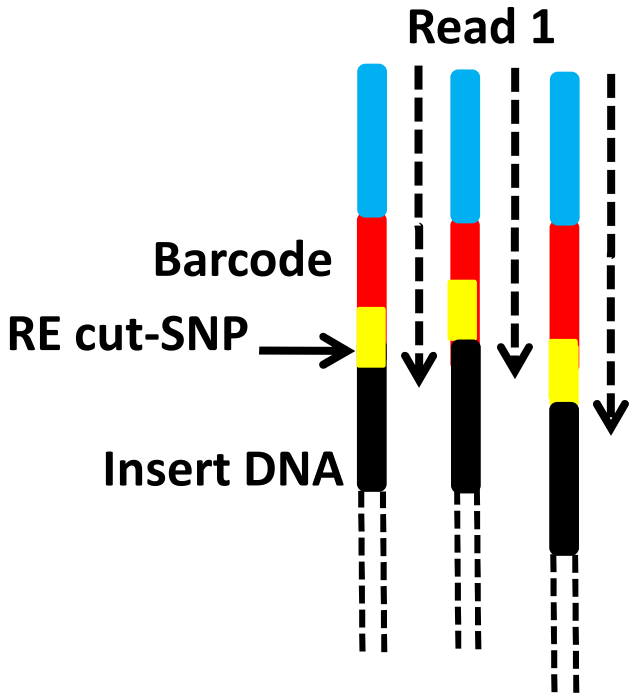
Variable Length GBS Barcodes Solves Sequence Phasing Issues

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



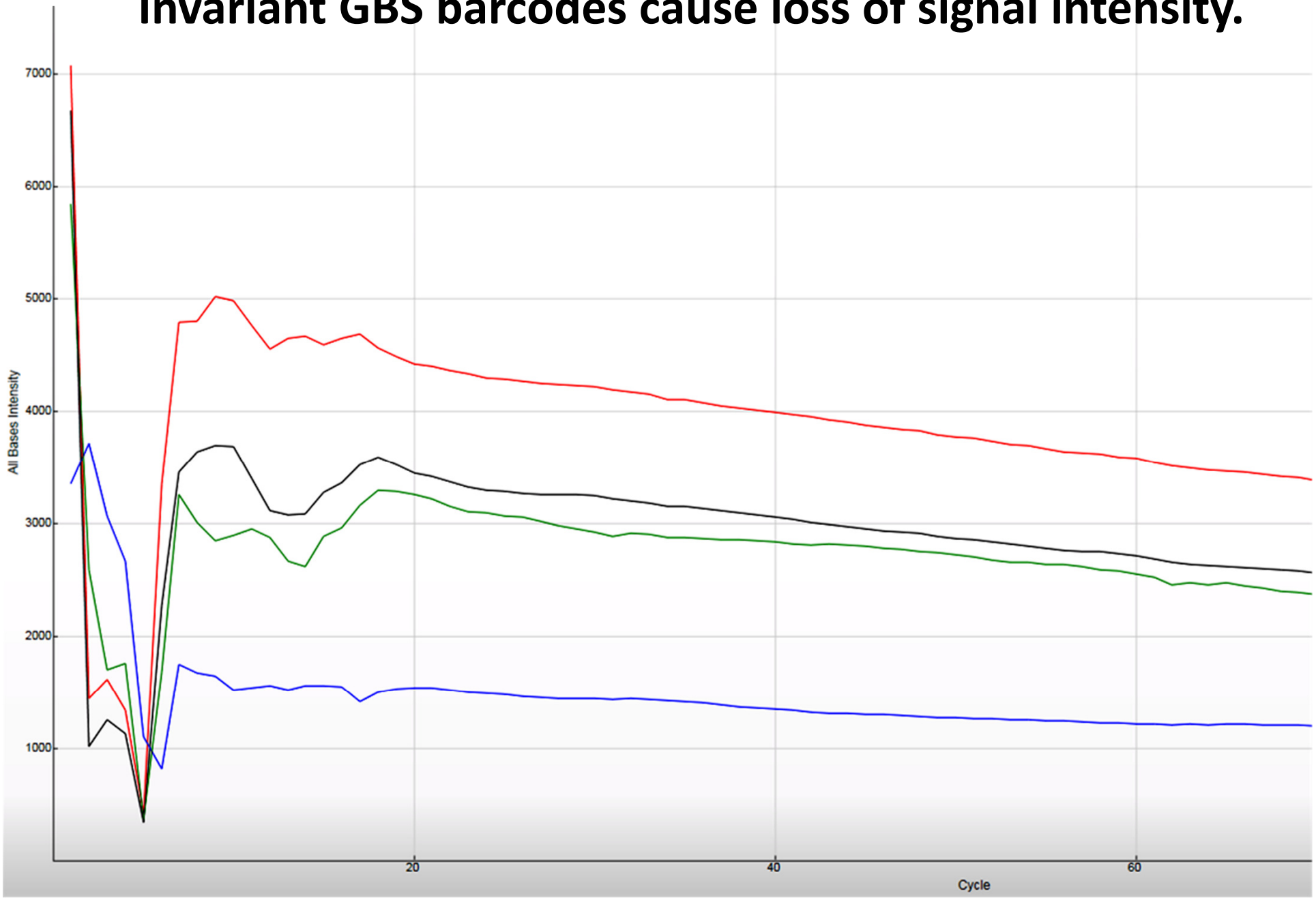
Illumina

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

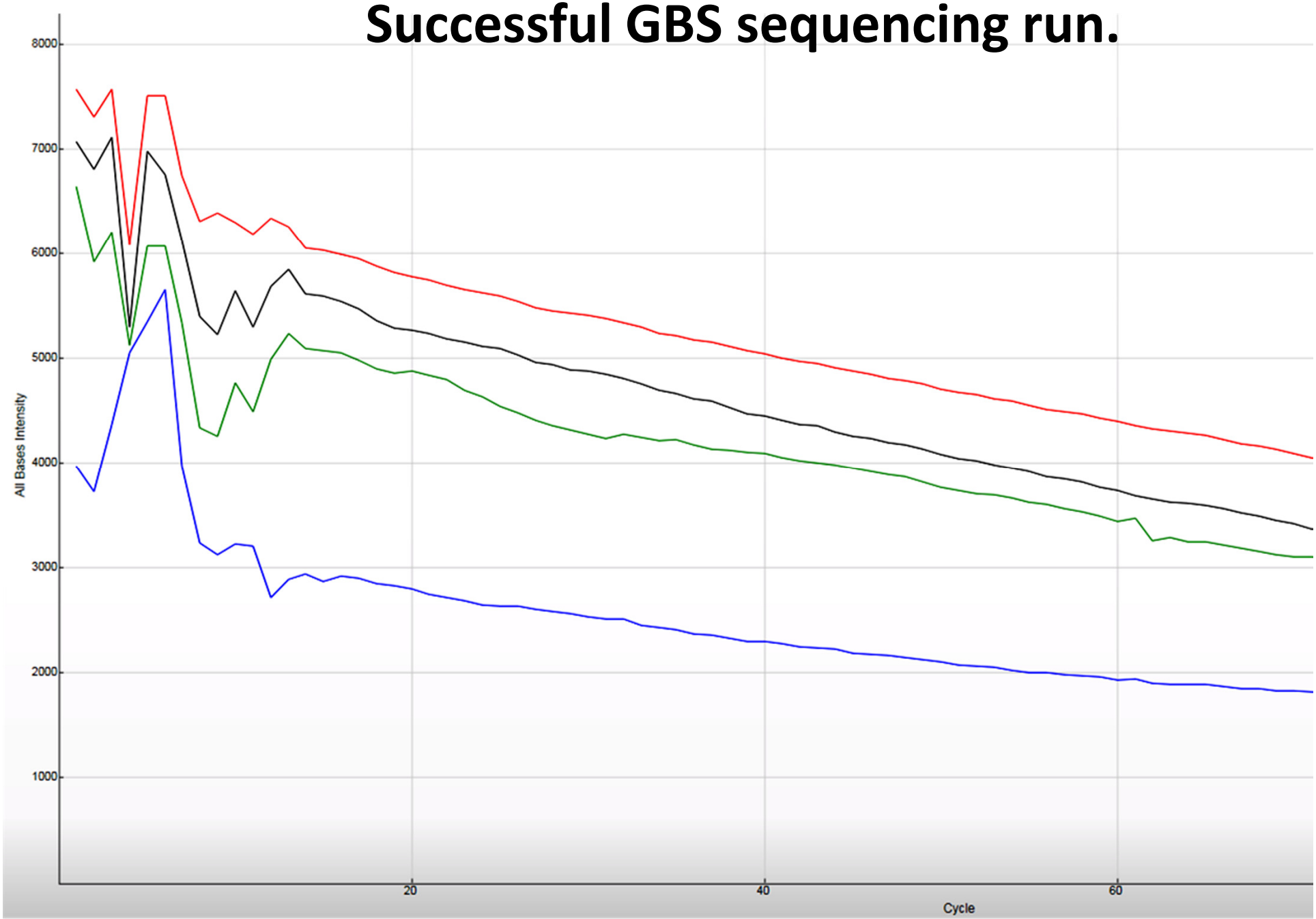


GBS

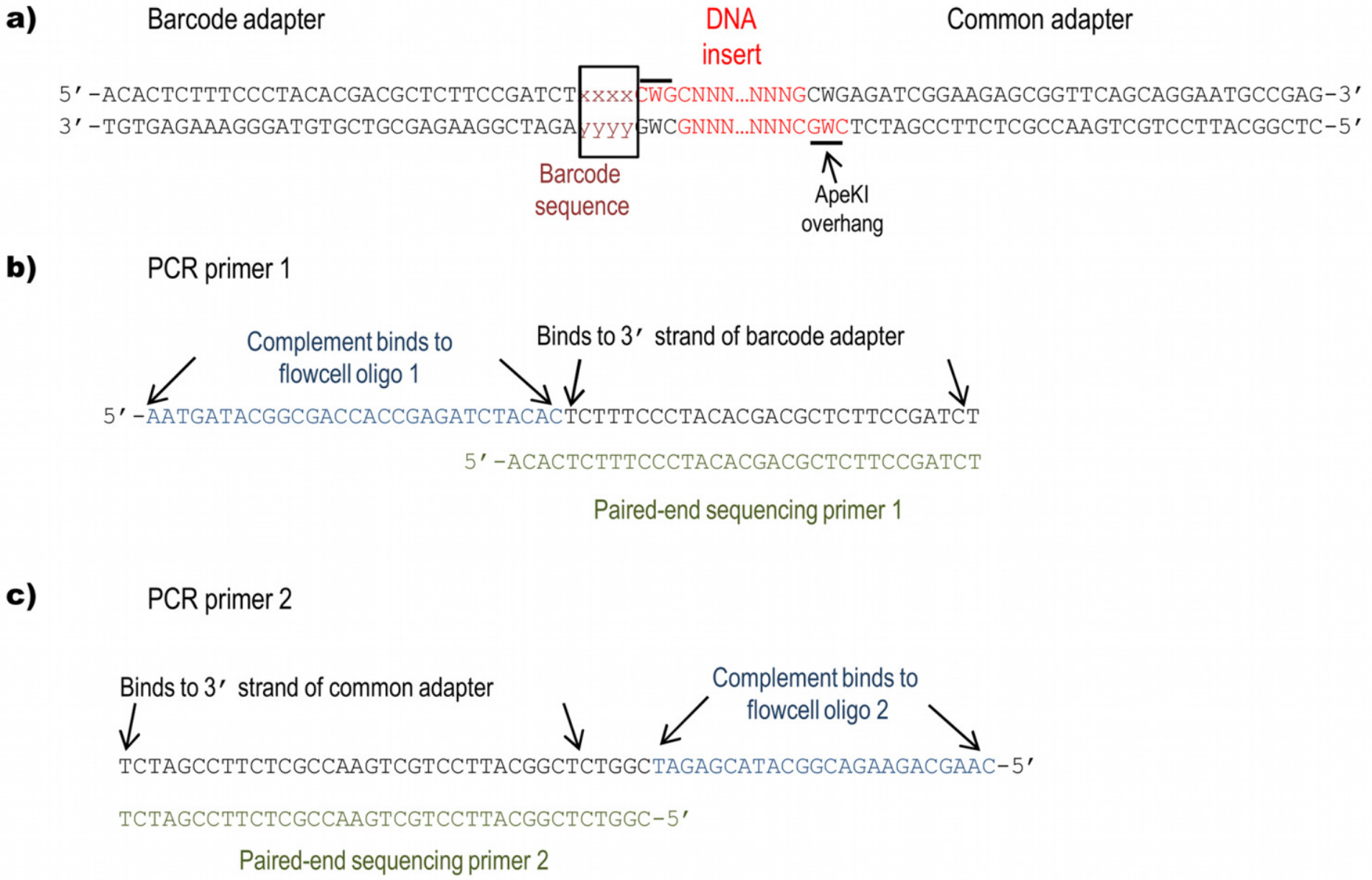
Invariant GBS barcodes cause loss of signal intensity.



Successful GBS sequencing run.



GBS Adapter Design



Barcode Design Considerations

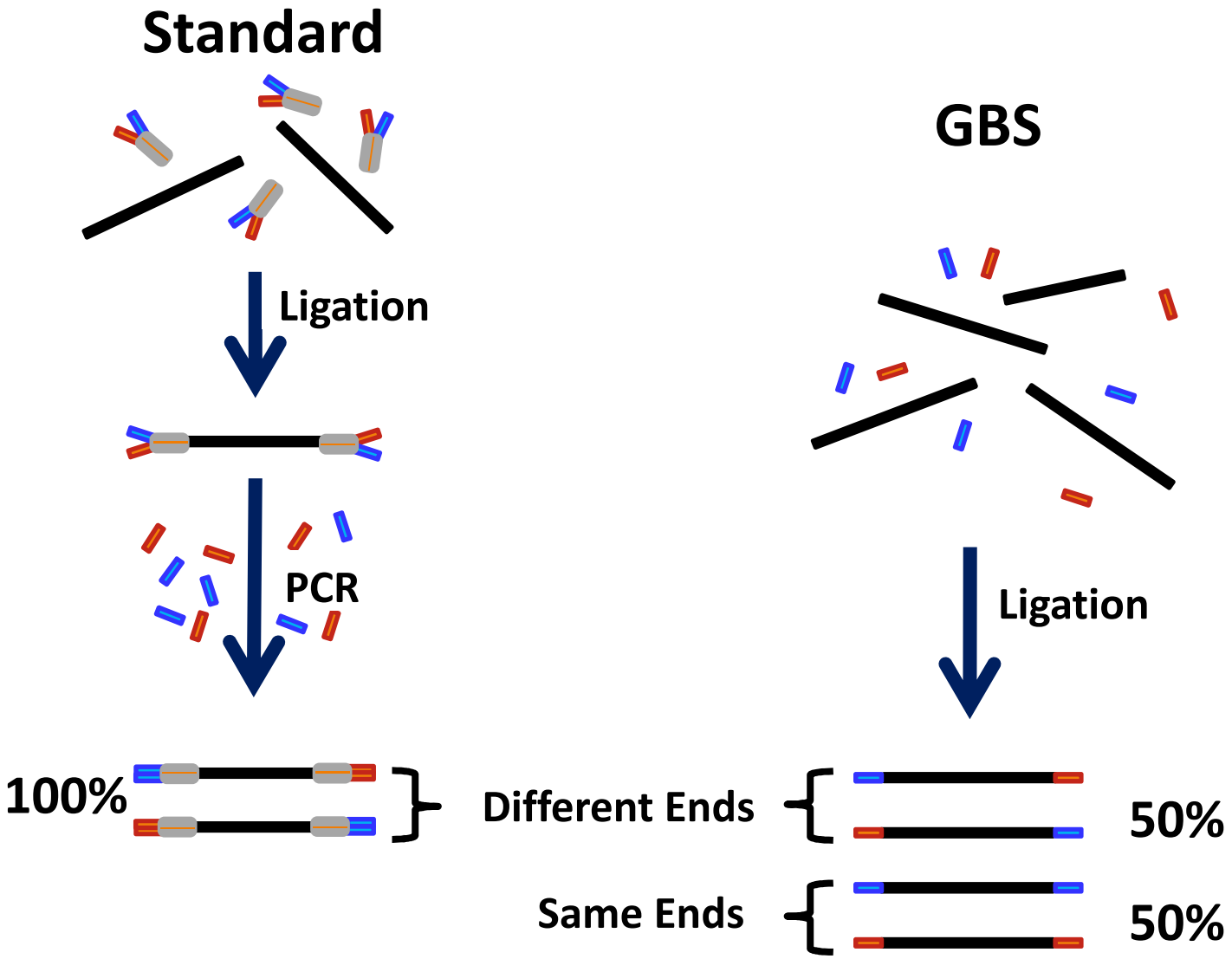
- **Barcode sets are enzyme specific**
  - **Must not recreate the enzyme recognition SNP**
  - **Must have complementary overhangs**
- **Sets must be of variable length**
- **Bases must be well balanced at each position**
- **Must different enough from each other to avoid confusion if there is a sequencing 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

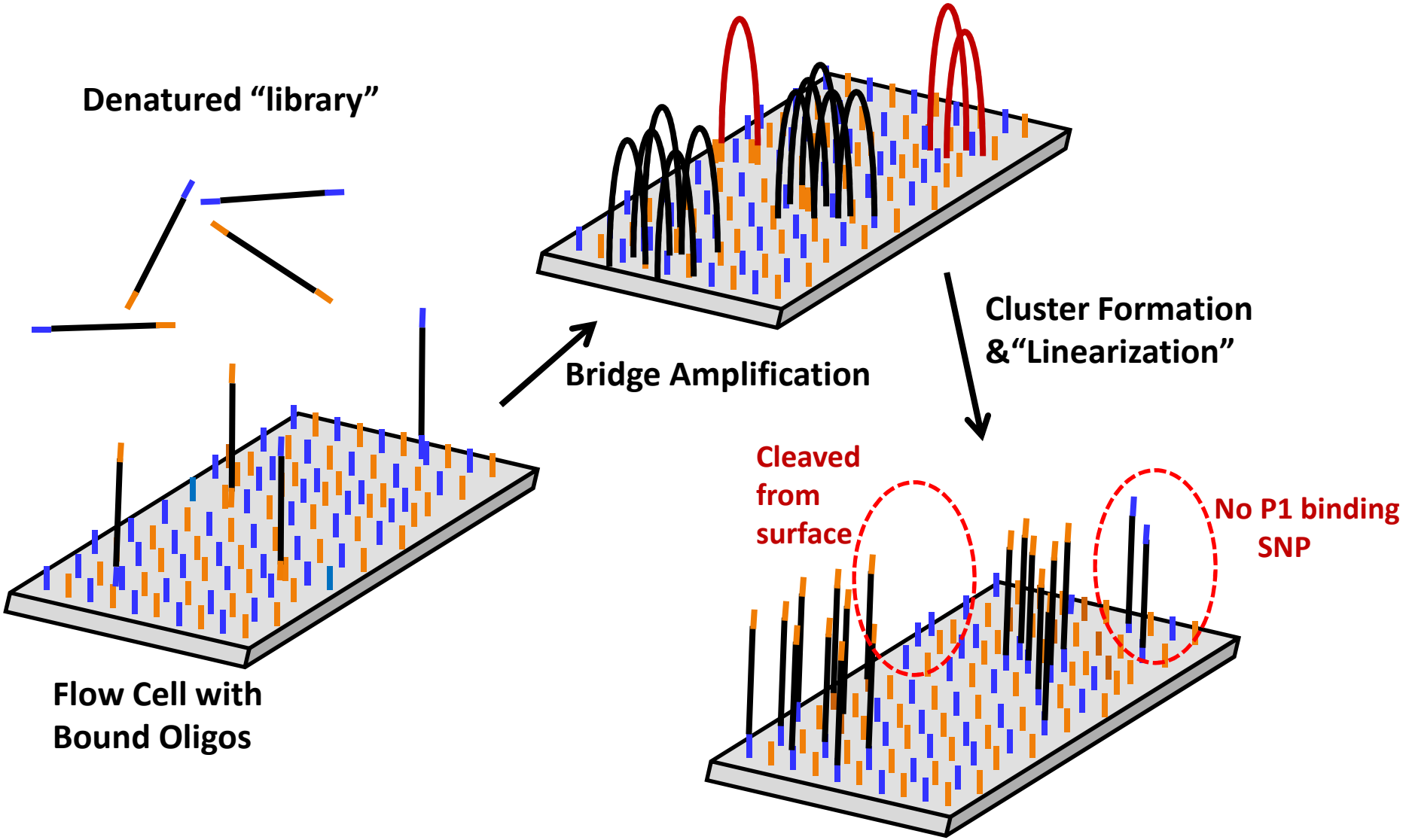
Most significant GBS technical issues?

- DNA quality
- DNA quantification

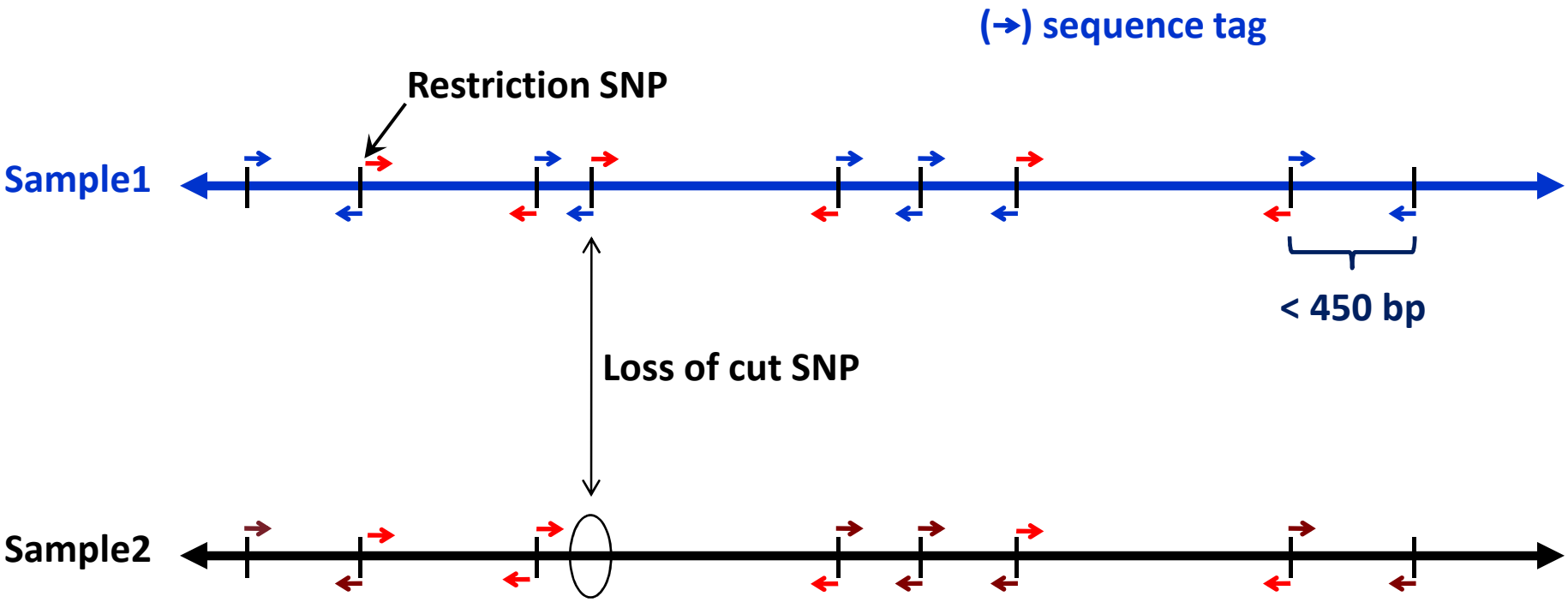
GBS does not use standard “Y” adapters



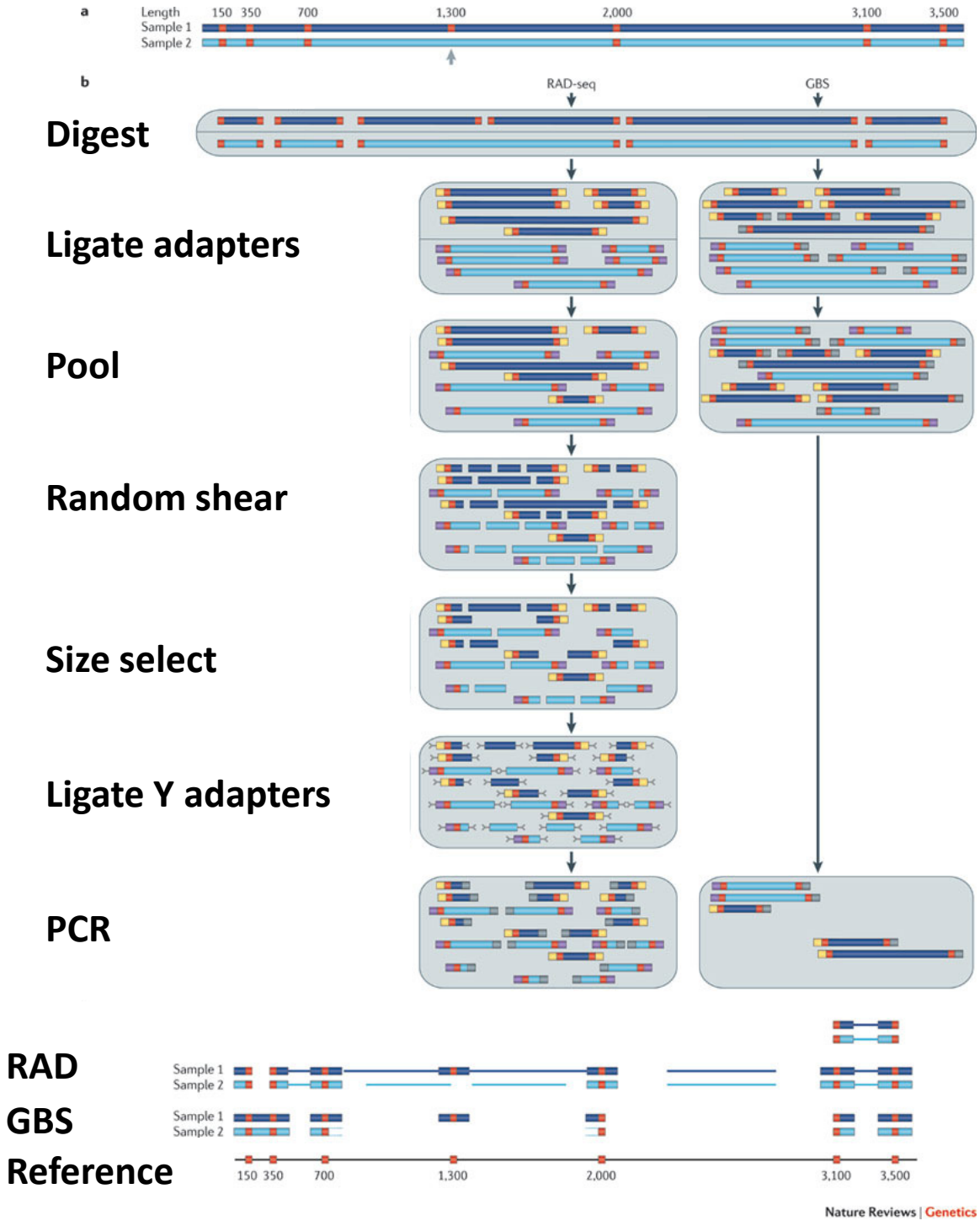
Same-ended Fragments Do Not Form Clusters



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

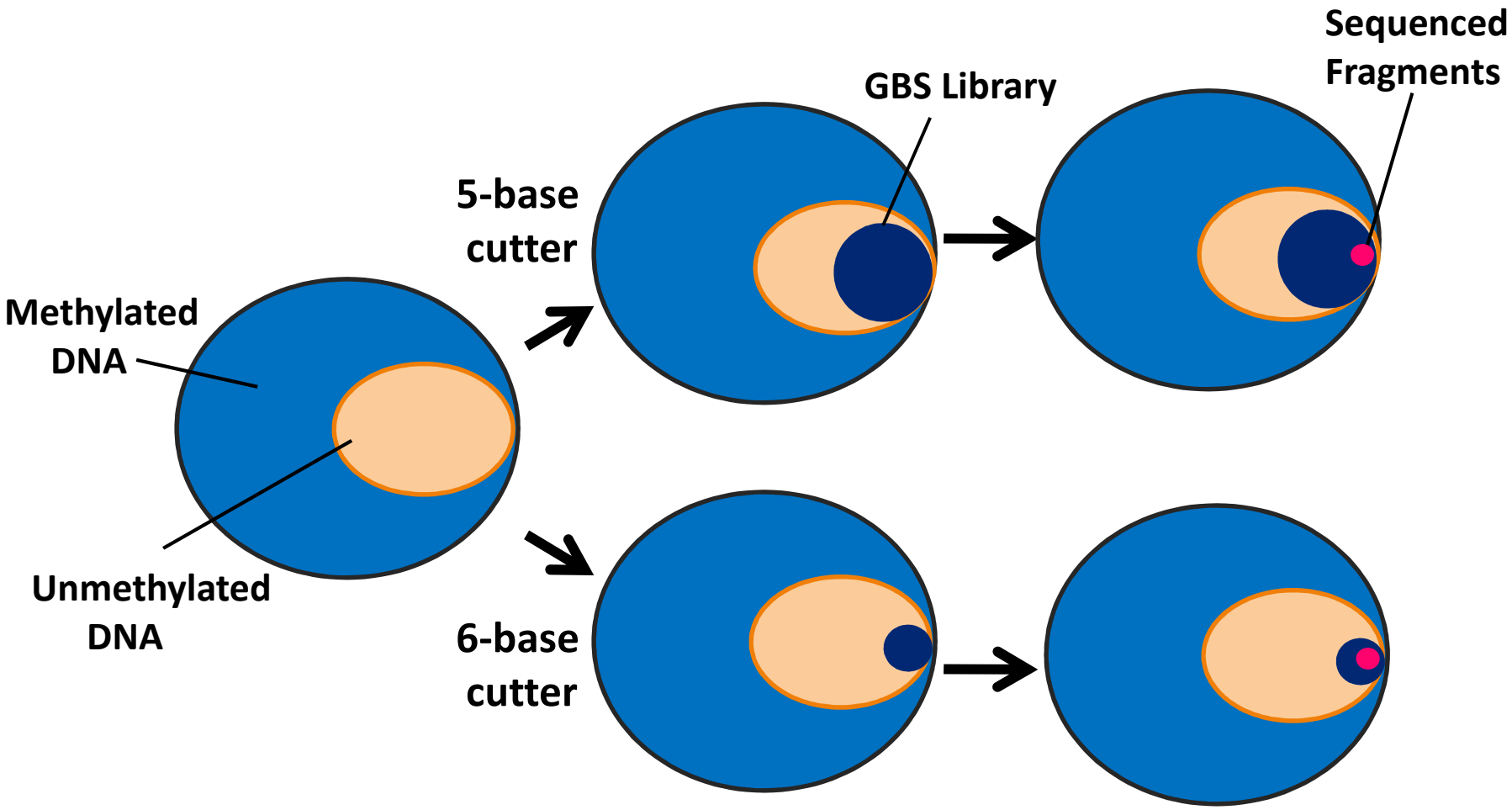
Dependent on Factors that Affect Diversity:

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

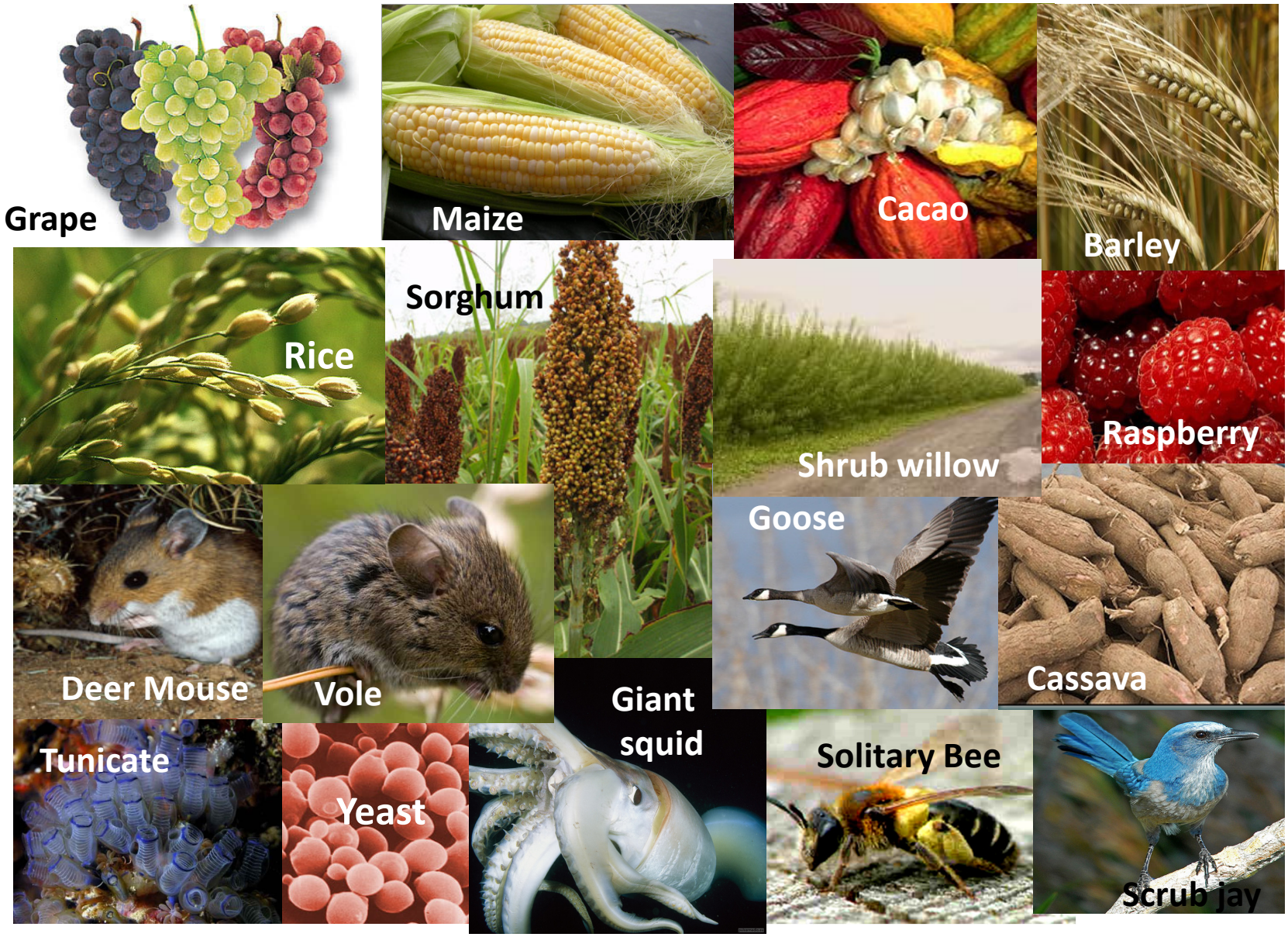
Other Factors

- **Genome size**
  - The size of the genome has some bearing on the number of fragments in the sequencing pool.
  - Amount of repetitive DNA directly correlated with genome size.
- **Genome composition**
  - The base composition of the genome can affect the frequency and distribution of the cut SNP s.
  - How repetitive DNA is organized in the genome affects library profiles.

Sampling large genomes with methylation-sensitive restriction enzymes.



Optimizing GBS in New Species



Pine
Spruce

Conifers

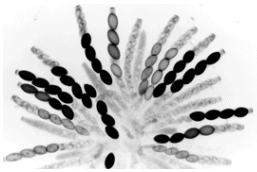
Maize
Sorghum
Rice
Barley
Switchgrass
Bracypodium
Pearl Millet
Teosinte
Lily
Andropogon
Fonio
Finger Millet

Strawberry
Ragweed
Silene
Sunflower
Safflower
Soybean
Goldenberry
Jatropha

Grape
Cassava
Cacao
Watermelon
Apple
Hop

Pepper
Cucumber
Squash
Pea
Gourd
Arabidopsis
Willow
Tea
Potato
Cherry
Flax

Flowering Plants







Neurospora
Verticillium

Solitary Bee
Corn Ear Worm
Plant Bug

Mexican Tetra
Killifish

Scrub Jay
Goose
Chickadee

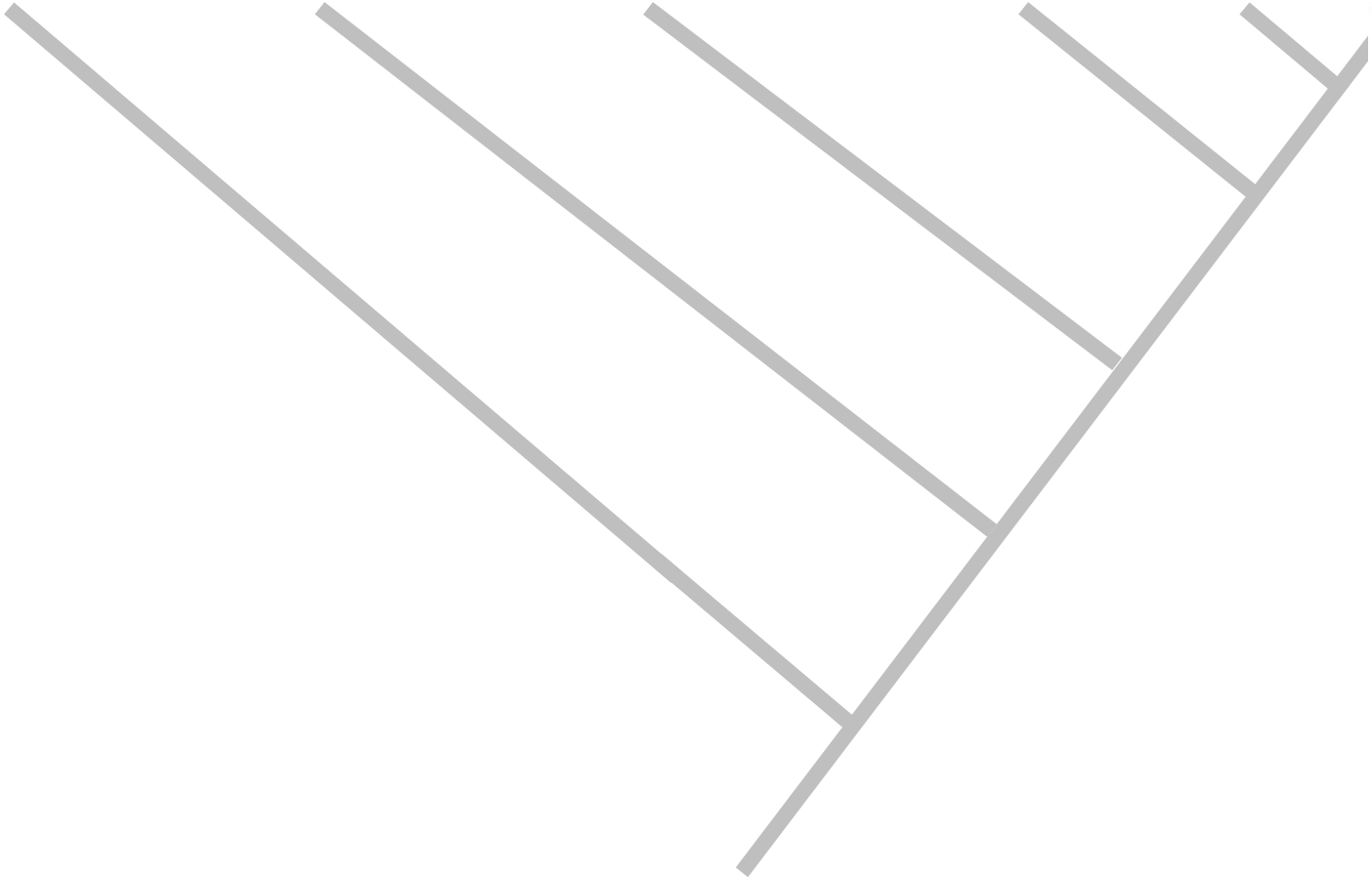
Deer Mouse
Vole

Cow

Pig

Killer Whale

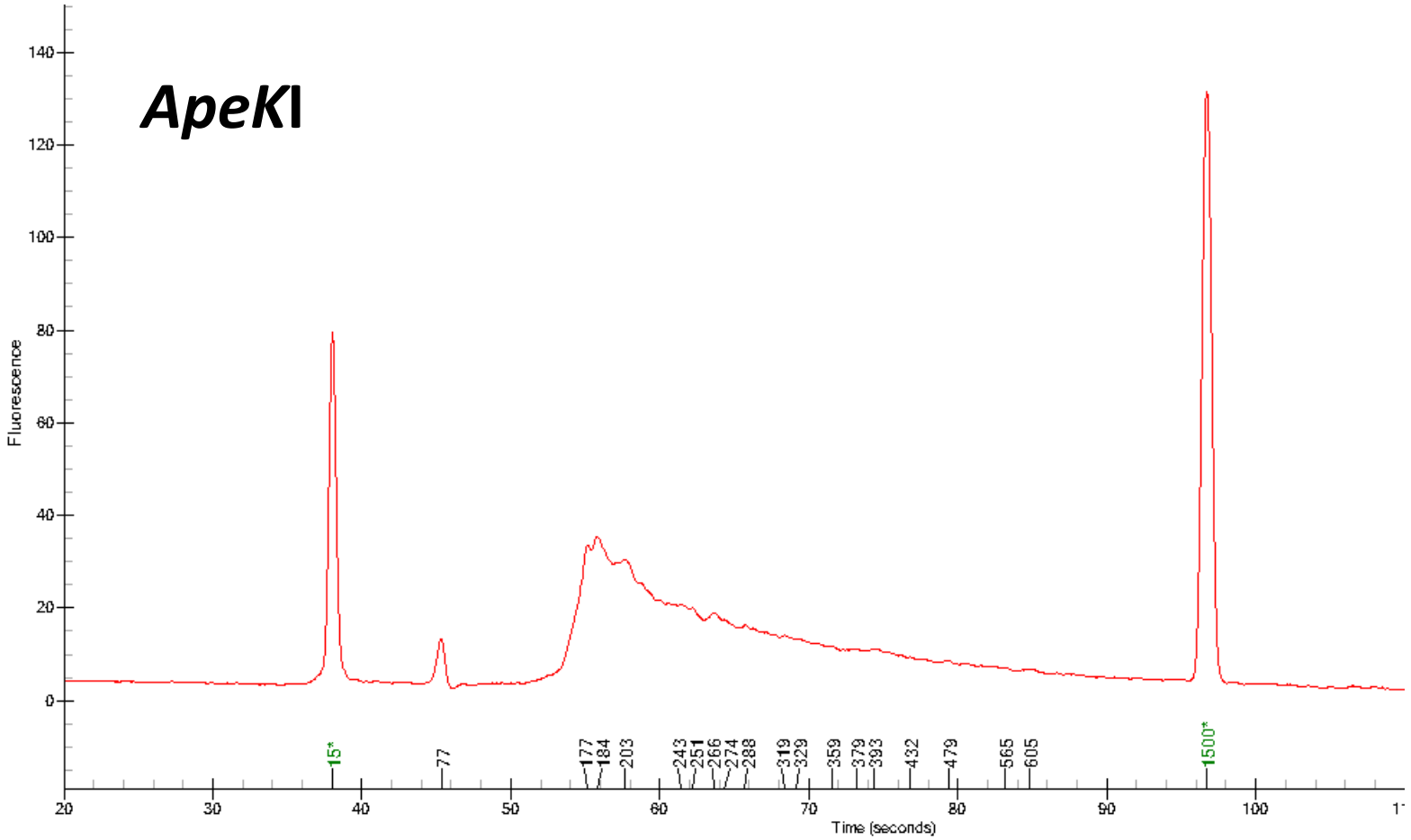
Fox



Choosing Appropriate Restriction Enzymes: Generalizations from the Bench

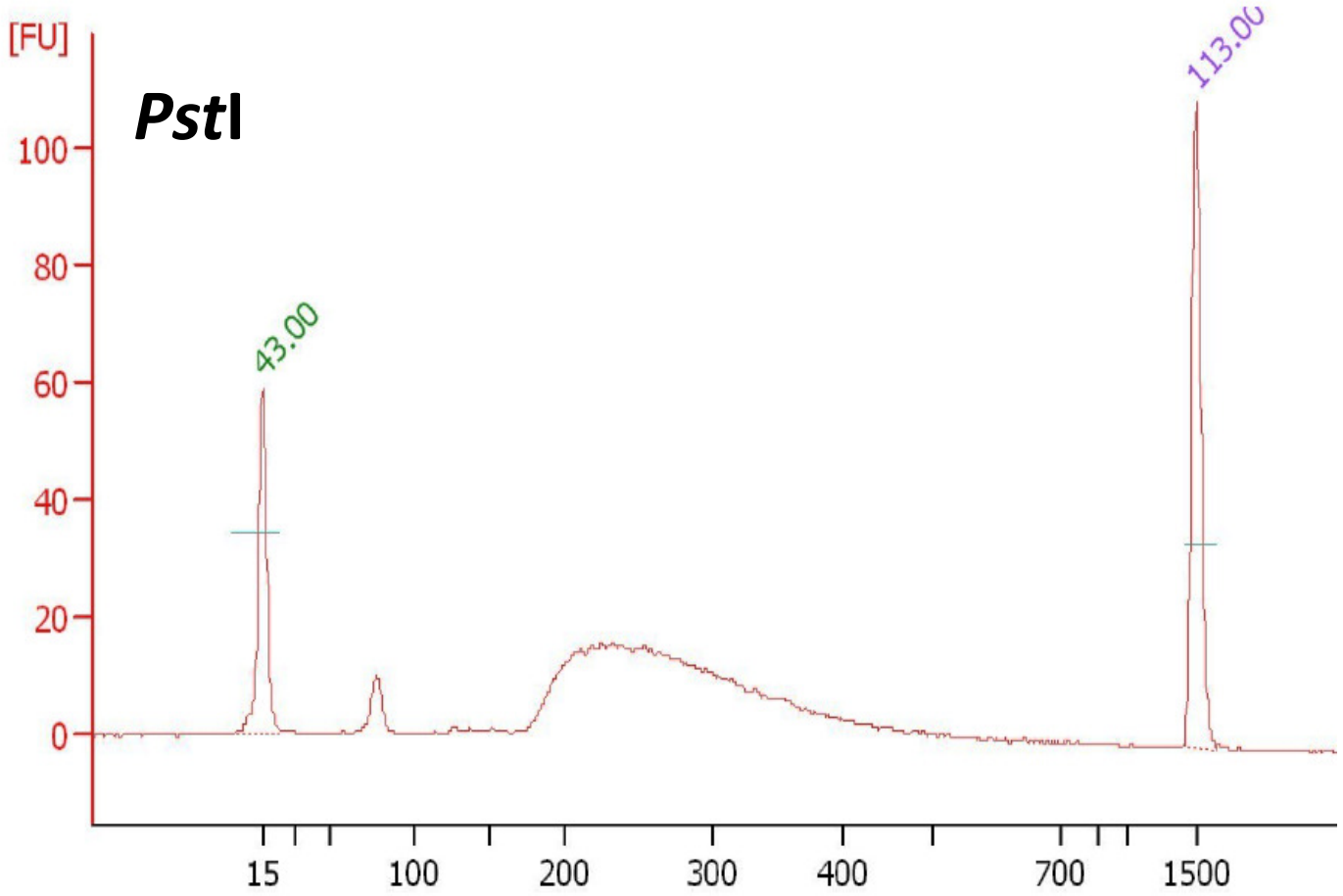
*Ape*KI works well for grasses.

Maize, sorghum, teosinte, rice, barley, millet, switchgrass, brachypodium.



*Pst*I works well for most mammals.

Deer mouse, vole, cow, pig.



Most frequently asked question for new species:

How many SNPs will I get?

Answer: **It depends……**

- **Genome size and expected heterozygosity affects size of fragment pool for desired amount of sequence coverage (enzyme choice and multiplex level).**
- **Amount of extant diversity and how well your sample reflects that diversity.**
- **Reference genome sequence? 3-4X more SNPs attained by aligning to a reference sequence.**

How many SNPs will I get?

Species	Genome Size (Mb)	Enzyme	Sample Size	No. SNPs
Maize	2,600	*Ape*KI	33,000	1,200K
Grape	500	*Ape*KI	1000	200K
Cow	3,000	*Pst*I	48	64K
Rice	400	*Ape*KI	850	60K
Pine*	16,000	*Ape*KI	12	63K
Vole*	3,400	*Pst*I	283	53K
Willow*	460	*Ape*KI	459	23K
Fox*	2,400	*Eco*T22I	48	16K
Verticilliflorum (fungus isolates)	40	*Ape*KI	2	10K

*No reference genome. UNEAK analysis pipeline used for analysis. To avoid homology/paralogy issues this pipeline calls SNPs very conservatively.

SNP calls in *Sorghum bicolor*- Lots of Missing Data

	alleles	Taxa1	Taxa2	Taxa3	Taxa4	Taxa5	Taxa6	Taxa7	Taxa8	Taxa9	Taxa10	Taxa11	Taxa12	Taxa13	Taxa14	Taxa15	Taxa16	Taxa17	Taxa18	Taxa19
SNP 1	C/A	A	N	A	C	A	N	N	C	C	N	N	N	N	N	N	N	N	C	N
SNP 2	A/C	C	N	C	A	C	N	N	A	A	N	N	N	N	N	N	N	N	A	N
SNP 3	T/C	C	N	C	T	C	N	N	T	T	N	N	N	N	N	N	N	N	T	N
SNP 4	C/G	N	N	C	N	G	N	N	N	N	N	N	N	N	N	C	N	N	N	N
SNP 5	T/C	T	T	N	N	T	N	T	N	N	N	N	C	N	N	N	N	N	N	N
SNP 6	C/T	C	C	N	N	C	N	C	N	N	N	N	T	N	N	N	N	N	N	N
SNP 7	G/A	G	A	G	A	G	N	G	G	N	N	G	G	G	R	N	A	N	R	G
SNP 8	G/A	G	N	N	A	G	A	G	N	N	G	G	G	N	N	A	N	G	G	G
SNP 9	T/C	T	C	T	C	T	C	T	N	N	T	T	N	N	N	C	N	N	T	T
SNP 10	T/C	T	C	N	N	N	N	N	T	T	N	N	N	N	N	N	N	N	N	N
SNP 11	G/A	G	N	N	A	G	N	N	N	N	N	N	N	G	A	A	N	G	G	N
SNP 12	G/A	G	A	N	A	G	A	N	N	N	G	N	G	N	N	N	N	G	G	N
SNP 13	G/A	G	A	N	A	N	N	N	G	N	N	G	N	N	N	N	N	G	N	G
SNP 14	T/G	G	T	N	T	G	N	G	N	N	G	G	N	G	T	T	T	N	G	N
SNP 15	C/T	T	C	N	C	T	N	T	N	N	T	T	N	T	C	C	C	N	T	N
SNP 16	G/A	G	A	G	N	G	A	G	G	G	N	G	N	G	A	A	N	G	G	G
SNP 17	C/G	N	G	C	S	C	G	C	C	C	N	C	C	N	C	G	G	N	C	N
SNP 18	G/A	N	A	G	A	G	A	G	G	N	N	G	G	G	N	A	A	G	G	N
SNP 19	C/T	C	N	N	T	N	N	N	N	C	N	C	N	N	N	N	N	N	C	C
SNP 20	T/G	T	N	T	N	T	G	T	N	T	T	T	T	T	N	N	G	T	N	T
SNP 21	T/G	T	G	T	G	T	G	N	T	T	T	T	T	T	N	G	G	T	T	T
SNP 22	G/A	N	A	N	A	N	N	G	G	G	N	G	N	N	N	A	N	N	N	N
SNP 23	G/T	G	T	G	T	G	N	N	G	G	G	N	N	N	N	N	T	G	G	G
SNP 24	C/T	C	T	C	T	C	N	N	C	C	C	N	N	N	N	N	T	C	C	C
SNP 25	T/C	T	C	T	C	T	C	N	N	T	T	N	N	T	N	N	N	T	N	N
SNP 26	C/A	N	N	N	N	A	N	A	N	N	N	N	N	N	N	N	N	N	N	N
SNP 27	G/A	N	N	N	N	A	N	A	N	N	N	N	N	N	N	N	N	N	N	N
SNP 28	C/T	N	N	C	N	N	N	N	C	T	C	N	N	C	T	N	N	C	N	C
SNP 29	T/C	T	N	T	N	N	N	N	N	N	N	N	N	N	N	N	N	N	T	T
SNP 30	G/T	G	N	G	N	N	N	N	G	G	G	N	N	N	N	N	N	N	N	N
SNP 31	A/T	A	T	N	N	A	N	N	A	A	A	N	A	A	N	T	T	A	A	A
SNP 32	A/T	A	T	A	T	N	N	N	A	N	A	N	A	N	N	T	T	A	N	A
SNP 33	C/T	N	N	C	N	C	N	C	C	C	N	N	N	C	T	N	N	C	N	C
SNP 34	C/T	C	T	C	T	C	N	N	C	N	C	C	N	C	T	T	N	C	C	N
SNP 35	A/C	A	C	A	N	A	N	A	A	A	A	A	A	A	C	C	N	A	A	A
SNP 36	T/C	T	C	T	C	T	C	N	T	T	T	T	N	T	N	C	N	T	T	T

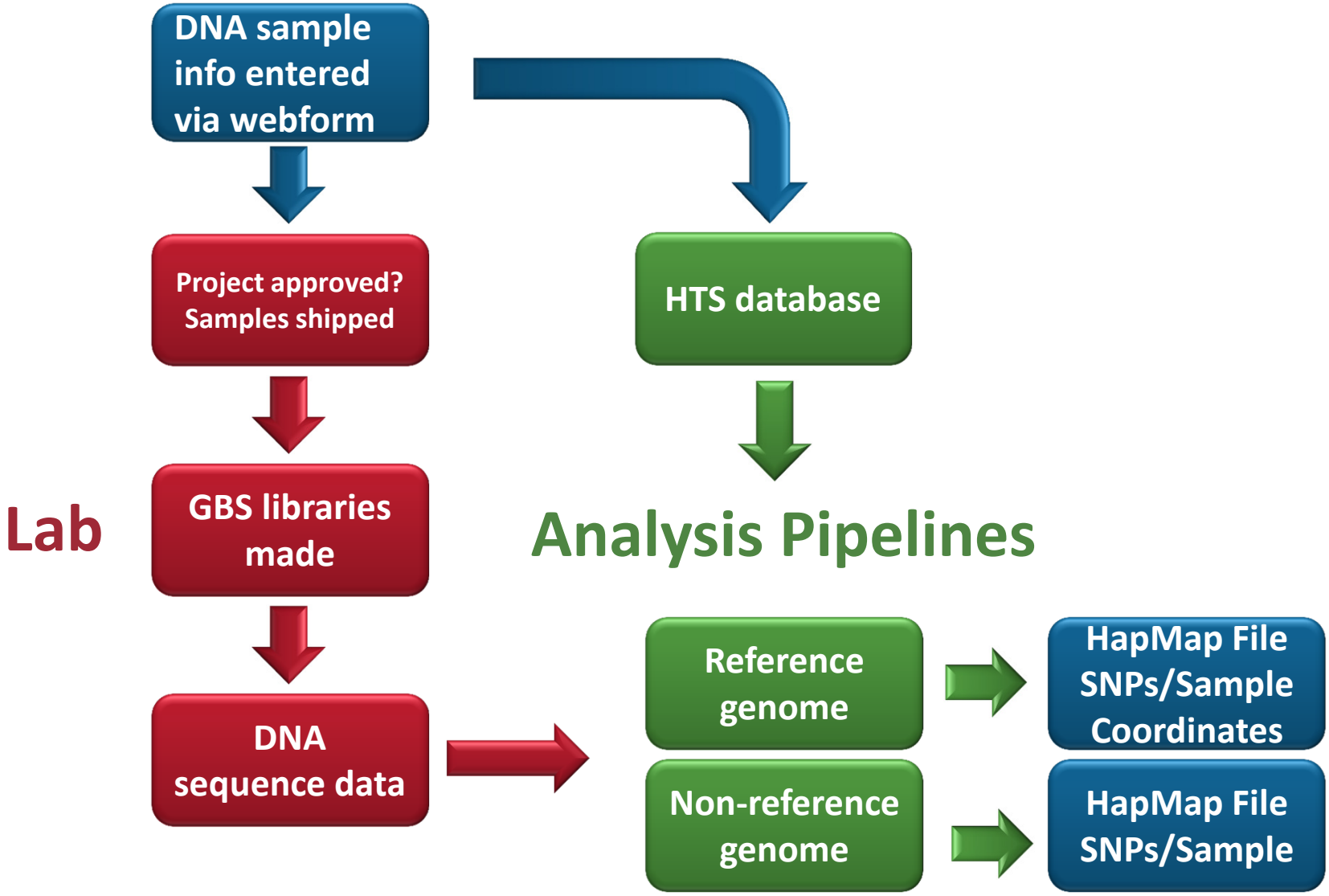
Filtering SNPs to remove most of the missing data.

Will be covered later in discussion of
TASSEL (http://www.maizegenetics.net/)

Missing Data Strategies

- **Impute Missing SNPs.**
  - **Many algorithms for doing this.**
- **Technical Options**
  - **Reduce the multiplexing level**
  - **Sequence the same library multiple times**
- **Molecular Options**
  - **Choose less frequently cutting enzymes**

GBS workflow at IGD



http://www.igd.cornell.edu/index.cfm/page/projects/GBS.htm

GBS Team

Method Development

Rob Elshire

Ed Buckler

Sharon Mitchell

Laboratory/Production

Charlotte Acharya

Wenyan Zhu

Lisa Blanchard

Shane Cieri

Bioinformatics

Jeff Glaubitz

Qi Sun

Katie Hyma

Fei Lu

Workshop Coordinator

Theresa Fulton