RADSEQ/GENOTYPING BY SEQUENCING WORKSHOP

Shichen Wang, Bioinformatics Scientist
Genomics and Bioinformatics Service
Texas A&M University AgriLife Research

11.29.2017

Genomics and Bioinformatics Service

Providing Genomics and Bioinformatics Services to the Texas A&M System, Texas, and the World

"How do I start a new sequencing project?"

Questions related to your samples, or the submission process

Search

Welcome

Who We Are

News

Research Highlights

Bioinformatics

Personnel

Publications

FAQ

Contacts

OUTLINE

- What is RADseq/Genotyping By Sequencing (GBS)
  - different protocols
    - Dr. Retz's protocol
  - experiment design

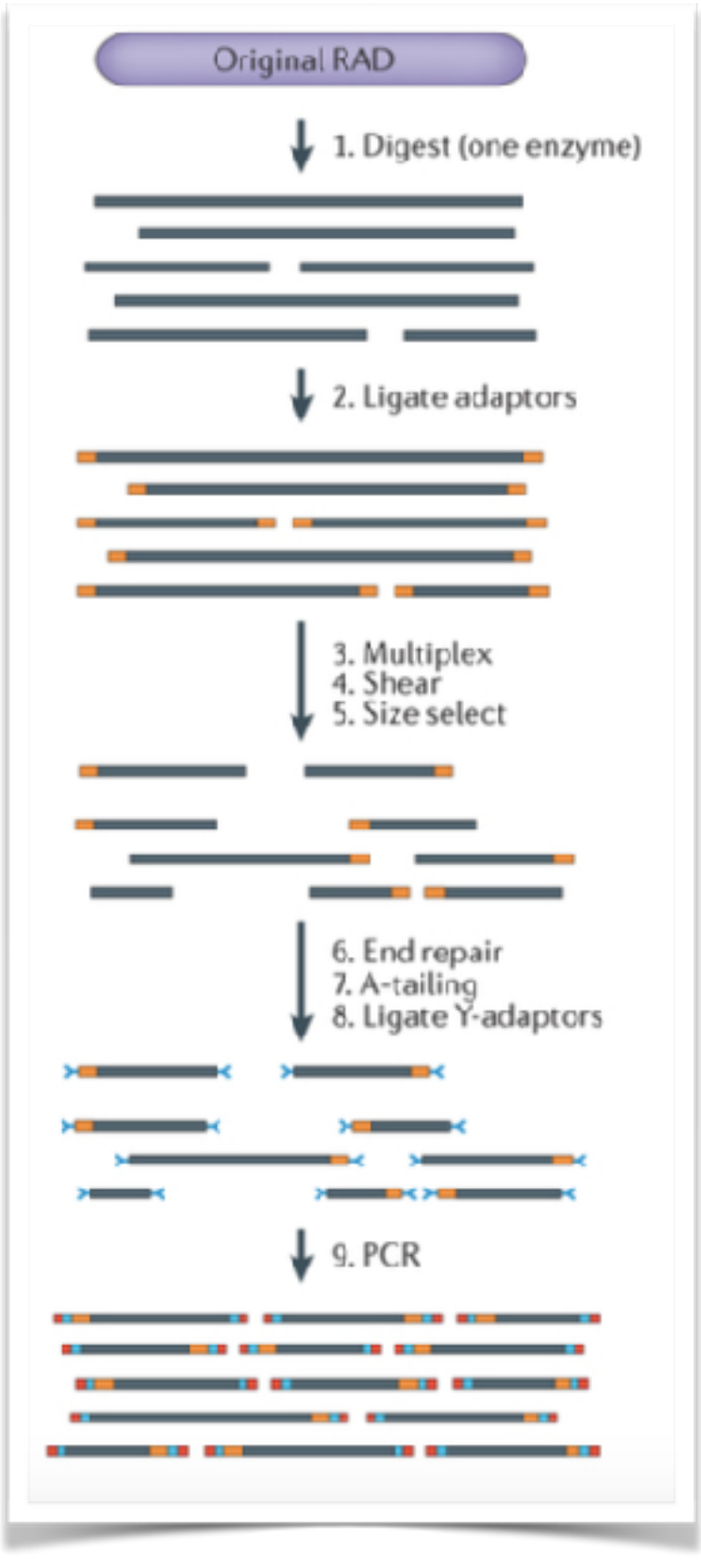
OUTLINE

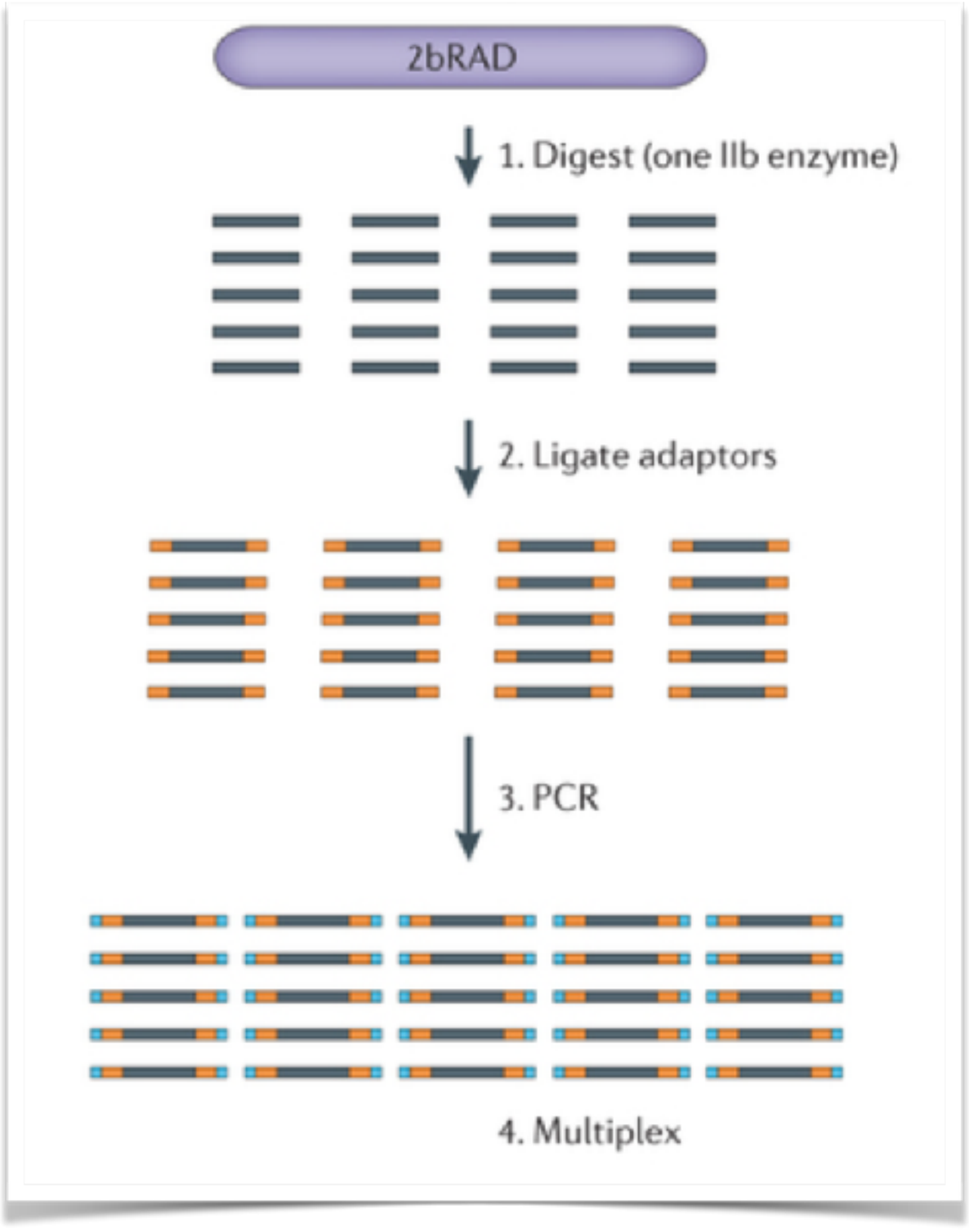
- Examples of RADseq/GBS applications
  - 1. For genome-wide variation discovery;
  - 2. Genetic Maps, QTL mapping
  - 3. Genome wide association study (GWAS), Genomic Selection (GS)
  - 3. Population genetics study
- Data analysis
  - general approach for NGS data analysis
  - Several GBS pipelines
- Hands on exercises

RADSEQ/GENOTPING BY SEQUENCING

- History
  - 2007, RAD markers
    - Miller et al. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers: RAD tags with microarray
    - Van et al. Complexity Reduction of Polymorphic Sequences: RAD tags sequencing with 454 sequencer. Genotyping using Keygene SNPWave (patent application)
  - 2008, RADseq
    - Baird et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers (1118 citations): RAD tags sequencing with Illumina platform.
- GBS, ddRAD, 2-b RAD, bsRADseq
  - Elshire et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.
  - Poland et al. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach.
  - Peterson et al. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species
  - Wang et al. 2b-RAD: a simple and flexible method for genome-wide genotyping
  - ....

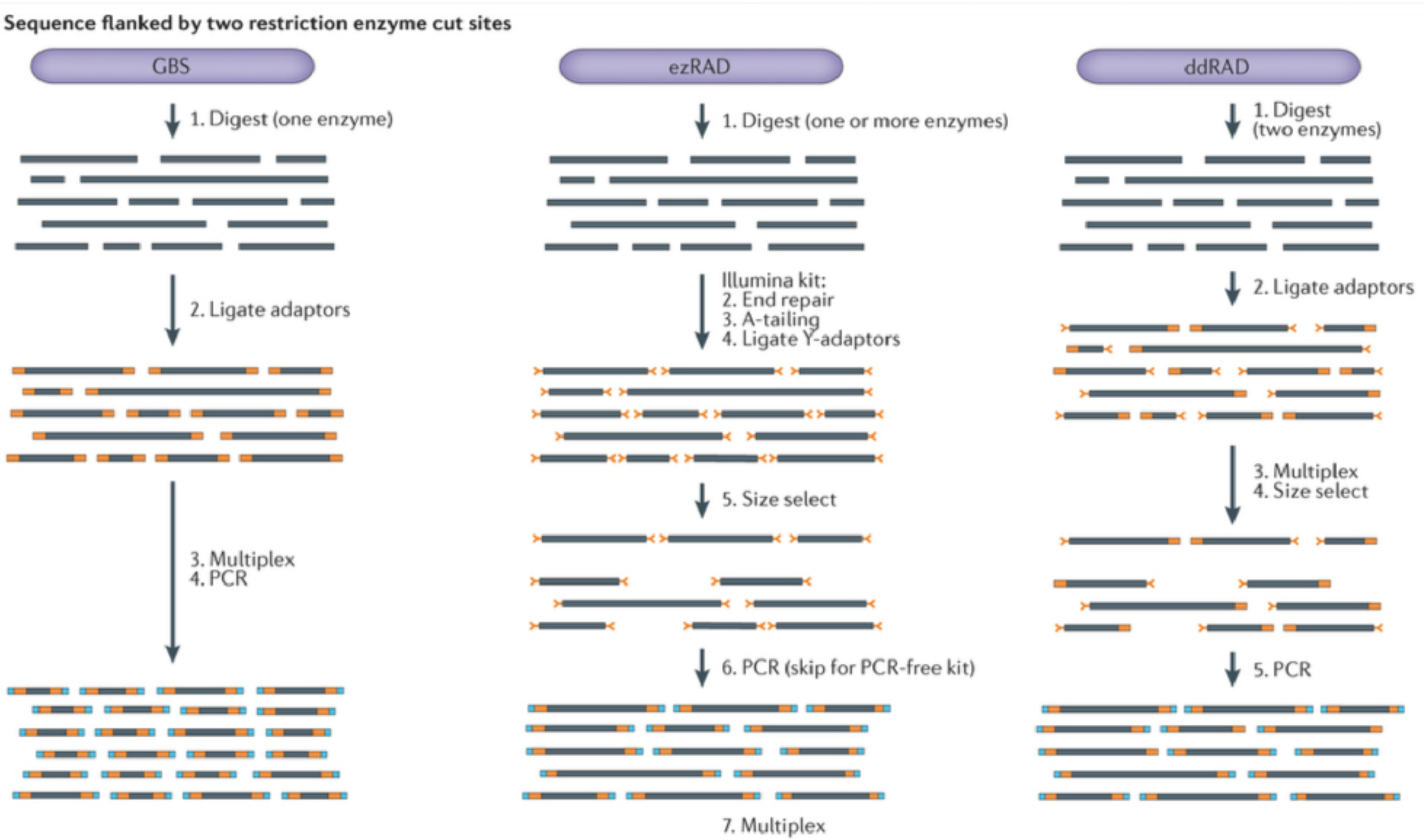
RADSEQ/GBS PROTOCOLS:



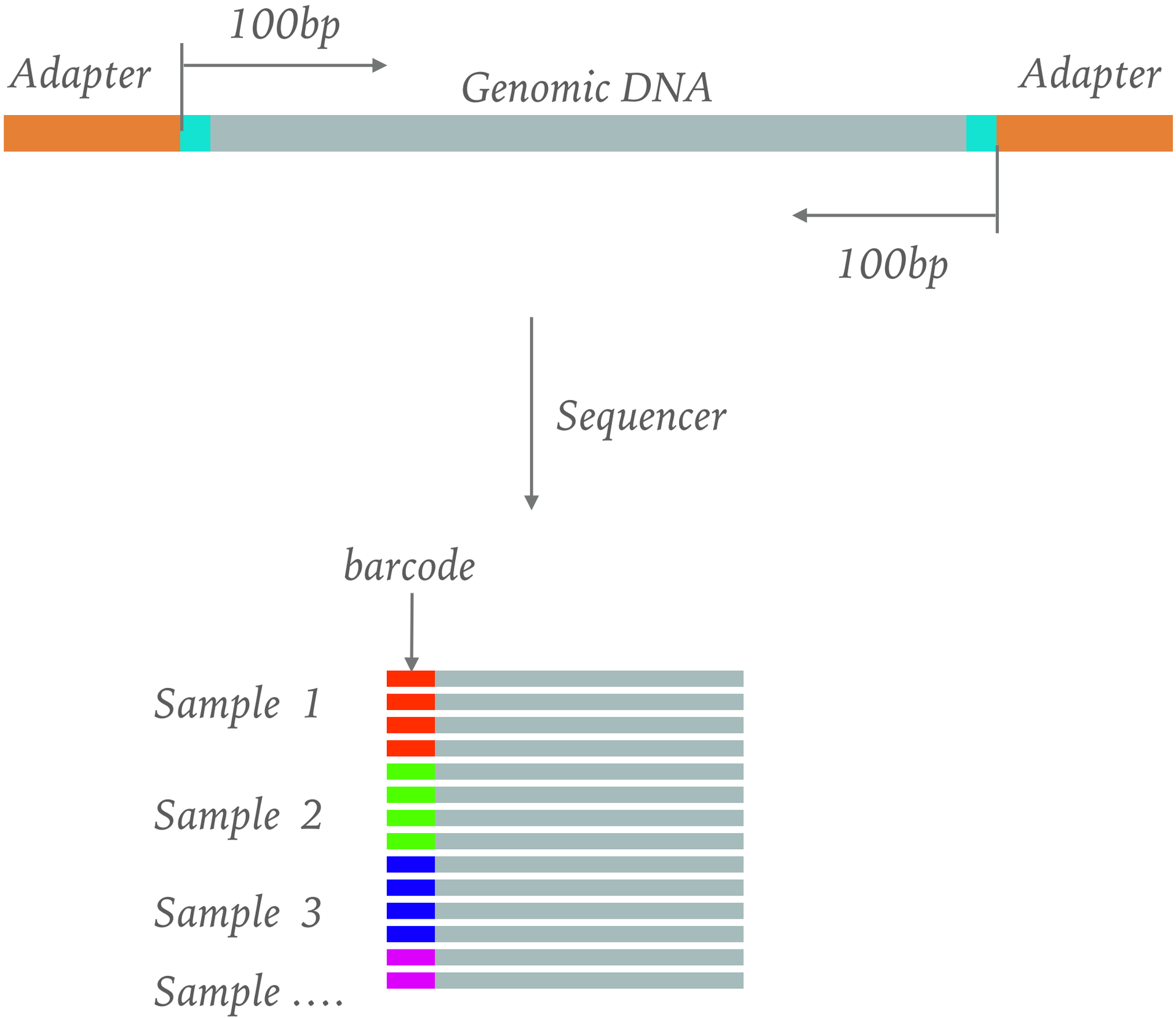


Nature Reviews Genetics **17**, 81–92 (2016) doi:10.1038/nrg.2015.28

PROTOCOLS:

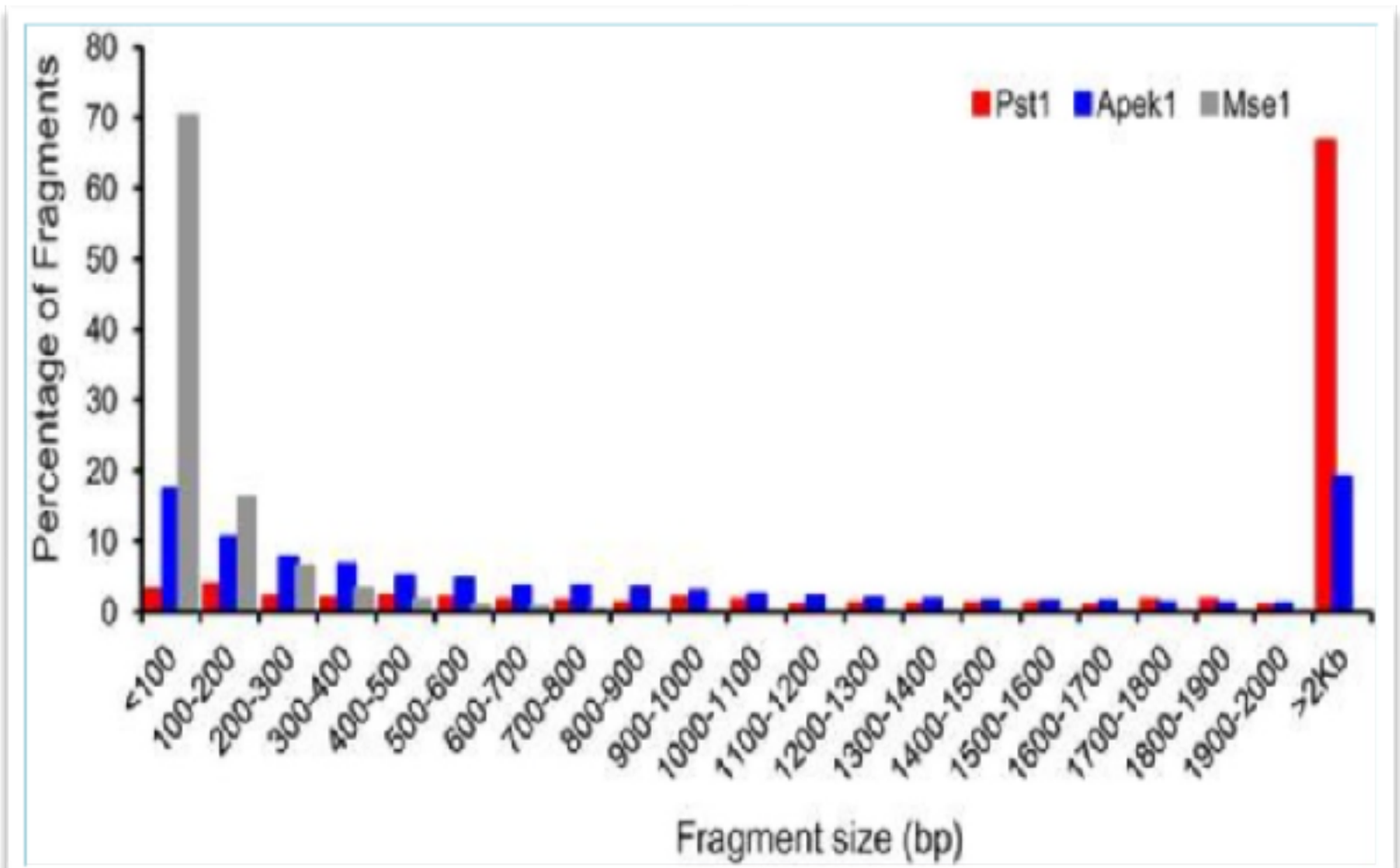


Nature Reviews Genetics **17**, 81–92 (2016) doi:10.1038/nrg.2015.28


100bp
Adapter
Genomic DNA
Adapter
100bp
Sequencer
barcode
Sample 1
Sample 2
Sample 3
Sample ....

RADSEQ/GBS EXPERIMENT DESIGN

- Choose the protocol
  - Pick the most suitable restriction enzyme(s) for the targeting genome
    - Methylation-sensitive restriction enzymes
    - Double digestion



In silico digestion genome coverage for fragment between 100-600 bp

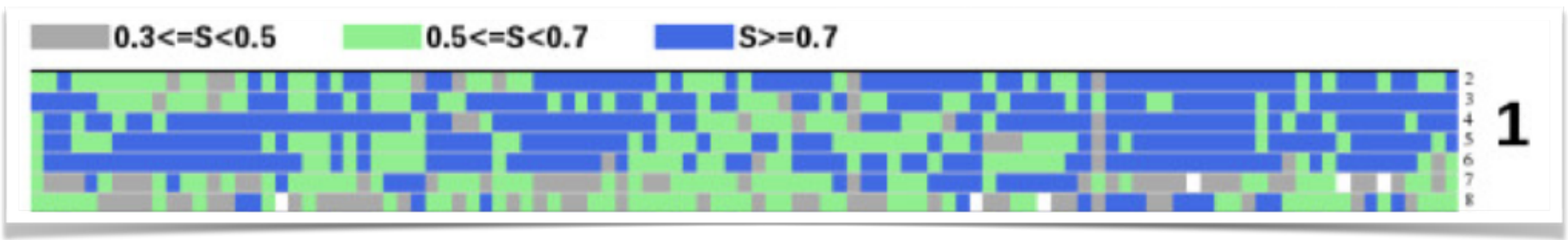
	Plant								Animal									
Enzyme pair	Arabidopsis	Cottonwood	Medicago	Winegrape	Soybean	Rice	Sorghum	Maize	C. elegans	Fruit fly	Honey bee	Stickleback	Pike	Zebra fish	Turkey	Zebra finch	Dog	Housecat
genome size (Mbp)	120	379	297	426	950	382	659	2060	83	158	220	401	377	1340	1040	1021	2328	2419
CviAII+HinfI	35.1	32.5	27.9	31.4	32.5	31.1	30.8	33.8	31.9	30.1	24.6	35.0	31.0	31.0	30.8	34.5	35.9	35.5
CviAII+DdeI	34.4	30.8	26.6	30.6	30.9	30.3	31.8	32.9	27.0	29.7	16.1	35.1	32.9	33.3	31.9	35.9	35.3	34.9
BfaI+HinfI	32.5	32.5	26.0	30.9	31.4	32.0	31.8	34.6	27.9	24.5	17.5	24.9	30.2	27.9	29.3	31.8	36.3	35.2
BfaI+DdeI	32.5	31.4	25.0	30.4	30.4	31.2	32.3	34.2	23.8	24.0	11.8	25.1	30.8	29.1	28.7	30.9	33.6	32.9
CviAII+TfiI	33.4	27.9	24.3	27.2	27.7	25.8	26.2	27.6	29.3	25.8	24.5	26.6	23.2	23.7	24.7	26.6	28.4	29.7
BfaI+TfiI	31.8	29.1	23.6	27.9	28.2	27.0	27.3	29.1	25.9	21.4	17.8	20.5	24.0	22.4	24.4	25.8	30.0	31.8
MluCI+HinfI	25.3	18.0	14.1	19.8	19.2	24.9	28.1	29.1	13.5	20.7	14.1	33.0	28.4	24.2	25.1	26.6	29.6	29.4
MluCI+DdeI	24.4	17.3	13.6	19.4	18.6	24.0	27.5	29.0	10.2	19.6	7.4	33.3	30.1	26.0	26.5	28.4	28.6	29.4
CviAII+ApeKI	19.3	17.2	12.9	13.4	14.2	25.3	22.6	23.8	18.5	27.7	11.0	33.7	25.6	26.6	28.1	31.8	24.9	23.9
HinfI+MseI	28.1	22.0	18.7	24.7	24.3	28.0	28.8	30.6	27.0	7.9	4.3	10.4	9.8	21.6	8.4	28.3	28.8	28.7
HinfI+HpyCH4IV	29.4	22.7	21.6	18.0	23.1	27.9	27.2	27.9	30.5	14.3	11.6	16.1	14.8	24.8	11.7	16.0	19.4	22.7
DdeI+MseI	27.1	20.8	17.9	24.1	23.2	26.9	29.3	30.3	21.5	8.3	3.7	10.2	9.3	22.6	6.0	29.5	29.0	29.1
MluCI+TfiI	23.7	15.0	11.8	16.6	15.9	19.9	23.3	24.6	11.8	16.9	13.1	24.8	21.0	18.4	19.6	19.8	22.7	25.3
DdeI+HpyCH4IV	29.2	22.2	20.8	17.5	22.6	27.2	27.2	27.5	26.3	14.9	9.3	15.9	13.8	26.0	8.6	14.6	18.0	20.7
ApeKI+BfaI	19.5	18.3	13.1	14.9	15.1	26.3	23.3	24.3	16.8	12.6	5.9	13.8	14.7	25.2	13.5	29.8	26.6	23.9
TfiI+HpyCH4IV	29.0	20.7	19.7	16.8	21.5	24.2	24.2	23.7	28.3	13.9	12.3	15.2	13.2	20.2	10.9	13.3	16.8	20.7
NlaIII+MluCI	25.7	19.4	16.0	20.2	19.6	25.4	27.0	29.0	13.4	5.2	2.4	10.8	7.7	27.1	6.3	28.7	29.3	30.2
TfiI+MseI	26.5	19.0	16.3	21.5	20.9	23.6	24.3	26.3	24.5	7.8	5.0	9.7	8.5	16.0	7.7	21.4	22.2	23.9
CviAII+AvaII	15.6	14.0	12.3	14.5	14.6	18.7	20.8	24.7	12.4	15.9	7.0	22.9	19.6	13.4	14.5	17.7	20.4	20.9
ApeKI+MseI	17.2	14.1	9.9	13.2	12.8	25.1	22.9	23.2	15.8	7.2	3.8	10.0	9.5	19.5	7.0	26.7	21.8	20.4
AvaII+BfaI	16.1	15.3	12.3	15.8	15.5	20.6	21.8	26.2	11.3	9.0	3.9	11.6	12.3	13.0	9.0	17.3	22.5	21.4

SUMMARY

- Many protocols to be chosen from
- *In silico* digestion helps to pick the restriction enzymes
- Depends on the protocol, the subsequent data analysis might require adding/removing steps

RADSEQ/GBS APPLICATION EXAMPLES

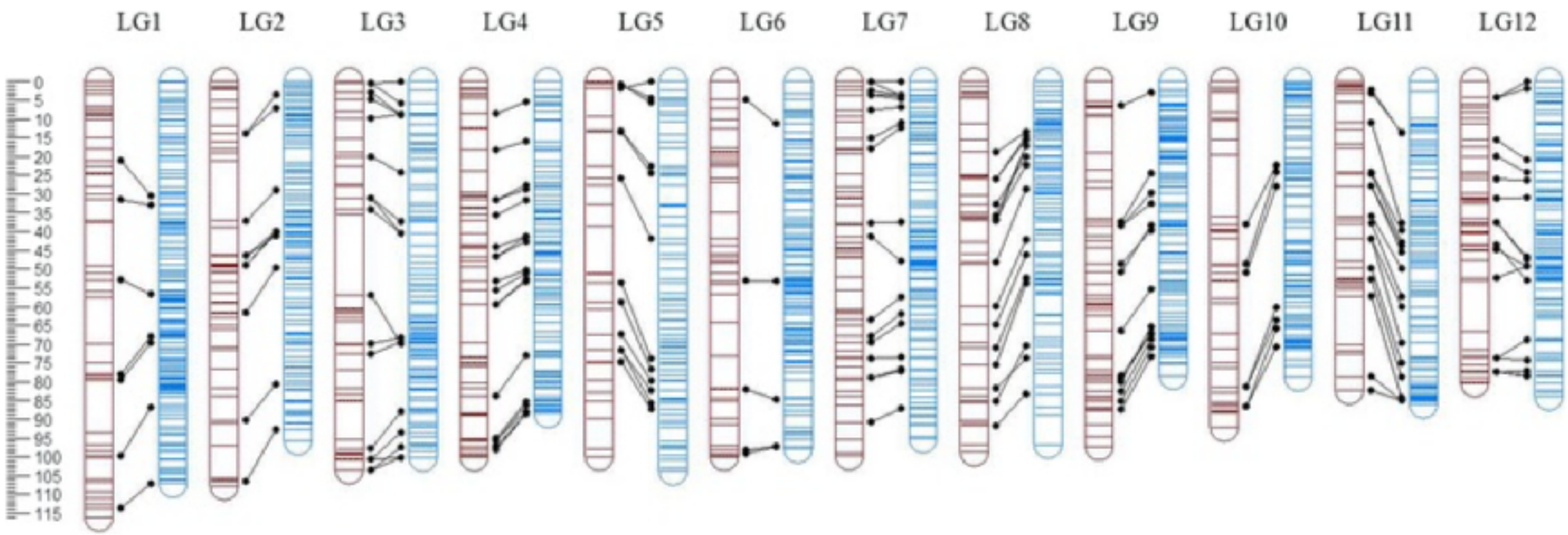
- Variation discovery
  - Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. BMC Genomics. 2011
    - Sequencing from libraries constructed to limit sequencing to start at defined restriction sites led to genotyping 10-fold more SNPs..



Haplotype sharing patterns

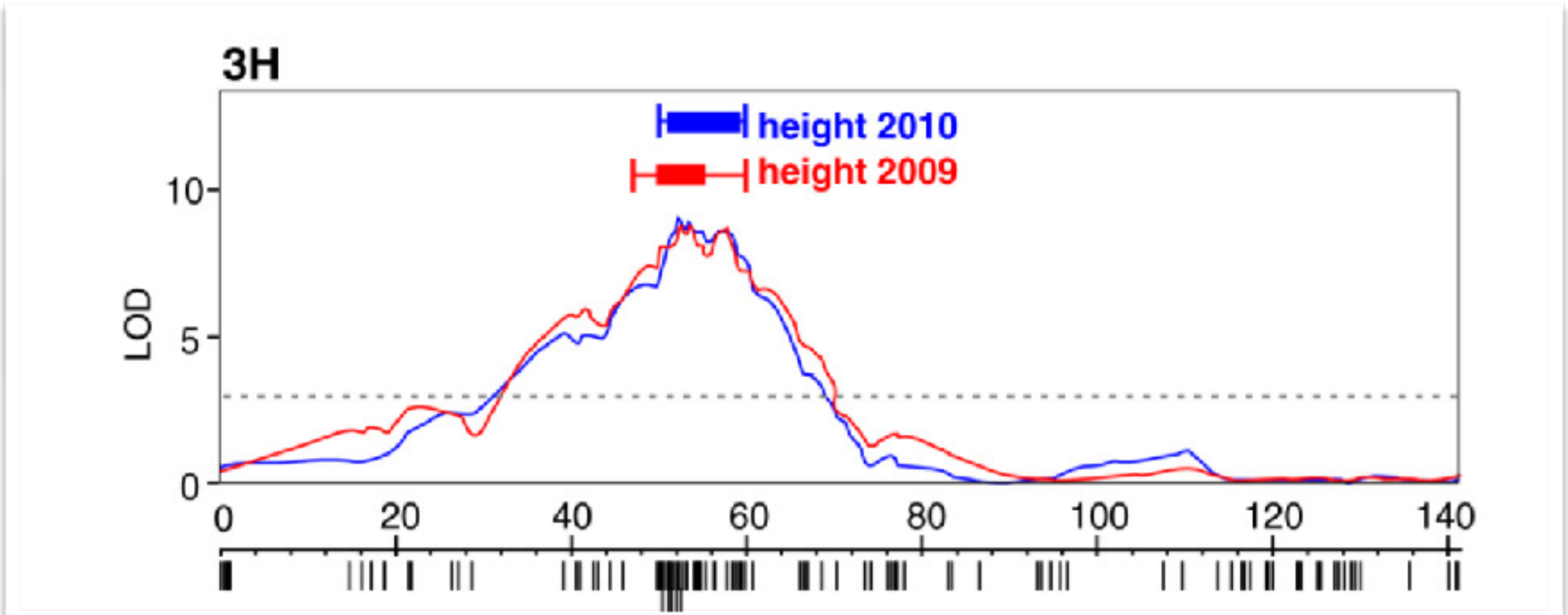
RADSEQ/GBS APPLICATION EXAMPLES

- Genetic Mapping
  - Exploiting genotyping by sequencing to characterize the genomic structure of the American cranberry through high-density linkage mapping, BMC Genomics. 2016
    - 10842 SNPs in total; 4849 markers were mapped.



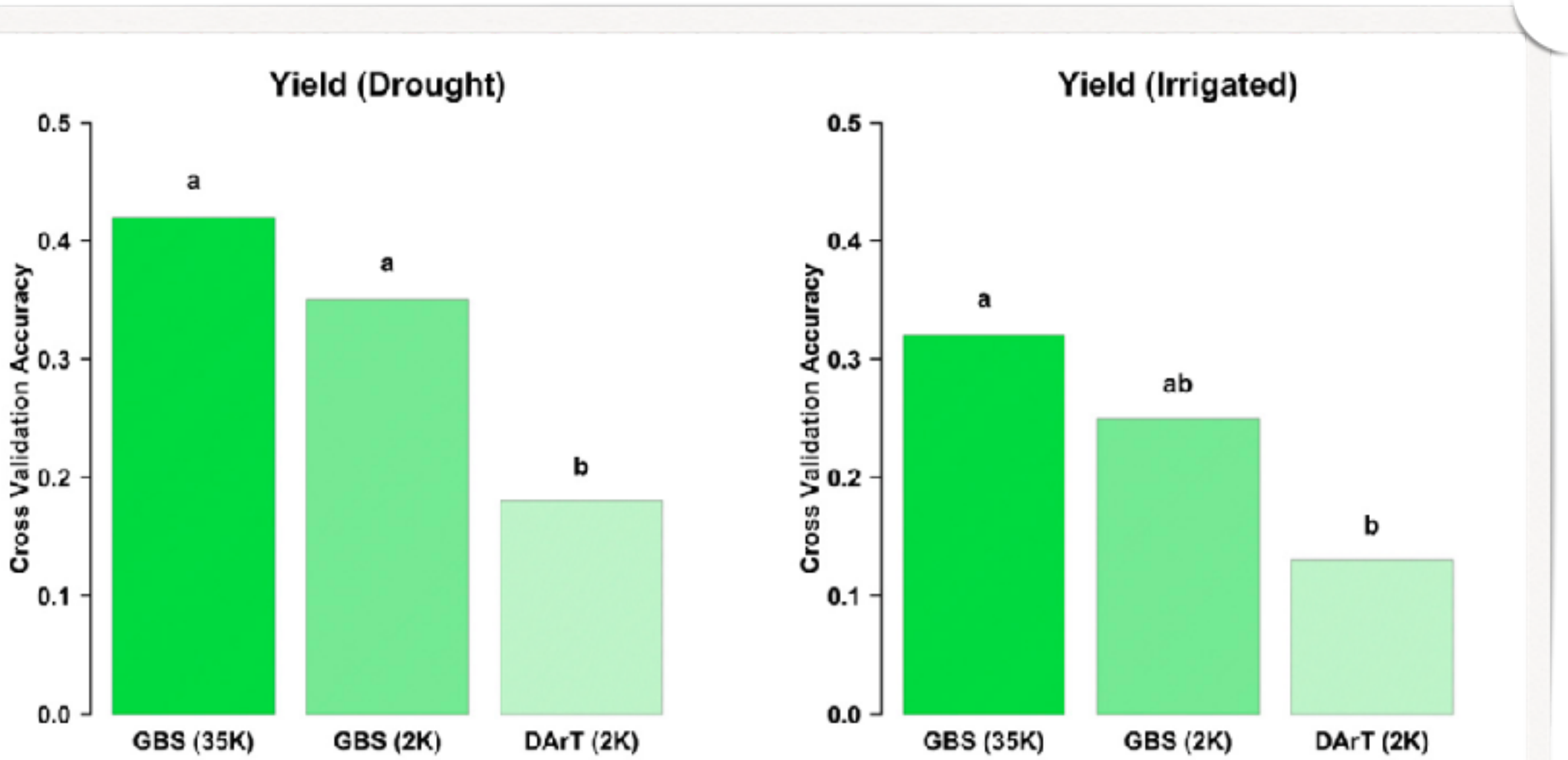
RADSEQ/GBS APPLICATION EXAMPLES

- QTL Mapping
  - An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley, BMC Genomics, 2014
    - In total, 461M categorized reads from the GPMx mapping population were mapped.… Using these highly conservative criteria, we identified an initial set of 1,949 co-dominant SNPs with robust allele calls across the population.



RADSEQ/GBS APPLICATION EXAMPLES

- Genomic Selection (GS)
  - GS uses genomewide molecular markers to predict complex, quantitative traits in animal and plant breeding.
  - Is GBS suitable for this task?



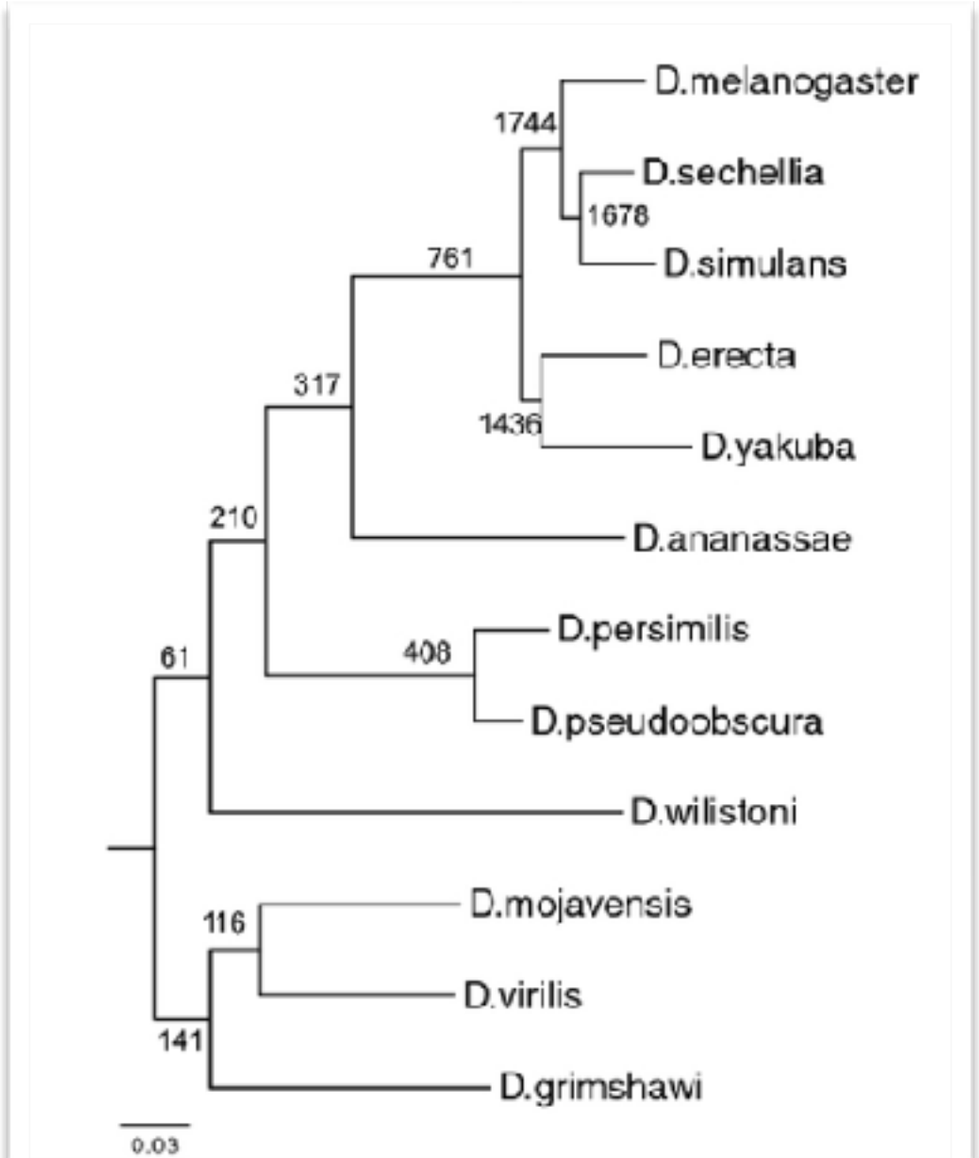
Jesse Poland, et al, 2012

RADSEQ/GBS APPLICATION EXAMPLES

- Phylogenetic inference
  - Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization, Ecology and Evolution, 2013
    - Inferring phylogenetic relationships between closely related taxa can be hindered by three factors: **(1) the lack of informative molecular variation at short evolutionary timescale; (2) the lack of established markers in poorly studied taxa; and (3) the potential phylogenetic conflicts among different genomic regions due to incomplete lineage sorting or introgression.**
    - **Clustering RAD-seq data using the BLASTN and SiLiX programs significantly improves the recovery of orthologous RAD loci compared with previously proposed approaches when comparing distant related species.**

RADSEQ/GBS APPLICATION EXAMPLES

- Phylogenetic inference
  - Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization, Ecology and Evolution, 2013

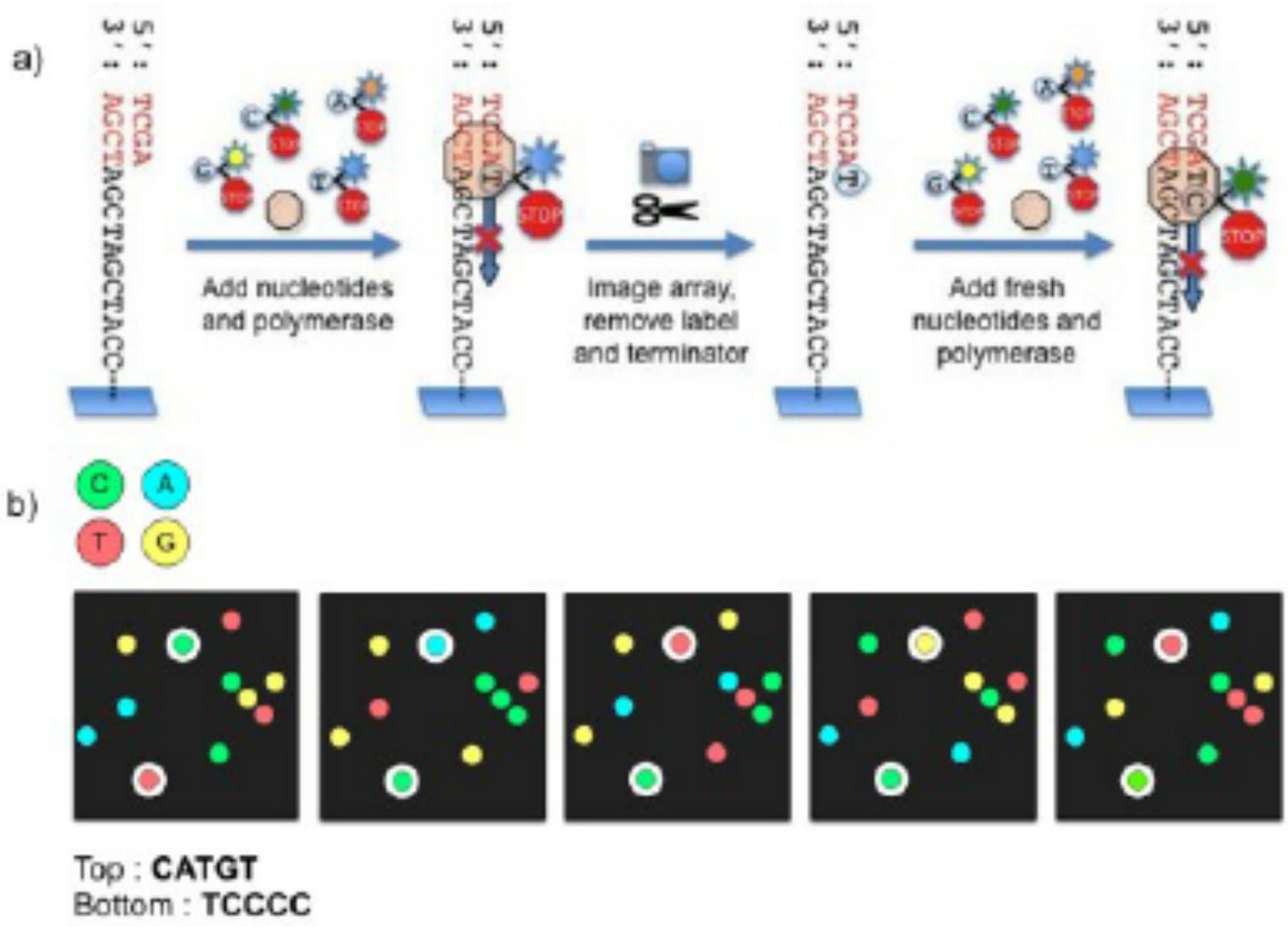


RAD-seq-based phylogeny of the 12 Drosophila species, based on 100-bp-long RAD-seq reads, inferred by maximum likelihood using PhyML 3.0.

?

RADSEQ/GBS DATA ANALYSIS

- What kind of data being generated?



Sequence by Synthesis, Anderson and Schrijver, 2010

RADSEQ/GBS DATA ANALYSIS

- What kind of data being generated?
  - fastq format
    - each read represented by 4 lines

line 1 `@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`
line 2 `TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATTTGTTGGGGGAG`
line 3 `+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`
line 4 `efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BB`

ATGTCGTGCAGACGATCGATGCTAGCTAGCTAGCTAGTCGATCGTAGCTAGCTAGCTAGCTAGCTGATCGATGCTAGCTAGCTAGC

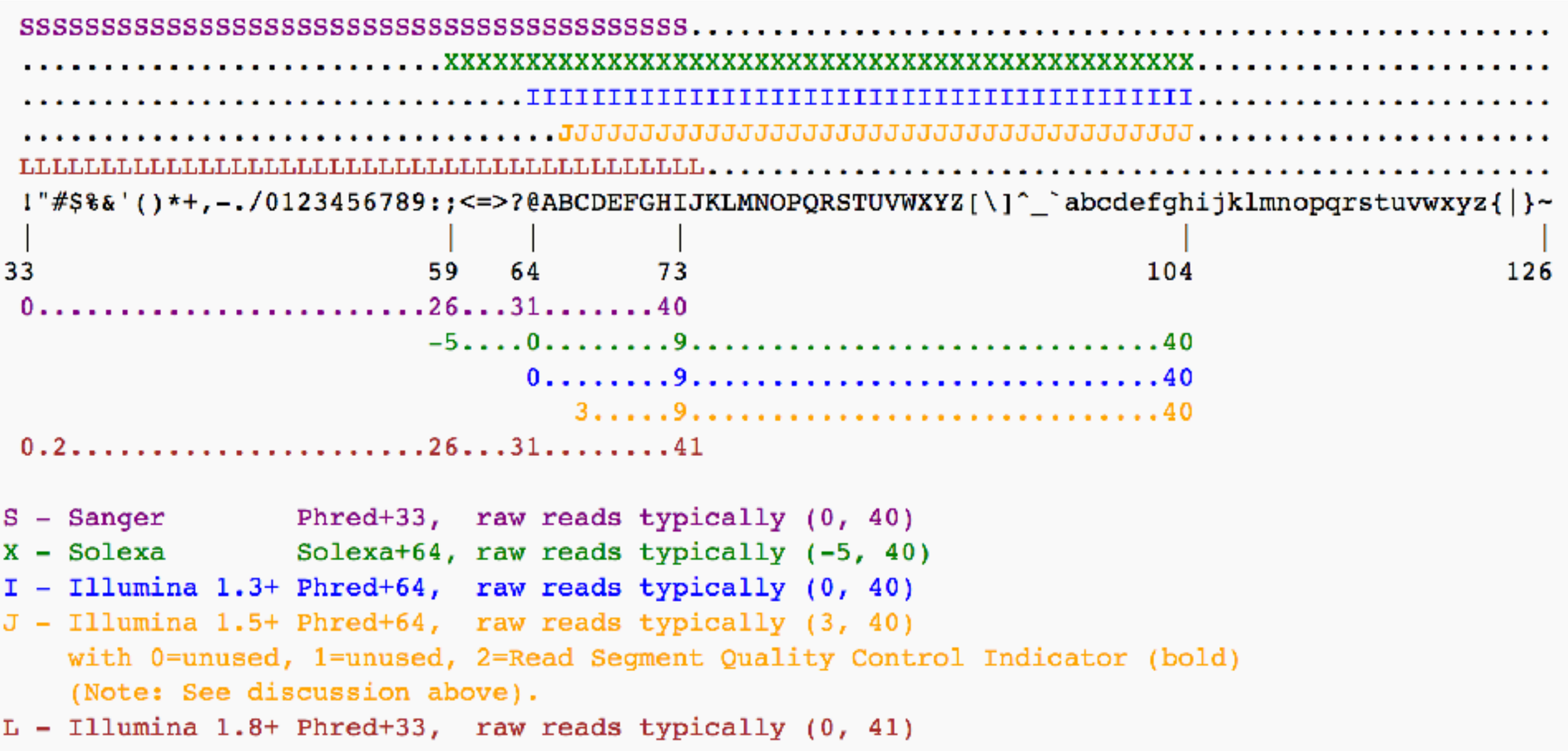
RADSEQ/GBS DATA ANALYSIS

- What kind of data being generated?
  - fastq format

```
[wangsc@login3 ~/fastq_]$ zcat VijayWSC4-Ind7_CAGATC_L004_R1_001.fastq.gz | head
@HISEQ:224:C8PEWACXX:4:1101:1269:1962 1:N:0:CAGATC
AAAAAGAACAAGATTGAGACTAGATCGTGAAGCGAATGTAAGCATACTCACCTTATACCATTCACCATCAAGTCCTGTTTTCCACACGGCTGTGATGACAT
+
BBBFFFFFFFFFFIIIIFBFFFIBBFFFB<0<FFFFFIFIIIFIFBFFFIIIBBFFIIIIFFIFFB<BFFFFF'BBFFFBBBBBBBBB'7B<B<B<BFBBB
@HISEQ:224:C8PEWACXX:4:1101:1835:1973 1:N:0:CAGATC
CTTGTTTAATACTTGCATGCATGTAGTCATAATGACAGTGTGCTACTTTCTTCTCATTCCTTCCTTGCATGCATGCAGACGTATTAATGATTCCAGATAAC
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFIFFIIIIFFFFFFFFFFFFFFFFFFFB
@HISEQ:224:C8PEWACXX:4:1101:2022:1963 1:N:0:CAGATC
GTTTATTGGTTGCTTGGCCTCTCTCAGCTGGTTCCTCAGCCGGCTGGGGGAGAGCGCCCTGCCCCTATATCAGCGGATGAAGAAGAAGACAACATTCGAGT
[wangsc@login3 ~/fastq_]$
[wangsc@login3 ~/fastq_]$
[wangsc@login3 ~/fastq_]$ zcat VijayWSC4-Ind7_CAGATC_L004_R2_001.fastq.gz | head
@HISEQ:224:C8PEWACXX:4:1101:1269:1962 2:N:0:CAGATC
CAACCCTCACTTGTCTGACATGGAAGAGNNNNNNNNNNNNNNNTTNNNNNNNNNNNNNNNNNNGNCNNANNNNNNNNNNGNNNNNCNNNNNNNNNNGCANCAN
+
BBBFB7FFFFFFFFFIIIBFFFFFB0BBF##########################################################################
@HISEQ:224:C8PEWACXX:4:1101:1835:1973 2:N:0:CAGATC
ACAAACTCAAATTACAGGTACCTAGGTAAATTTAATGACTACTTTAGAAGCCAACTTTTCTTCTTGTTATCTGGAATCATTAATACGTCTGCATGCATGCA
+
BBBFFFFFFFFFFIIIIIIFIIIIIIFFIFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFBFFFFFFFFF<
@HISEQ:224:C8PEWACXX:4:1101:2022:1963 2:N:0:CAGATC
TTGAGGTCGTGGAAAGCTTCGTCCGGCTGGTCGTTCAACTCGAATGTTGTCTTCTTCTTCATCCGCTGATATAGGGGCAGGGCGCTCTCCCCCAGCCGGCT
[wangsc@login3 ~/fastq_]$
```

RADSEQ/GBS DATA ANALYSIS

- What kind of data being generated?
  - fastq format



RADSEQ/GBS DATA ANALYSIS

- What kind of data being generated?
  - fastq format

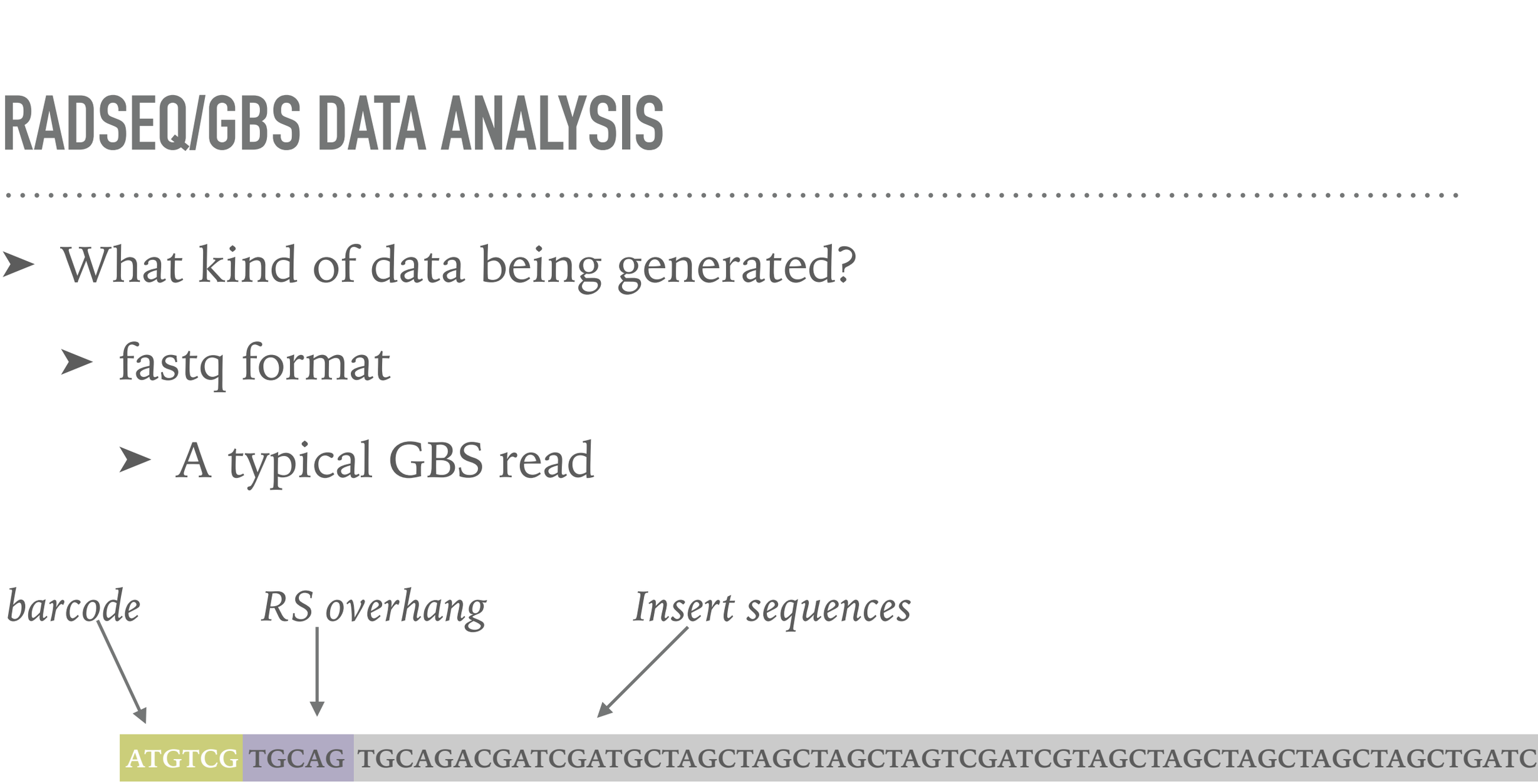
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                                    |                   |
33                        59   64       73                                   104                 126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31.......41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

A quality score Q is an integer mapping of p (the probability that the base casting is wrong.

$$Q_{\text{sanger}} = -10 \log_{10} p$$

RADSEQ/GBS DATA ANALYSIS

- What kind of data being generated?
  - fastq format
    - A typical GBS read



ATGTCGTGCAGACGATCGATGCTAGCTAGCTAGCTAGTCGATCGTAGCTAGCTAGCTAGCTAGCTGATCGATGCTAGCTAGCTAGC

RADSEQ/GBS DATA ANALYSIS

- **Mainly two types of pipeline**
  - Alignment-based
    - Treat GBS/RADseq data as usual NGS data
  - Clustering-based
    - Cluster the reads that are from the same loci, then discover variations within clusters (i.e. multiple sequence alignment)

RADSEQ/GBS DATA ANALYSIS

- **Mainly two types of pipeline**
  - Alignment-based
    - Results are easy to interpret; Imputation made easier
    - Require Ref; Computation intense; Miss novel variations
  - Clustering-based
    - Reduce the computation cost by clustering
    - Might lead to high false positive rate, or removing too many true variations with stringent filtering criteria

RADSEQ/GBS DATA ANALYSIS

- Features of the two types of pipelines
  - Alignment-based
    - require reference
    - compute intensive
    - More accurate
    - Imputation might be easier
  - Clustering-based
    - reference not required
    - Reduce the computation cost by clustering
    - Might lead to large false positive, or removing too many variation with stringent filtering criteria

Reference genome available?

Yes

Polyploidy?

Yes

Alignment-based Pipeline

No

Distant related individuals?

Yes

Clustering-based Pipeline

No

Either would work well

No

Polyploidy?

Yes

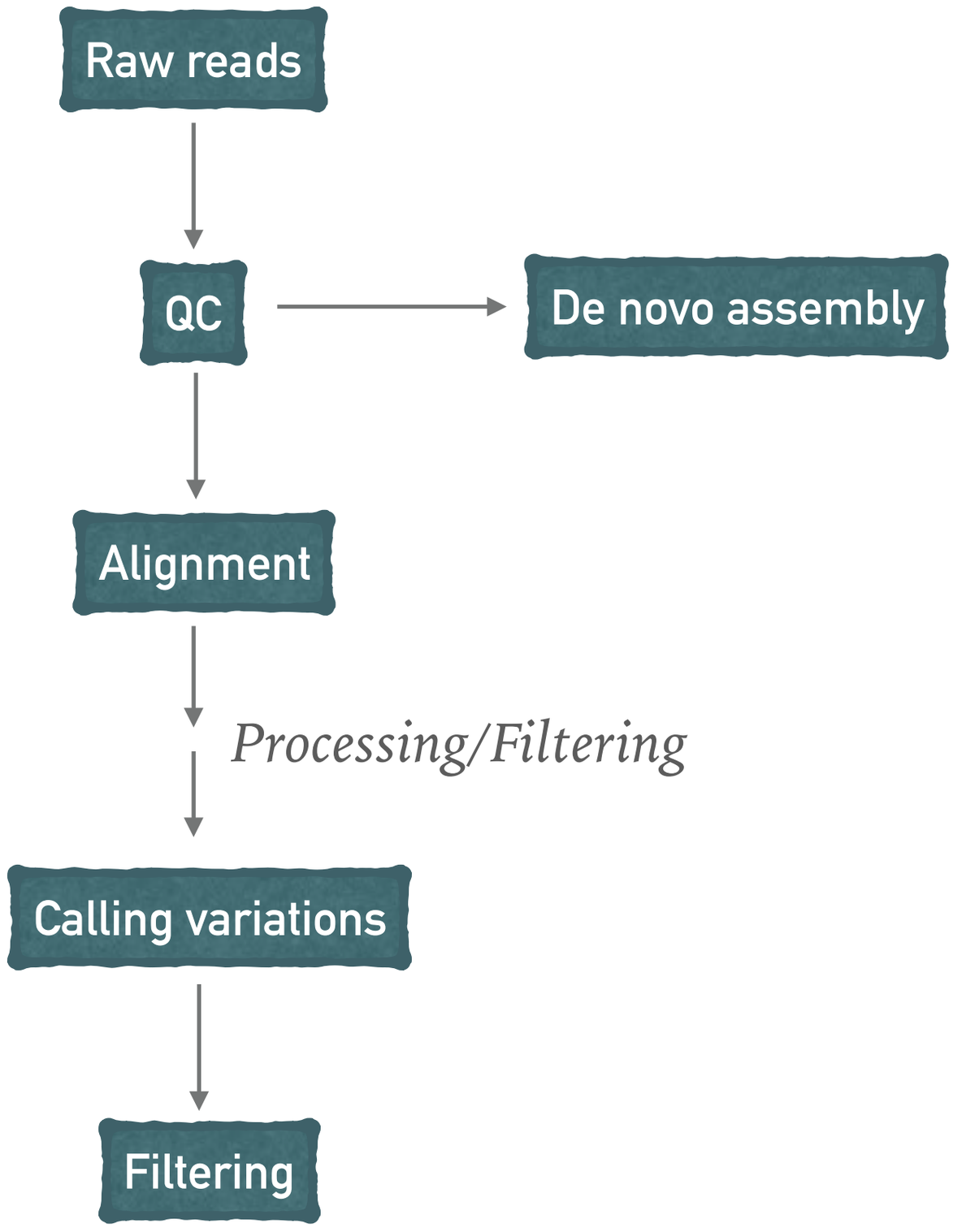
Build a reference from the reads

Alignment-based Pipeline

No

Clustering-based Pipeline

General NGS analysis workflow for variation discovery

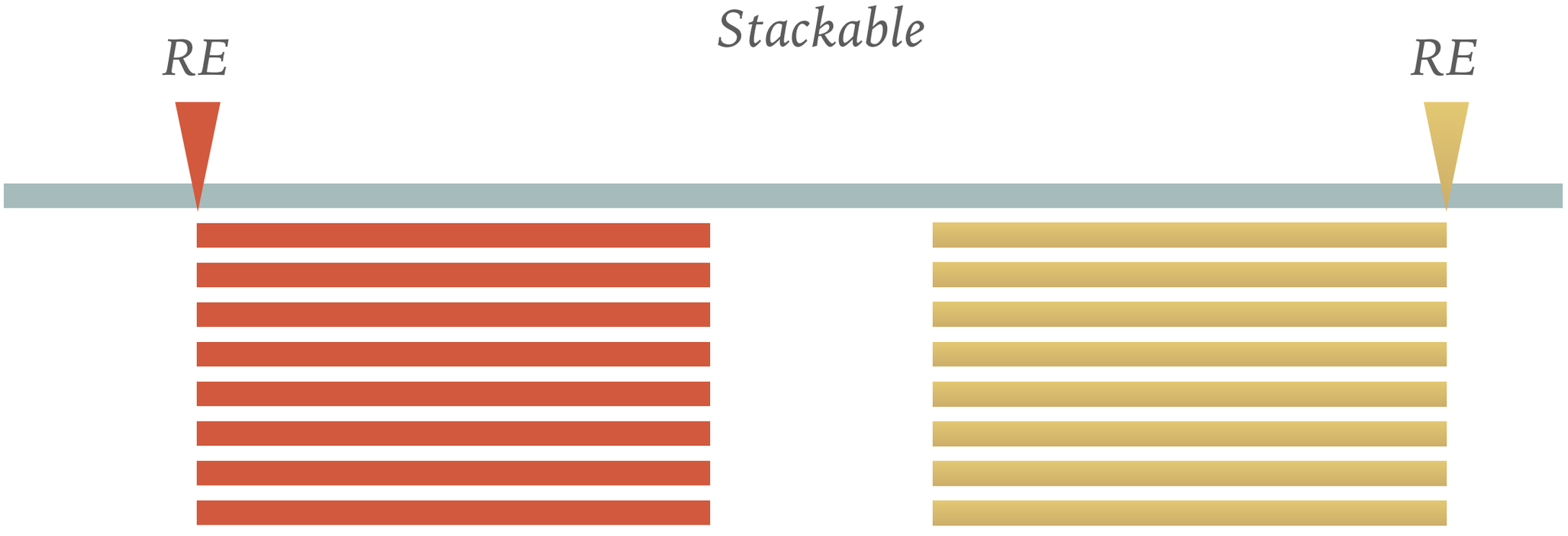


RADSEQ/GBS DATA ANALYSIS PIPELINE

Pipeline/ Progarm	Alignment	Clustering	Comment
Stacks	Y	Y	
TASSEL-GBS	Y	Y	Trim reads
UNEAK	N	Y	Trim reads
PyRAD	N	Y	
dDocent	Y	N	
AftrRAD	N	Y	

Stacks

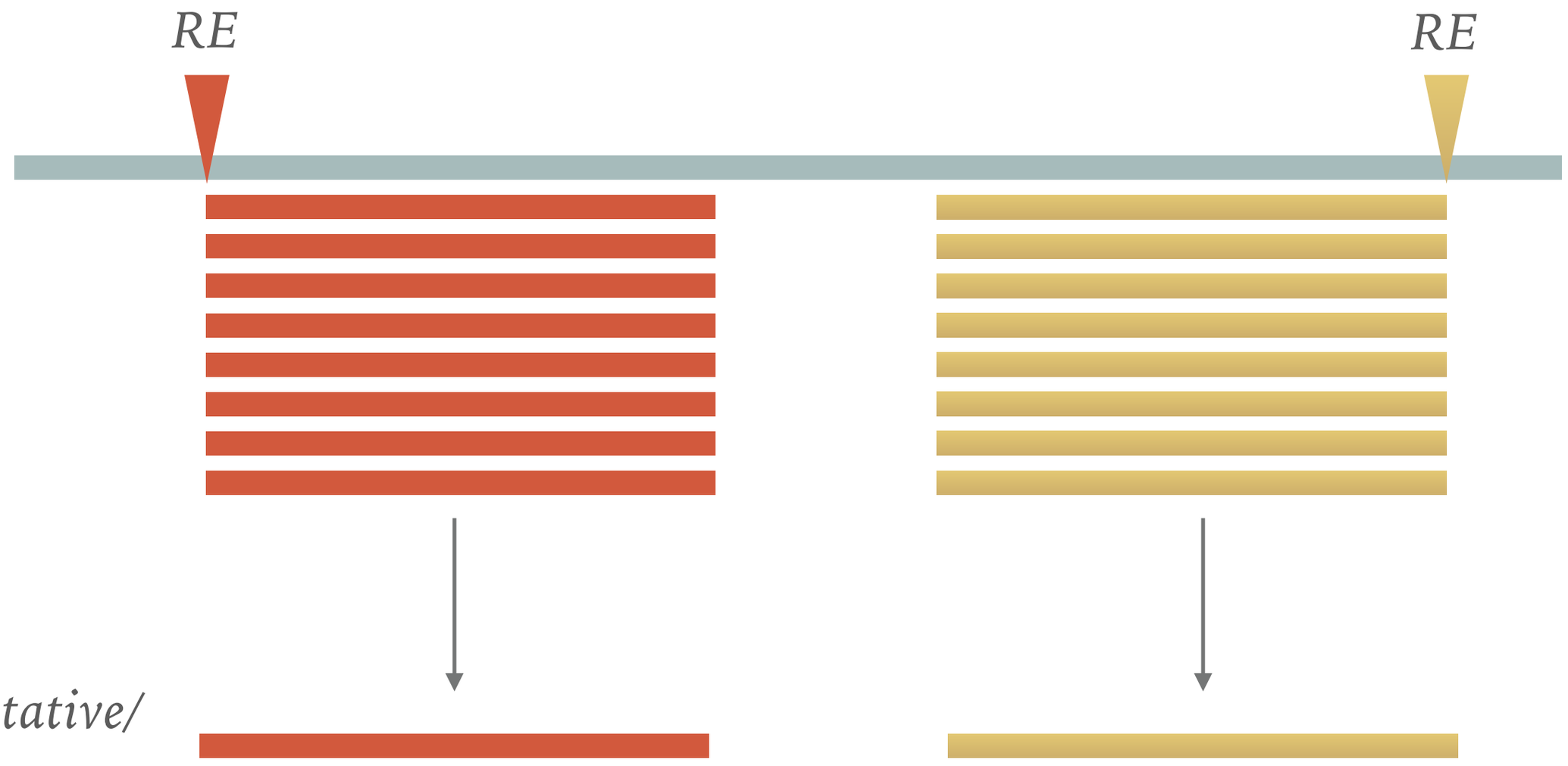
- Designed to work with short read (max 1024bp)
- Uniform length of reads
  - Ideal for Illumina
  - For Ion Torrent platform, reads would need to be truncated to a particular length
- Deal with most of the RADseq/GSB protocols
- *Stacks* is designed to process data that *stacks* together:
  - In the case of double-digest RAD, both the single-end and paired-end read are anchored by a restriction enzyme and can be assembled as independent loci;
  - In cases such as with the RAD protocol, where the molecules are sheared and the paired-end therefore does not stack-up, cannot be directly used.


Stackable
RE
RE

Non-Stackable
RE
RE

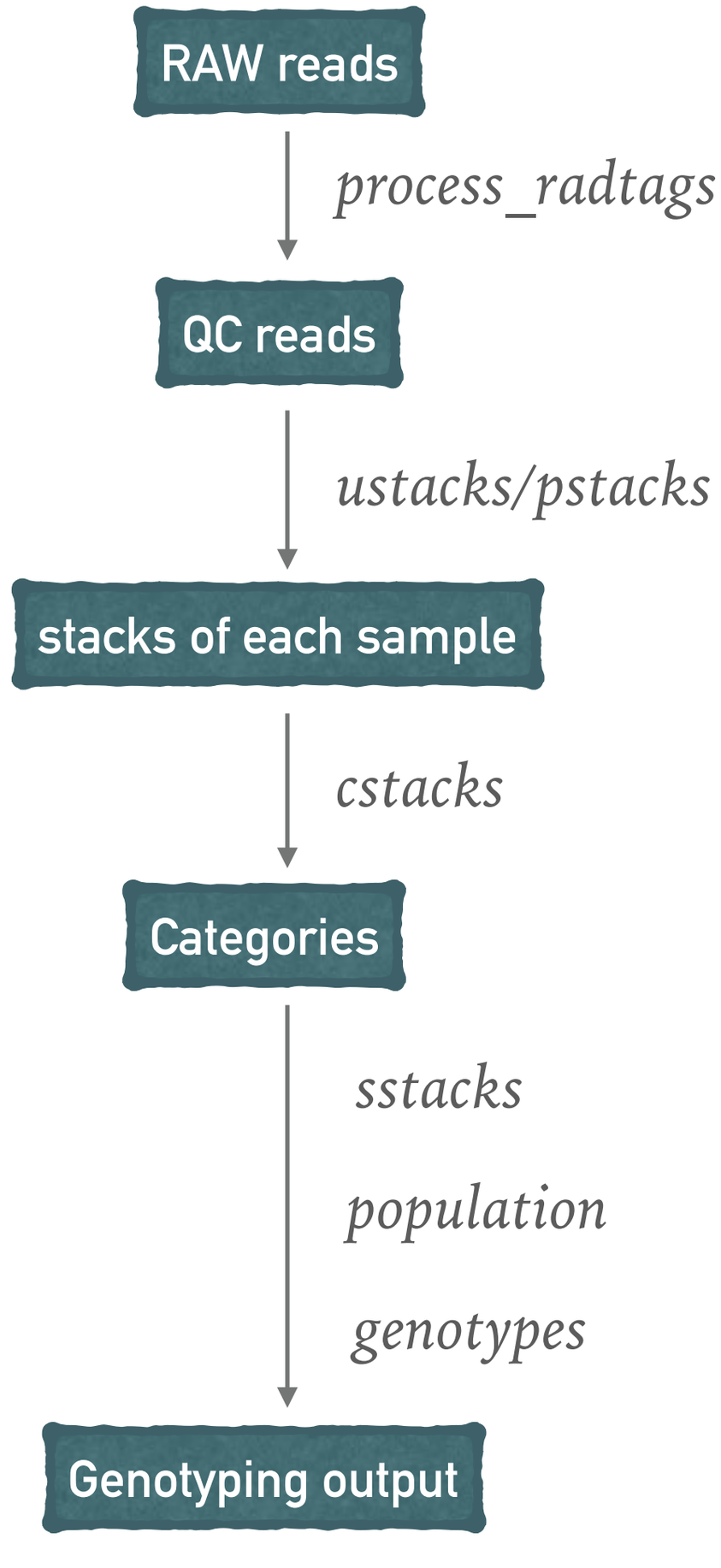
WHY CLUSTERING OR STACKING REDUCE THE COST OF COMPUTATION

TASSEL-GBS:



Instead of trying to map eight reads separately, we may just take one as representative.

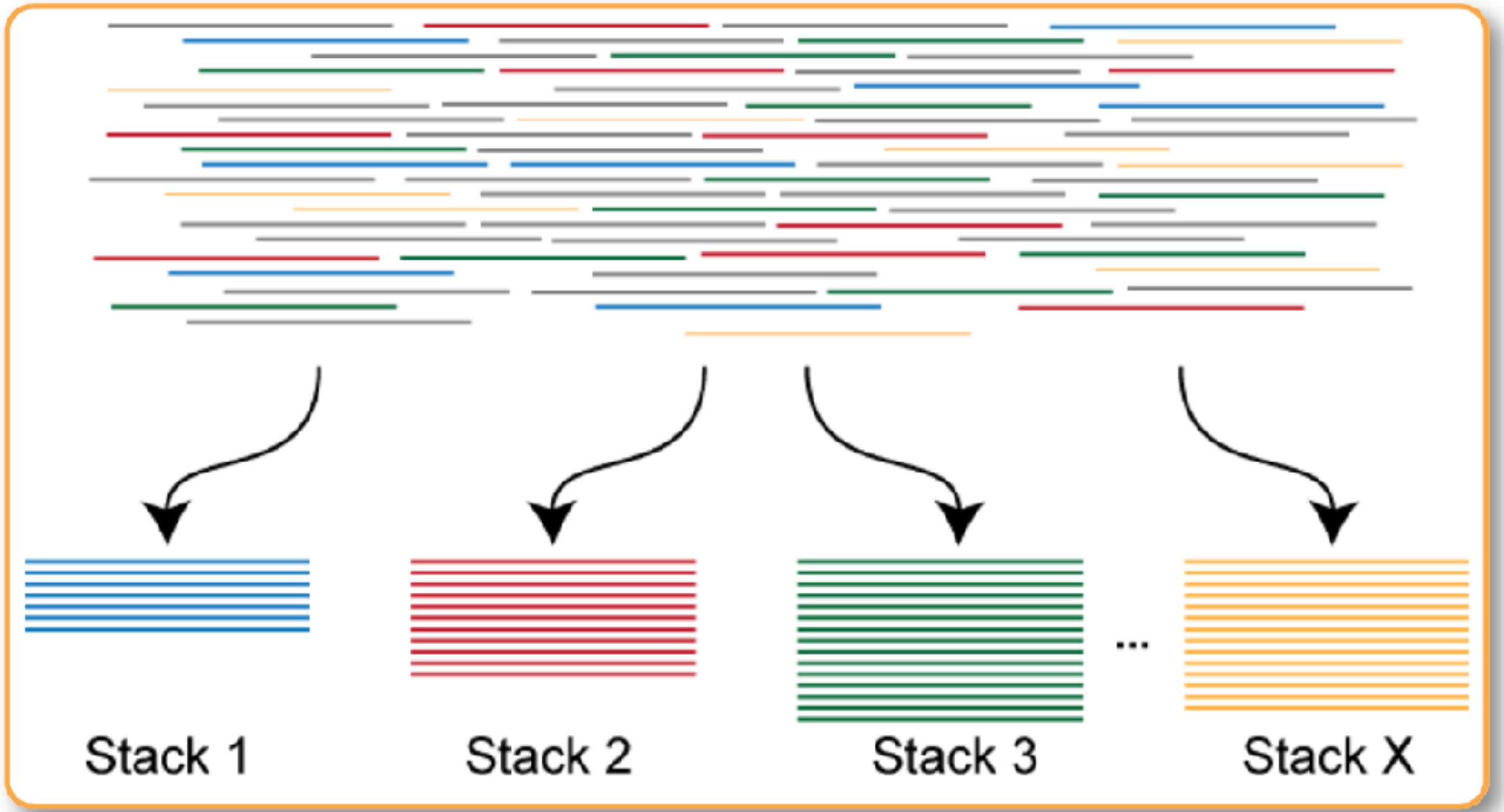
Stacks workflow



MAJOR PARAMETERS

Parameter Description	denovo_map.pl Parameter	Pipeline component	Component Parameter	Default Value
Minimum stack depth / minimum depth of coverage	-m	ustacks	-m	3
Distance allowed between stacks	-M	ustacks	-M	2
Distance allowed between catalog loci	-n	cstacks	-n	0

1. MINIMUM STACK DEPTH *-m 3*



1. MINIMUM STACK DEPTH *-m 3*

1 If set to a value of 3 then three or more identical reads must be found to consider those reads a stack. If a stack is formed with only two reads, then those reads are set aside (**secondary reads**) and a stack is not constructed.

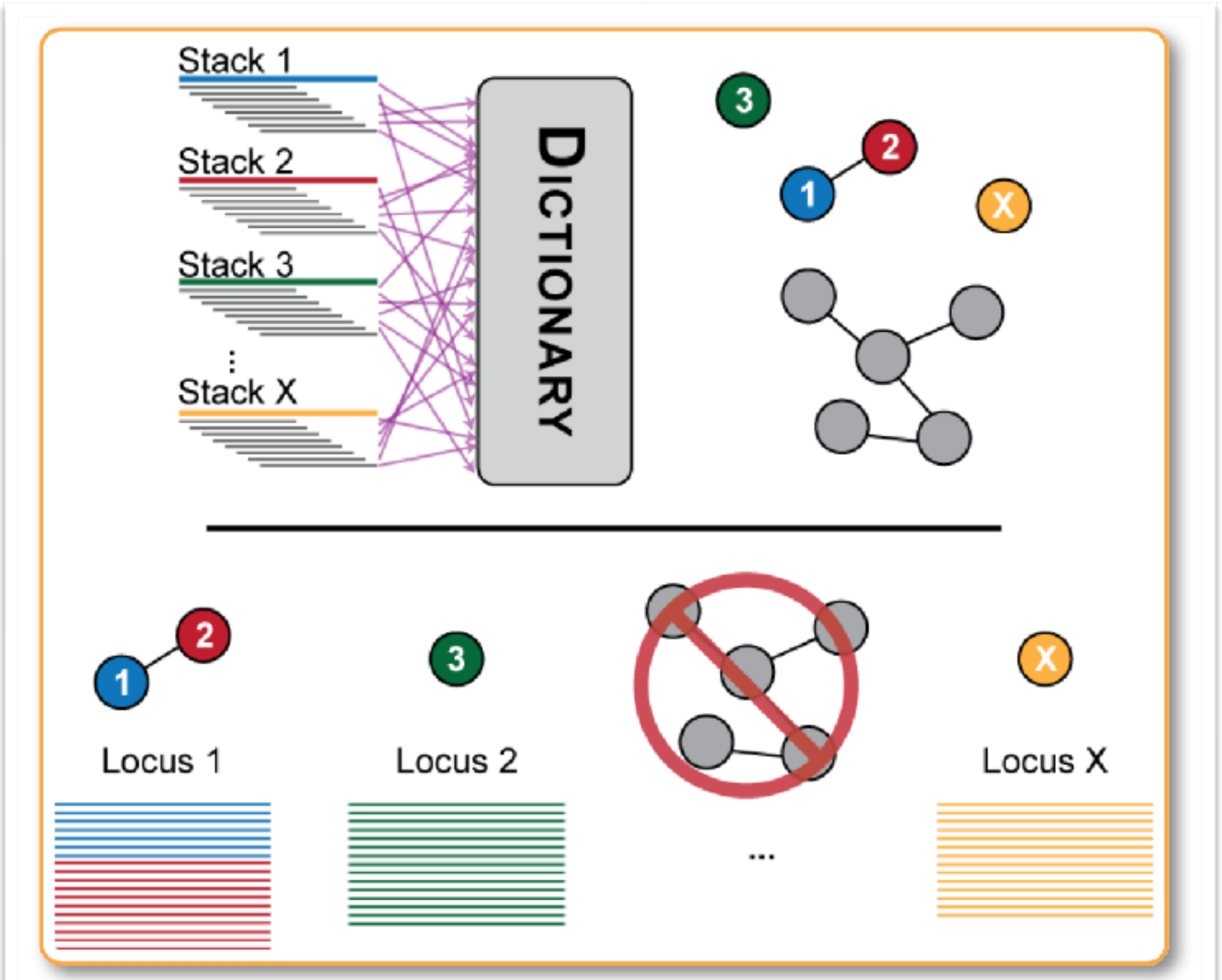
2 If this parameter is set too low, then reads with convergent sequencing errors are likely to be erroneously labled as stacks.

3 If this parameter too high, then true alleles will not be recorded and will drop out of the analysis.

4 If you have low sequencing depth for your samples, you will have to set this parameter to a relatively low value. Conversely, if you have very high sequencing coverage, you will want to increase this parameter.

5 If you have a high error rate in your sequencing lane, then you are likely to see convergent sequencing or PCR errors (errors that occur independently at the same nucleotide position in the same read) and should increase the minimum stack depth.

2. DISTANCE ALLOWED BETWEEN STACKS *-M 2*



2. DISTANCE ALLOWED BETWEEN STACKS *-M 2*

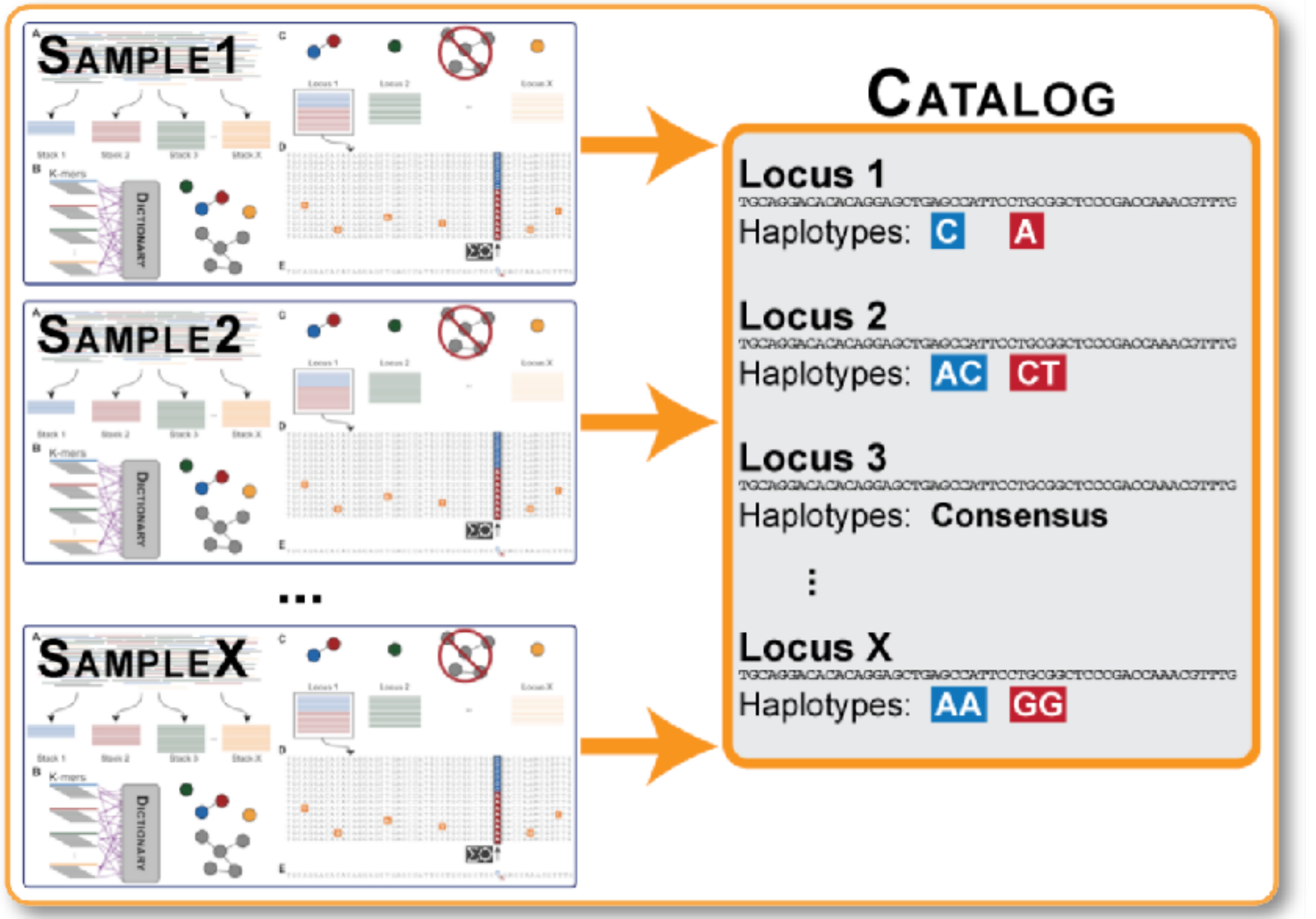
1 If you set this parameter too low, then some loci will fail to be reconstructed. This means the SNPs contained in that locus will not be identified and this locus will appear as two loci to the remainder of the pipeline.

2 Setting this parameter too high will allow repetitive sequence to chain together in to large, nonsensical loci. For example, if stack A is one nucleotide apart from stack B, which is one nucleotide apart from stack C, which is one nucleotide apart from stack D, then A, B, C, and D will be merged into a locus despite A and D being four nucleotides apart. These loci are not useful to the pipeline and at several points the pipeline will try to detect these and set them aside.

3 You will want to experiment with several different values of this parameter to see how many polymorphic loci you can construct.

3. DISTANCE BETWEEN CATALOG LOCI

-n 0



3. DISTANCE BETWEEN CATALOG LOCI *-n 0*

1 If you set this parameter too low, then some loci will fail to be reconstructed. This means the SNPs contained in that locus will not be identified and this locus will appear as two loci to the remainder of the pipeline.

2 Setting this parameter too high will allow repetitive sequence to chain together in to large, nonsensical loci. For example, if stack A is one nucleotide apart from stack B, which is one nucleotide apart from stack C, which is one nucleotide apart from stack D, then A, B, C, and D will be merged into a locus despite A and D being four nucleotides apart. These loci are not useful to the pipeline and at several points the pipeline will try to detect these and set them aside.

3 You will want to experiment with several different values of this parameter to see how many polymorphic loci you can construct.

OPTIMIZE THE PARAMETERS

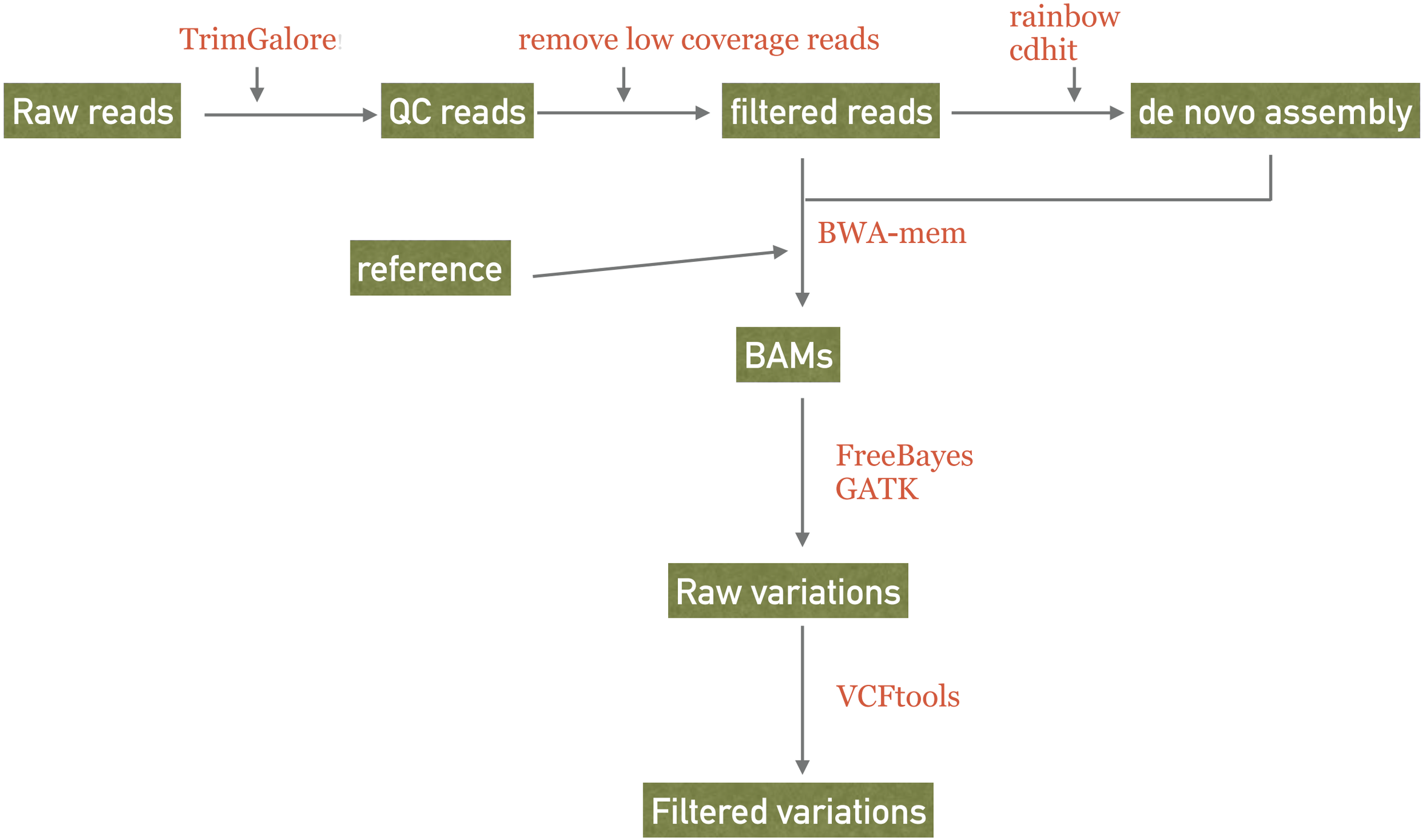
- How to optimize the parameters for my project?
  - Simulation
    - With reference genome available, simulate RADseq/GBS reads from the reference genome with predefined SNPs;
    - Call SNPs with different set of parameters, pick the one with the lowest FP and/ high TP.
  - Generate SNPs for multiple sets of parameters, then check the SNP accuracy

DDOCENT PIPELINE

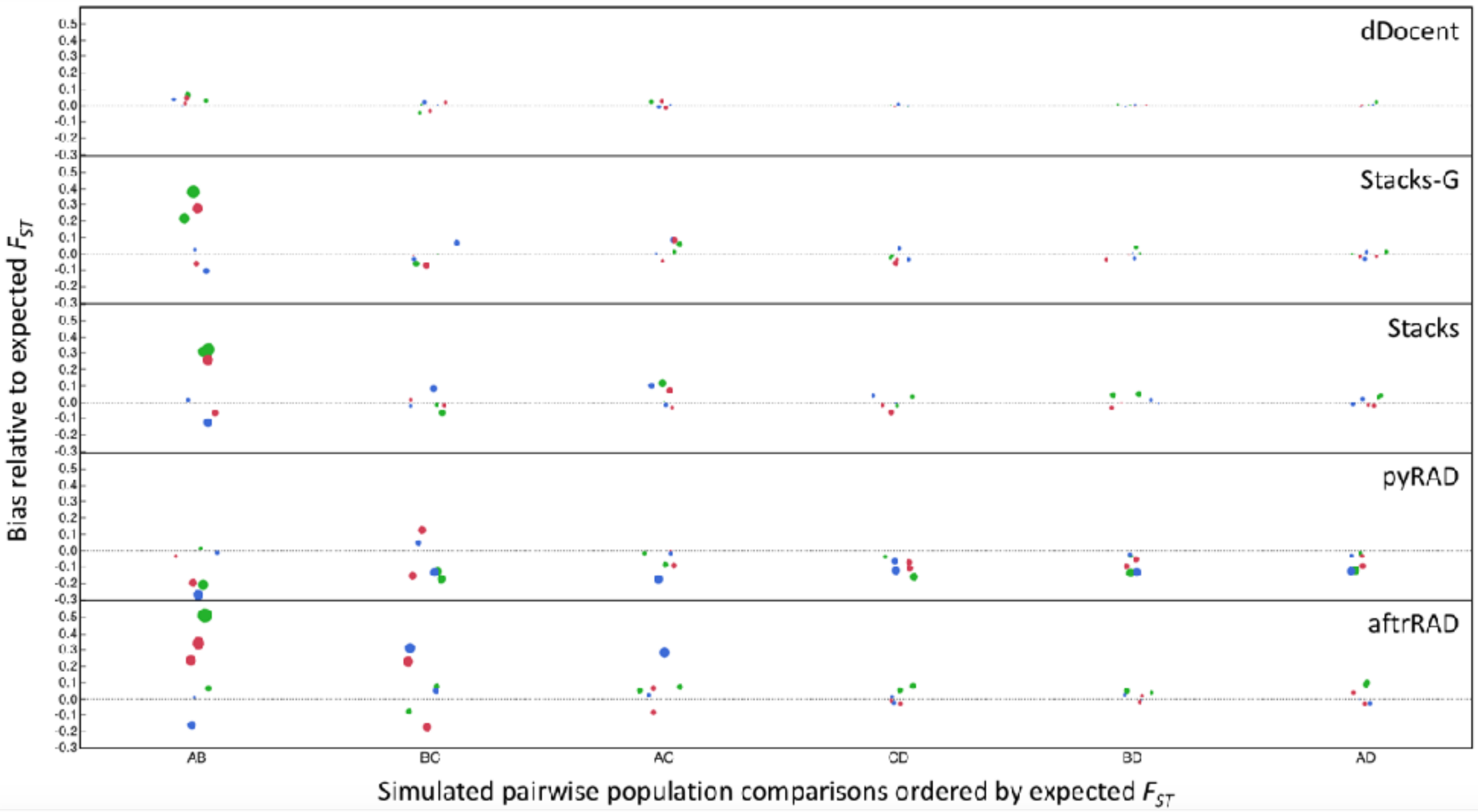
- dDocent relies almost entirely on third party software to complete every step of the analysis pipeline..

FreeBayes	https://github.com/ekg/freebayes
STACKS	http://creskolab.uoregon.edu/stacks/
PEAR	http://sco.h-its.org/exelixis/web/software/pear/
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
Mawk	http://invisible-island.net/mawk/
BWA	http://bio-bwa.sourceforge.net
SAMtools	http://samtools.sourceforge.net
VCFtools v.1.11**	http://vcftools.sourceforge.net/index.html
rainbow	http://sourceforge.net/projects/bio-rainbow/files/
seqtk	https://github.com/lh3/seqtk
CD-HIT	http://weizhong-lab.ucsd.edu/cd-hit/

DDOCENT PIPELINE



DDOCENT PIPELINE



Let’s give it a try!

Any questions?

DEMO OF GBS DATA ANALYSIS USING STACKS ON HPRC

login to ada: ssh your-tamu-netid@ada.tamu.edu

create a folder: mkdir stacks_tutorial && cd stacks_tutorial

get the tutorial data: git clone https://github.com/swang8/workshop

Create working directory:

cd workshop/material/tutorial_data/

mkdir RAD && cd RAD

tar xvf ../rad_test.tar

ls -l

DEMO OF GBS DATA ANALYSIS USING STACKS ON HPRCGALAXY

For testing only: https://hprcgalaxy.tamu.edu/fishcamp

For real projects: https://hprcgalaxy.tamu.edu/maroon