Genotyping-by-Sequencing (GBS): Applications in Sweetpotato

Bode Olukolu & Craig Yencho
Dept. of Horticultural Sciences
North Carolina State University (NCSU)



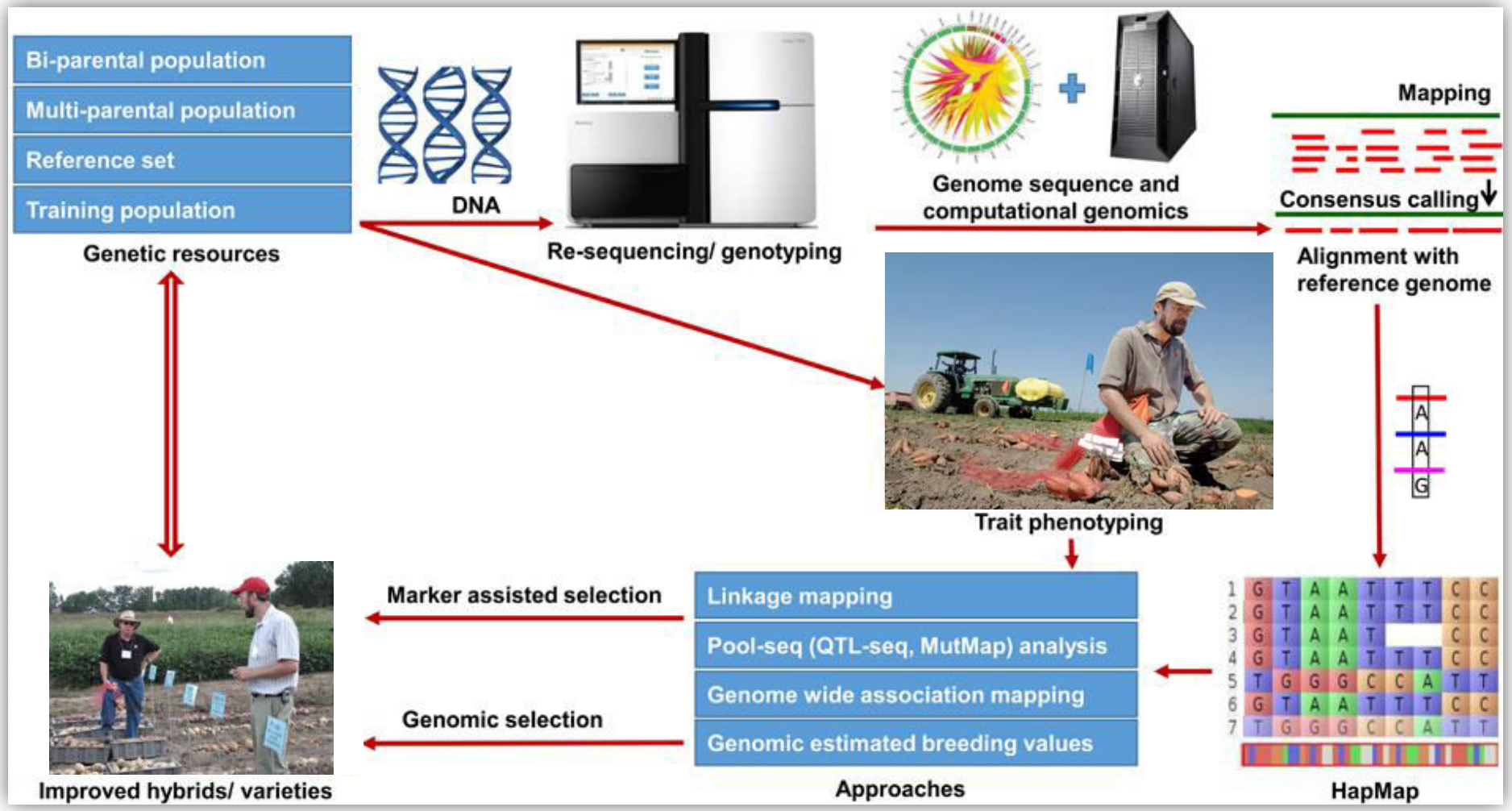

BILL & MELINDA GATES foundation

Acknowledgements

GT4SP
GENOMIC TOOLS FOR SWEETPOTATO IMPROVEMENT

- Craig Yencho (NCSU): Sharon Williamson; Bonnie Oloka; Victor Amankwaah; David Baltzegar; Hannah Huntley; Erin Young
- Awais Khan (CIP): Robert Mwanga; Dorcus Gemenet; David Maria
- NACRRI: Benard Yada
- Zhao-Zang Zeng (NCSU): Guilherme Da Silva Pereira; Marcelo Mollinari
- Lachlan Coin (University of Queensland)
- Zhangjun Fei (Cornell University)
- Robin Buell (Michigan State University)

Developing Genomic Resources: Interdisciplinary Research



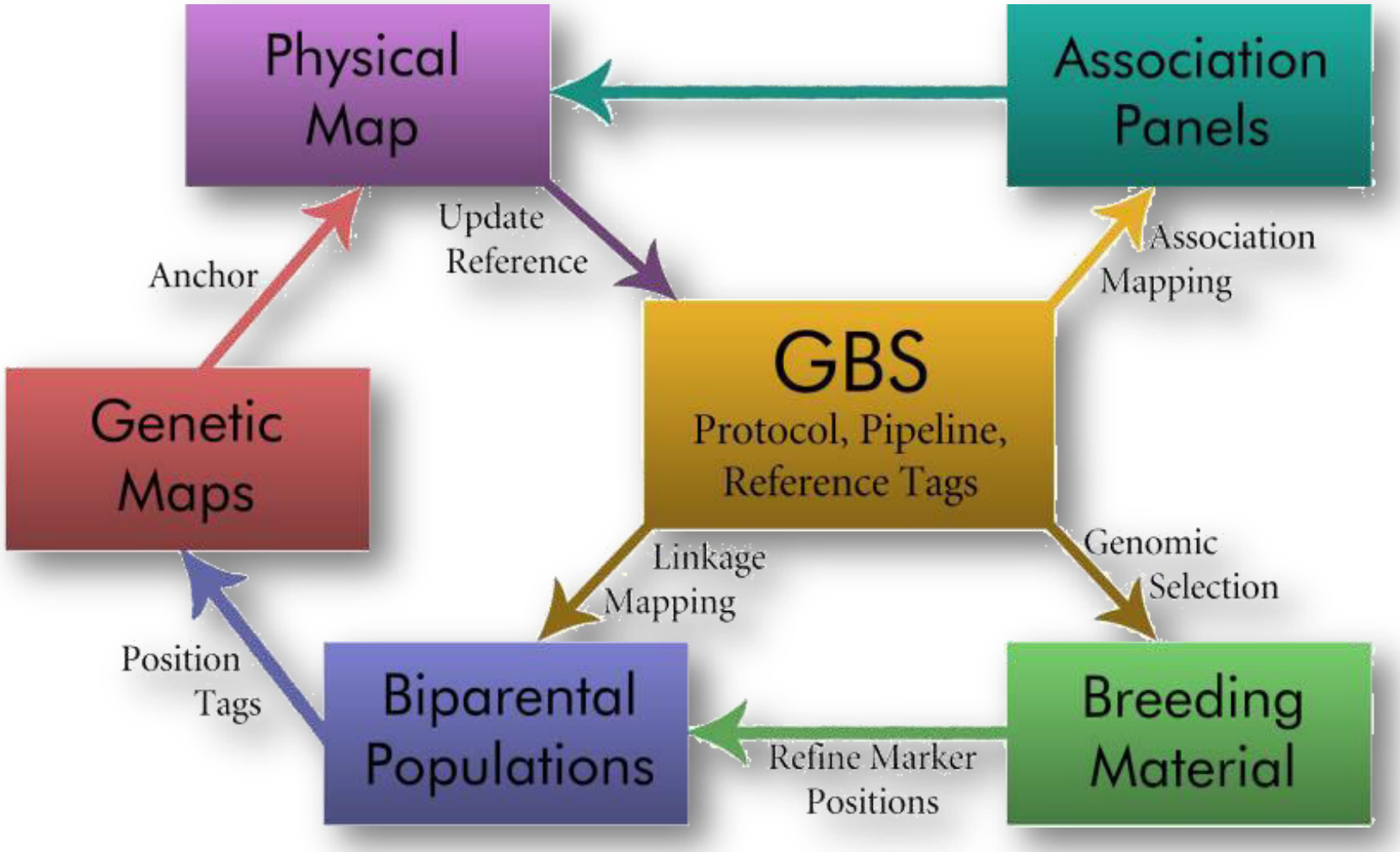
Challenge:

- Integrating knowledge from various disciplines in a seamless manner

- Developing & Incorporating Diagnostics tools into Analyses

Pazhamala et al. (2015) Front. Plant Sci.

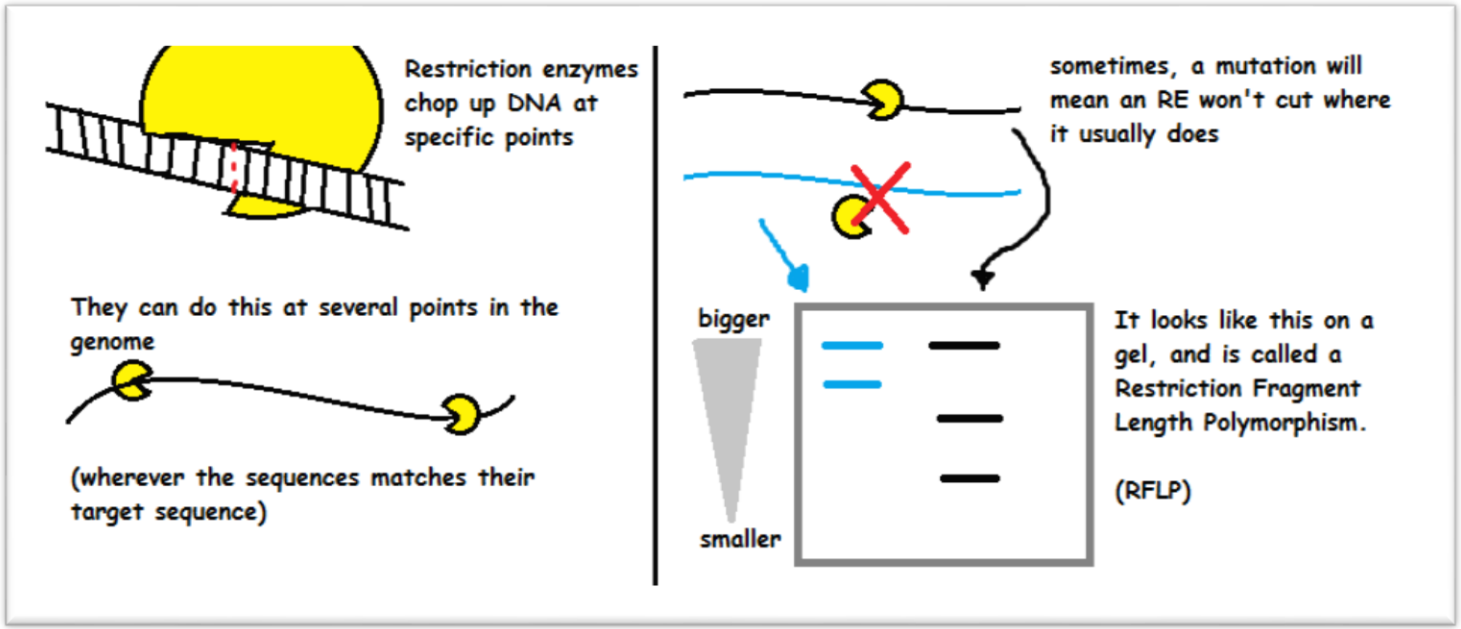
Molecular markers:
central to “genomic resources” and “genetic analyses”



Poland and Rife (2012) The Plant Genome

First generation molecular markers

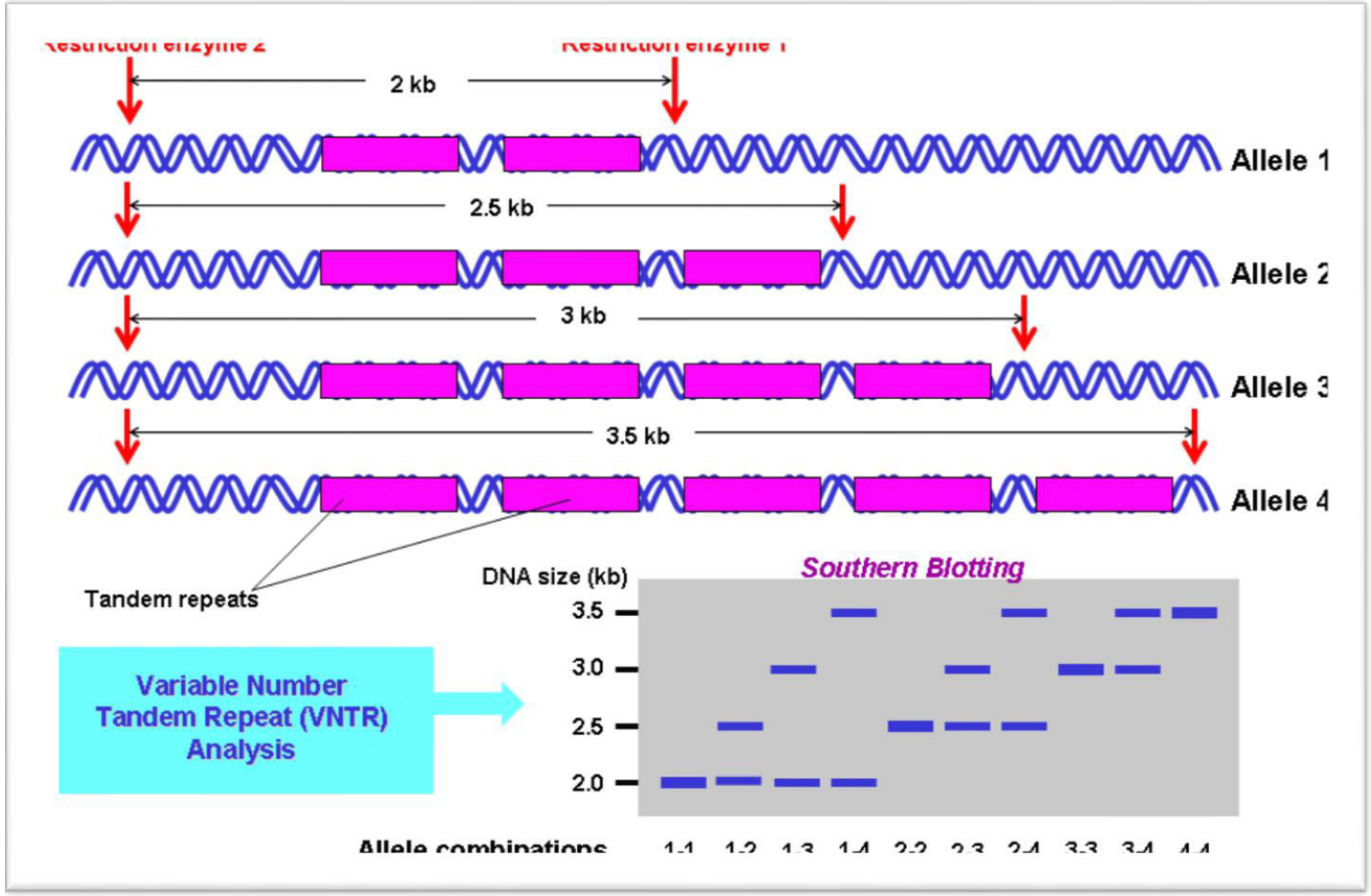
Year	Acronym	Nomenclature	Reference
1974	RFLP	Restriction Fragment Length Polymorphism	Grodzicker et al. (L974)
1985	VNTR	Variable Number Tandem Repeats	Jeffreys et al. (1985)
1986	ASO	Allele Specific Oligonucleotides	Saiki et al. (1986)
1988	AS-PCR	Allele Specific Polymerase Chain Reaction	Landegren et al. (1988)
1988	OP	Oligonucleotide Polymorphism	Beckmann (1988)
1989	SSCP	Single Stranded Conformational Polymorphism	Orita et al. (1989)
1989	STS	Sequence Tagged Site	Olsen et al. (1989)



Maheswaran (2004) Advanced Biotech

First generation molecular markers

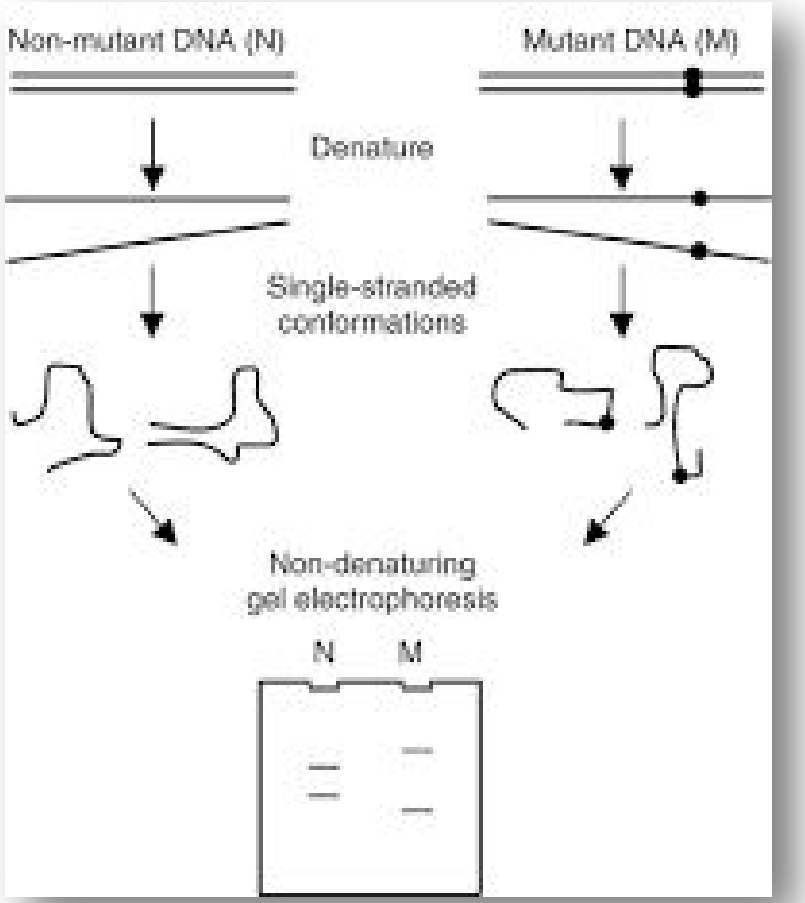
Year	Acronym	Nomenclature	Reference
1974	RFLP	Restriction Fragment Length Polymorphism	Grodzicker et al. (L974)
1985	VNTR	Variable Number Tandem Repeats	Jeffreys et al. (1985)
1986	ASO	Allele Specific Oligonucleotides	Saiki et al. (1986)
1988	AS-PCR	Allele Specific Polymerase Chain Reaction	Landegren et al. (1988)
1988	OP	Oligonucleotide Polymorphism	Beckmann (1988)
1989	SSCP	Single Stranded Conformational Polymorphism	Orita et al. (1989)
1989	STS	Sequence Tagged Site	Olsen et al. (1989)



Maheswaran (2004) Advanced Biotech

First generation molecular markers

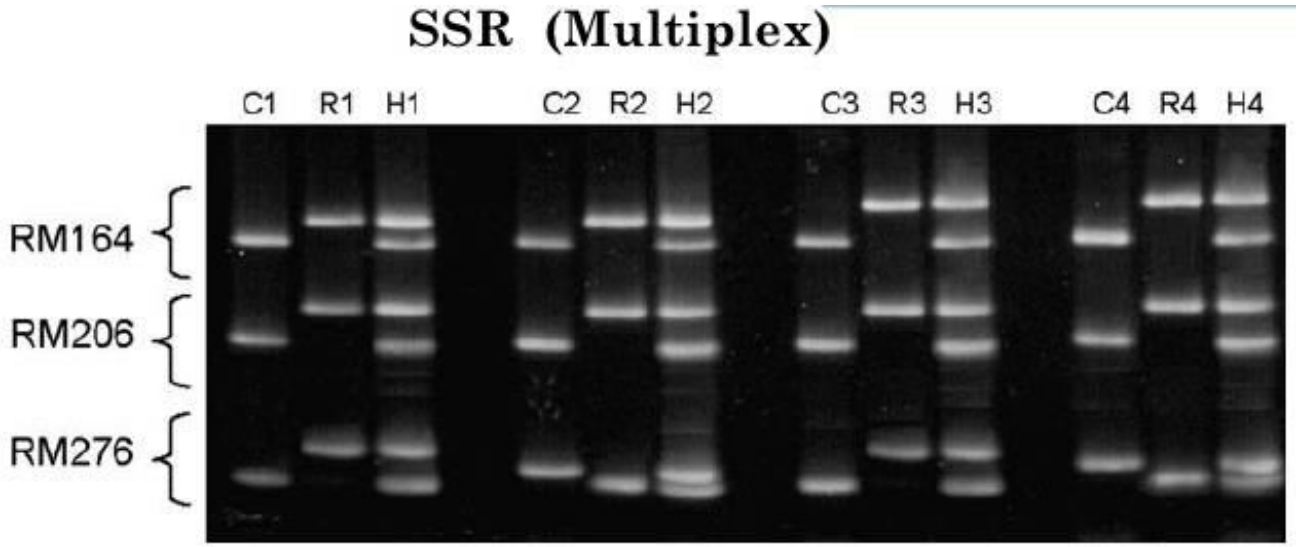
Year	Acronym	Nomenclature	Reference
1974	RFLP	Restriction Fragment Length Polymorphism	Grodzicker et al. (L974)
1985	VNTR	Variable Number Tandem Repeats	Jeffreys et al. (1985)
1986	ASO	Allele Specific Oligonucleotides	Saiki et al. (1986)
1988	AS-PCR	Allele Specific Polymerase Chain Reaction	Landegren et al. (1988)
1988	OP	Oligonucleotide Polymorphism	Beckmann (1988)
1989	SSCP	Single Stranded Conformational Polymorphism	Orita et al. (1989)
1989	STS	Sequence Tagged Site	Olsen et al. (1989)



Maheswaran (2004) Advanced Biotech

Second generation molecular markers

Year	Acronym	Nomenclature	Reference
1990	RAPD	Randomly Amplified Polymorphic DNA	Williams et al. (1990)
1990	AP-PCR	Arbitrarily Primed Polymerase Chain Reaction	Welsh and McClelland (1990)
1990	STMS	Sequence Tagged Micro Satellite Sites	Beckmann and Soller (1990)
1991	RLGS	Restriction Landmark Genome Scanning	Hatada et al. (1991)
1992	CAPS	Cleaved Amplified Polymorphic Sequence	Akopyanz et al. (1992)
1992	DOP-PCR	Degenerate Oligonucleotide Primer - PCR	Telenius (1992)
1992	SSR	Simple Sequence Repeats	Akkaya et al. (1992)
1993	MAAP	Multiple Arbitrary Amplicon Profiling	Caetano-Anollés et al. (1993)
1993	SCAR	Sequence Characterized Amplified Region	Paran and Michelmore (1993)

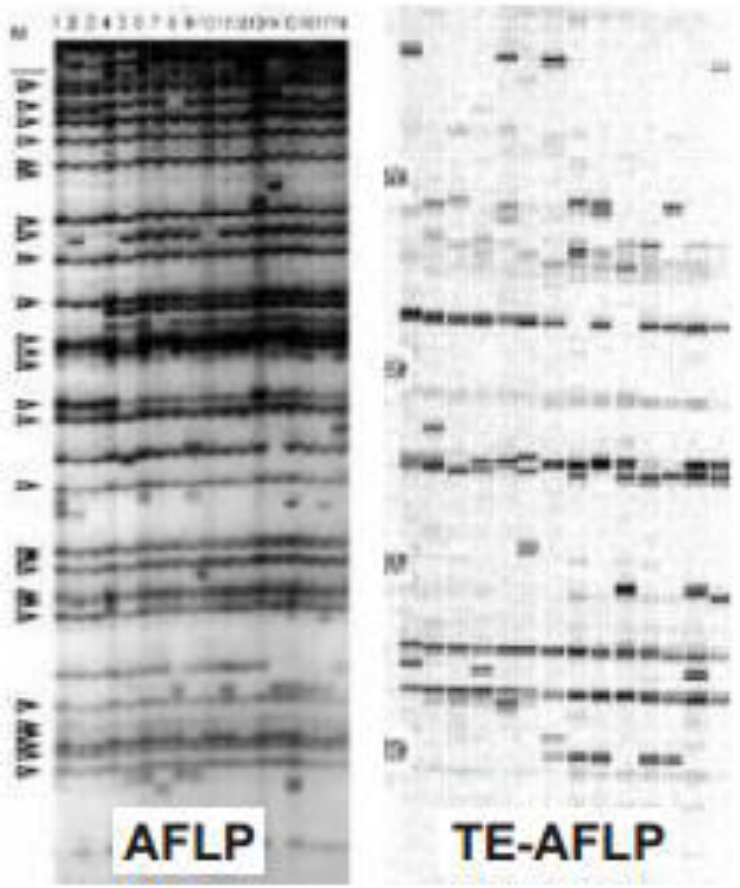


Maheswaran (2004) Advanced Biotech

Second generation molecular markers

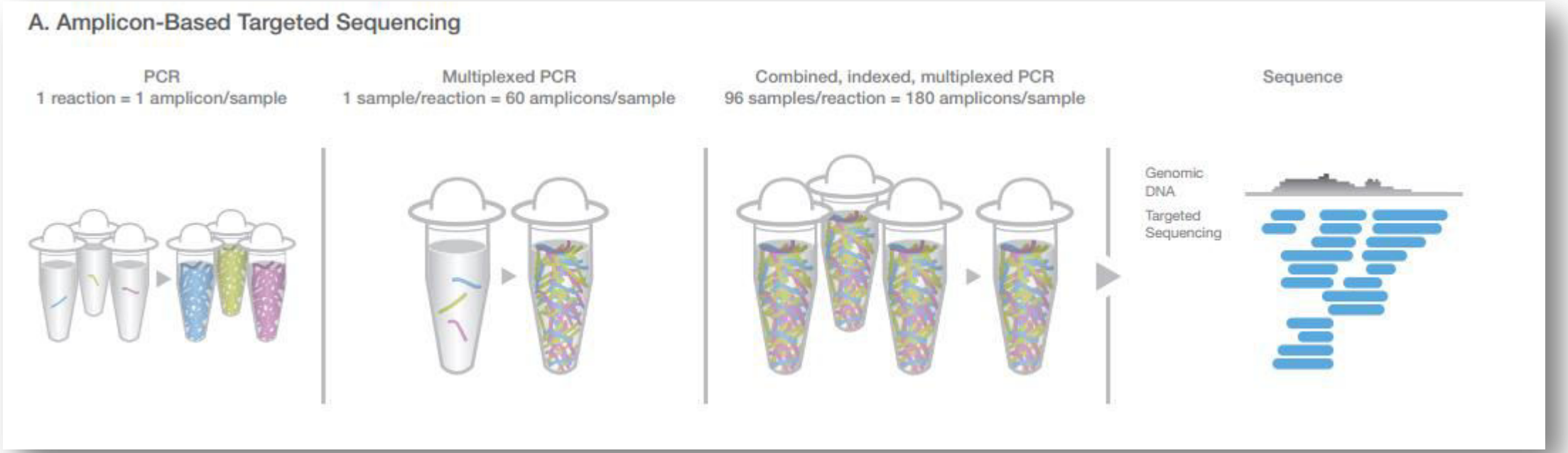
Year	Acronym	Nomenclature	Reference
1994	ISSR	Inter Simple Sequence Repeats	Zietkiewicz et al (1994)
1994	SAMPL	Selective Amplification Of Micro Satellite Polymorphic Loci	Morgante and Vogel, (1994)
1994	SNP	Single Nucleotide Polymorphisms	Jordan and Humphries (1994)
1995	AFLP (SRFA)	Amplified Fragment Length Polymorphism (selective Restriction Fragment Amplification)	Vos et al. (1995)
1995	ASAP	Allele Specific Associated Primers	Gu et al. (1995)
1996	CFLP	Clearvase Fragment Length Polymorphism	Brow (1996)
1996	ISTR	Inverse Sequence-tagged Repeats	Rhode (1996)
1997	DAMD-PCR	Directed Amplification Of Mini Satellite DNA-PCR	Bebeli et al. 1997

Year	Acronym	Nomenclature	Reference
1997	S-SAP	Sequence-specific Amplified Polymorphism	Waugh et al. (1997)
1998	RBIP	Retrotransposon Based Insertional Polymorphism	Flavell et al. (1998)
1999	IRAP	Inter-retrotransposon Amplified Polymorphism	Kalendar et al. (1999)
1999	REMAP	Retrotransposon-Microsatellite Amplified Polymorphism	Kalendar et al. (1999).
1999	MSAP	Methylation Sensitive Amplification Polymorphism	-----
2000	MITE	Miniature Inverted-repeat Transposable Element	Casa et al. (2000)
2000	TE-AFLP	Three Endonuclease AFLP	van der Wurff et al. (2000)
2001	IMP	Inter-MITE Polymorphisms	Chang et al. (2001)
2001	SRAP	Sequence-related Amplified Polymorphism	Li and Quiros (2001)



Maheswaran (2004) Advanced Biotech

Sequencing-based Genotyping methodologies

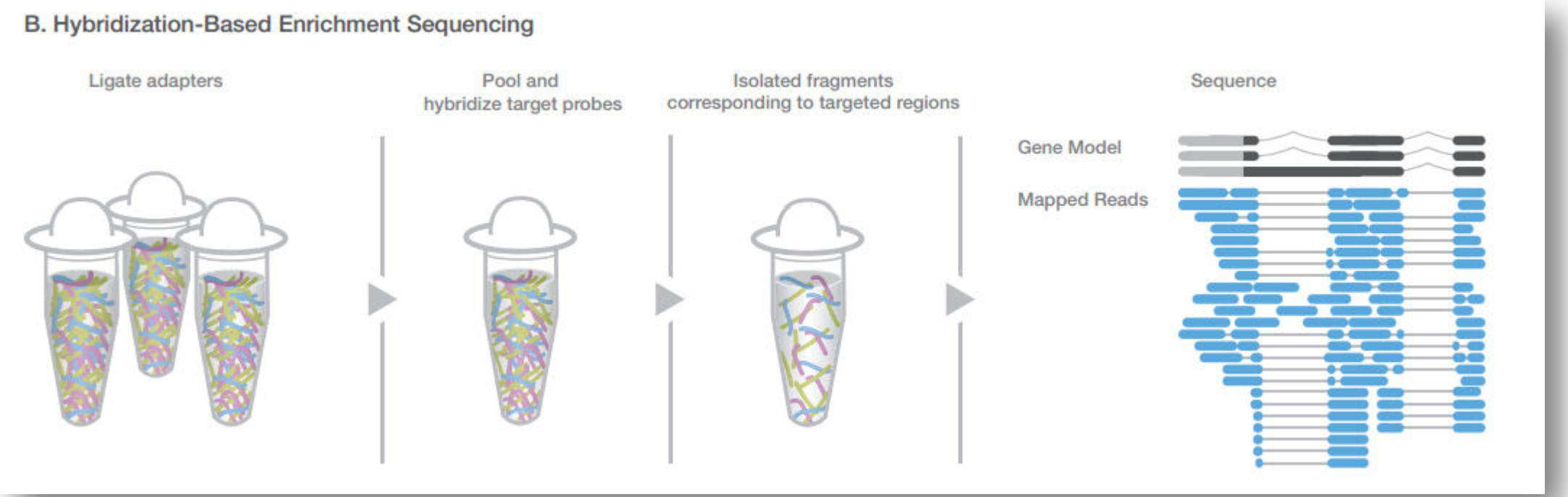


Sequencing-based genotyping

- Skim sequencing (no genome reduction)
- Enrichment/target capture/hybridization-based
- **Restriction enzyme methods (RE-GBS, RAD-Seq, ddRAD-Seq)**
- SNP chip arrays
- Transcriptome sequencing

http://www.illumina.com/Documents/products/appspotlights/app_spotlight_ngg_ag.pdf

Sequencing-based Genotyping methodologies

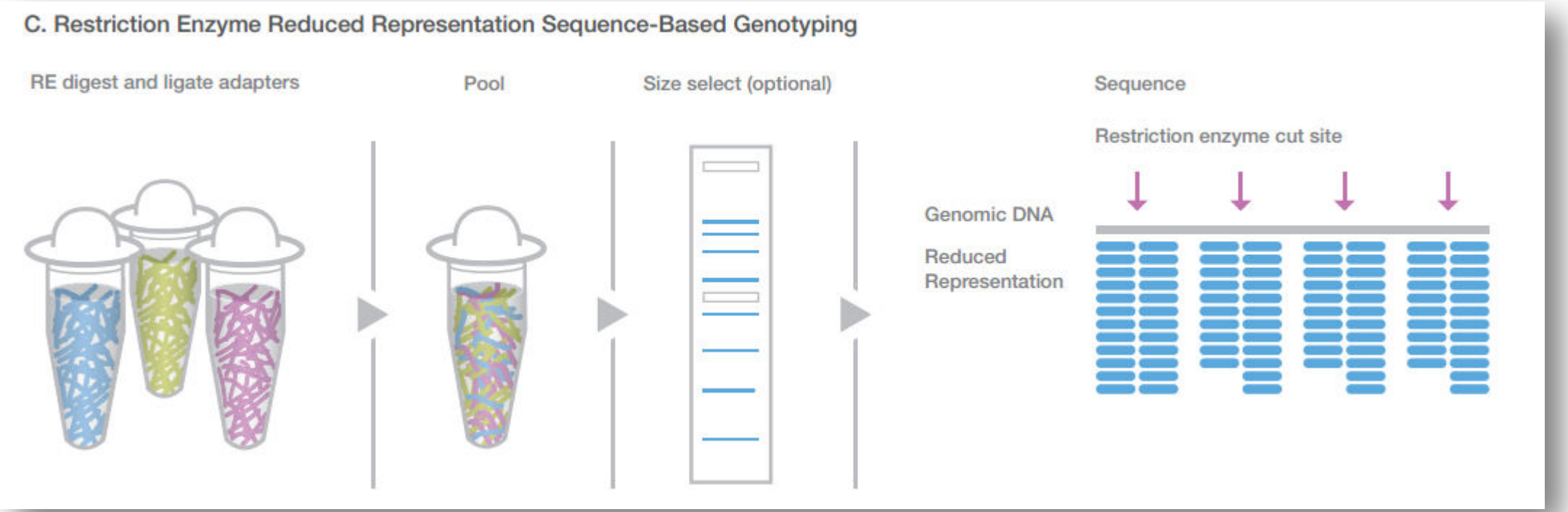


Sequencing-based genotyping

- Skim sequencing (no genome reduction)
- Enrichment/target capture/hybridization-based
- **Restriction enzyme methods (RE-GBS, RAD-Seq, ddRAD-Seq)**
- SNP chip arrays
-Transcriptome sequencing

http://www.illumina.com/Documents/products/appspotlights/app_spotlight_ngg_ag.pdf

Sequencing-based Genotyping methodologies



Sequencing-based genotyping

- Skim sequencing (no genome reduction)
- Enrichment/target capture/hybridization-based
- **Restriction enzyme methods (RE-GBS, RAD-Seq, ddRAD-Seq)**
- SNP chip arrays
- Transcriptome sequencing

http://www.illumina.com/Documents/products/appspotlights/app_spotlight_ngg_ag.pdf

Advantages of Genotyping by Sequencing

1) Sequences predetermined areas of genetic variation over many samples as far as:
   -reference genome -high-diversity samples
   -finely tuned coverage across multiplexed samples)

2) Reduces ascertainment bias compared to arrays

3) Identifies variants other than SNPs (i.e. small insertions, deletions, and microsatellites)

4) Provides a low cost per sample ($24/sample)

Poland and Rife (2012) The Plant Genome

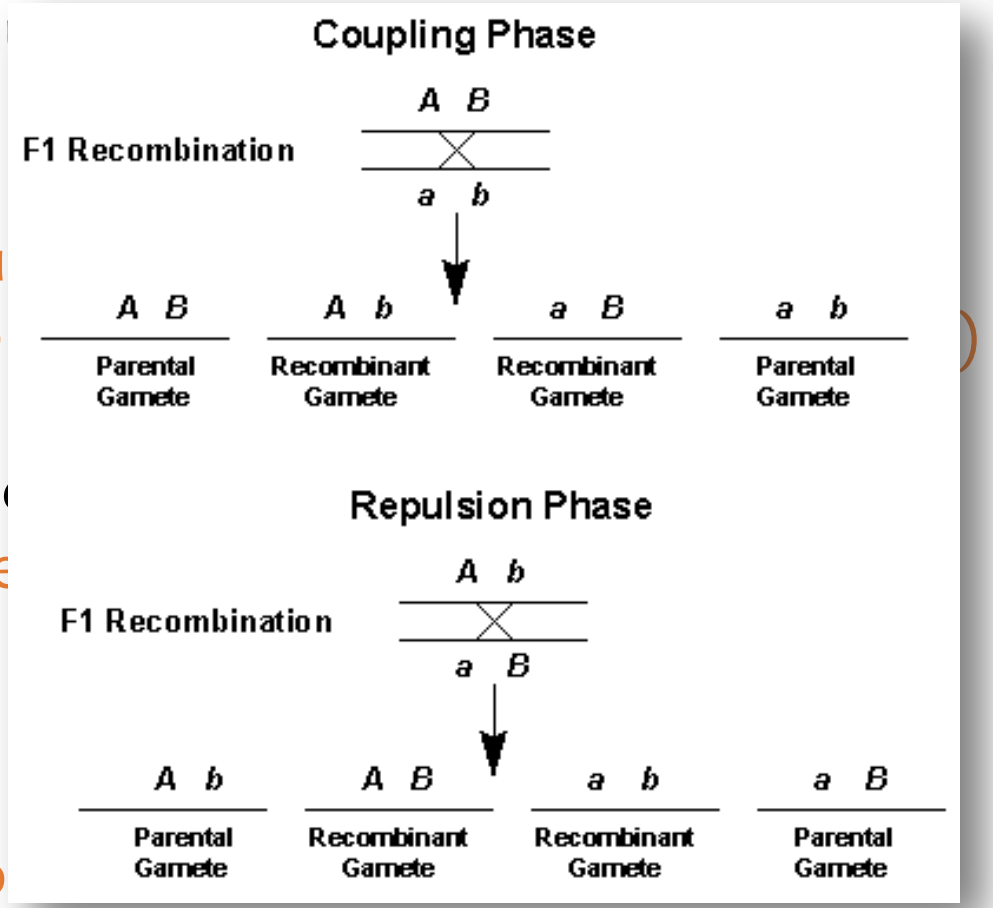
GBS pitfalls

1) Issues with erroneously calling heterozygotes as homozygotes (low coverage sequencing): - Increase sequencing depth -

2) Lots of missing data:
   - Error in de-multiplexing/barcodes (substitu
   - Coverage not uniform across loci: optimize

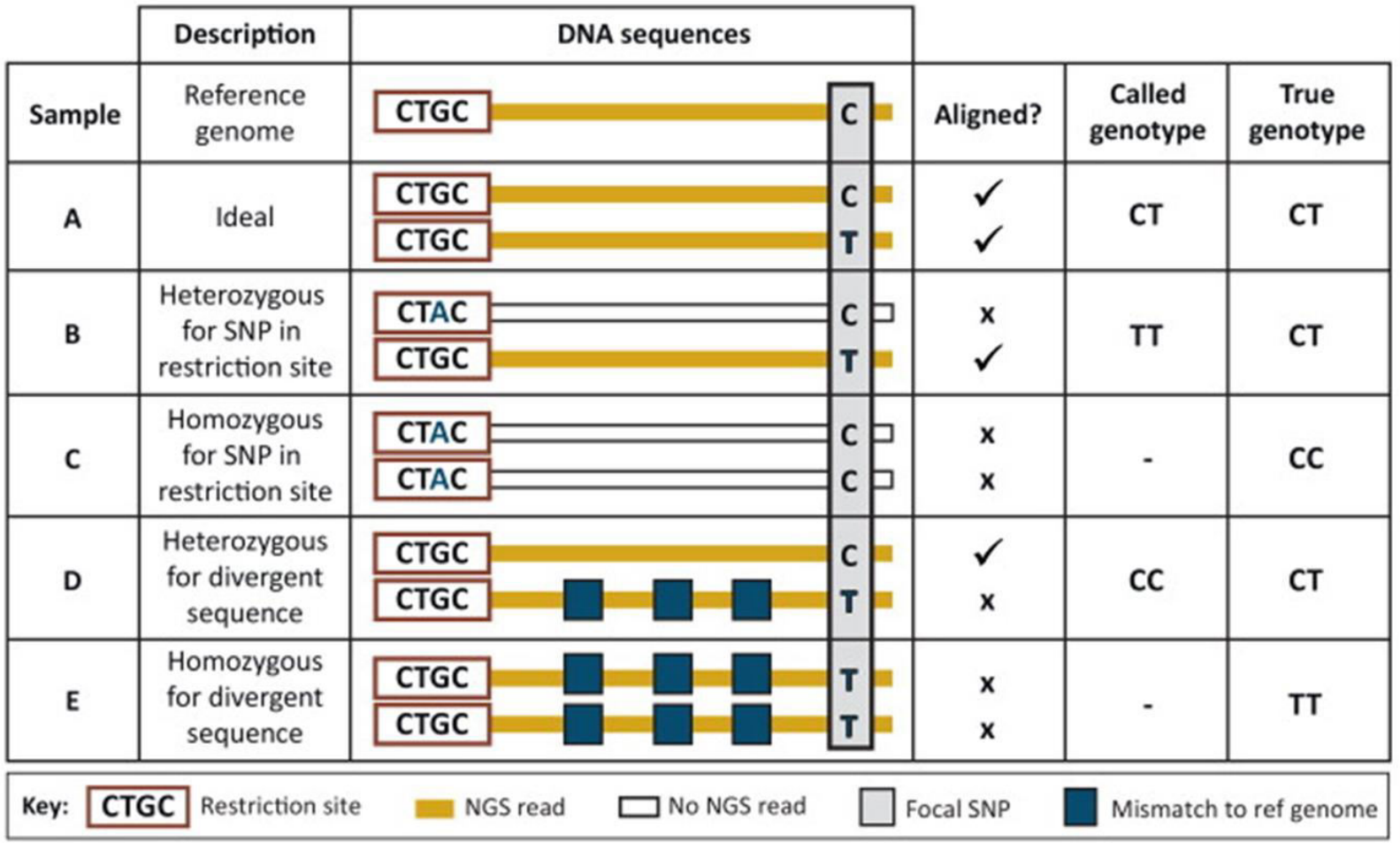
3) Repetitive sequences and paralogs introduce g
   - Exclude unusable repetitive sequences be
   - Develop algorithm to filter out paralogs

4) Imputation is still challenging
   - Imputation not require with our new proto
   - Mating design powerful for improving imputation and phasing



Poland and Rife (2012) The Plant Genome

GBS pitfalls



code as null allele, genotype accurately using read depth info i.e. genotype is "A-" and not "AA"

Completely missing alleles from one parent. Might still be useful if alternate parent is heterozygote

Problem with alignment of paralogs to the same locus. Filter and discard SNPs based on:

i) Read depth of paralogs that will be multiple times of the average read depth.

ii) Segregation distorted loci. Final data set should reveal distorted loci in few localized clusters.

Other technical issues not particularly due to GBS:

- illumina sequencing error
- chloroplast contanimation

Outline: New GBS pipeline

1) Pre-library prep:

*DNA quality check *select enzyme combination
*barcode/adapter design

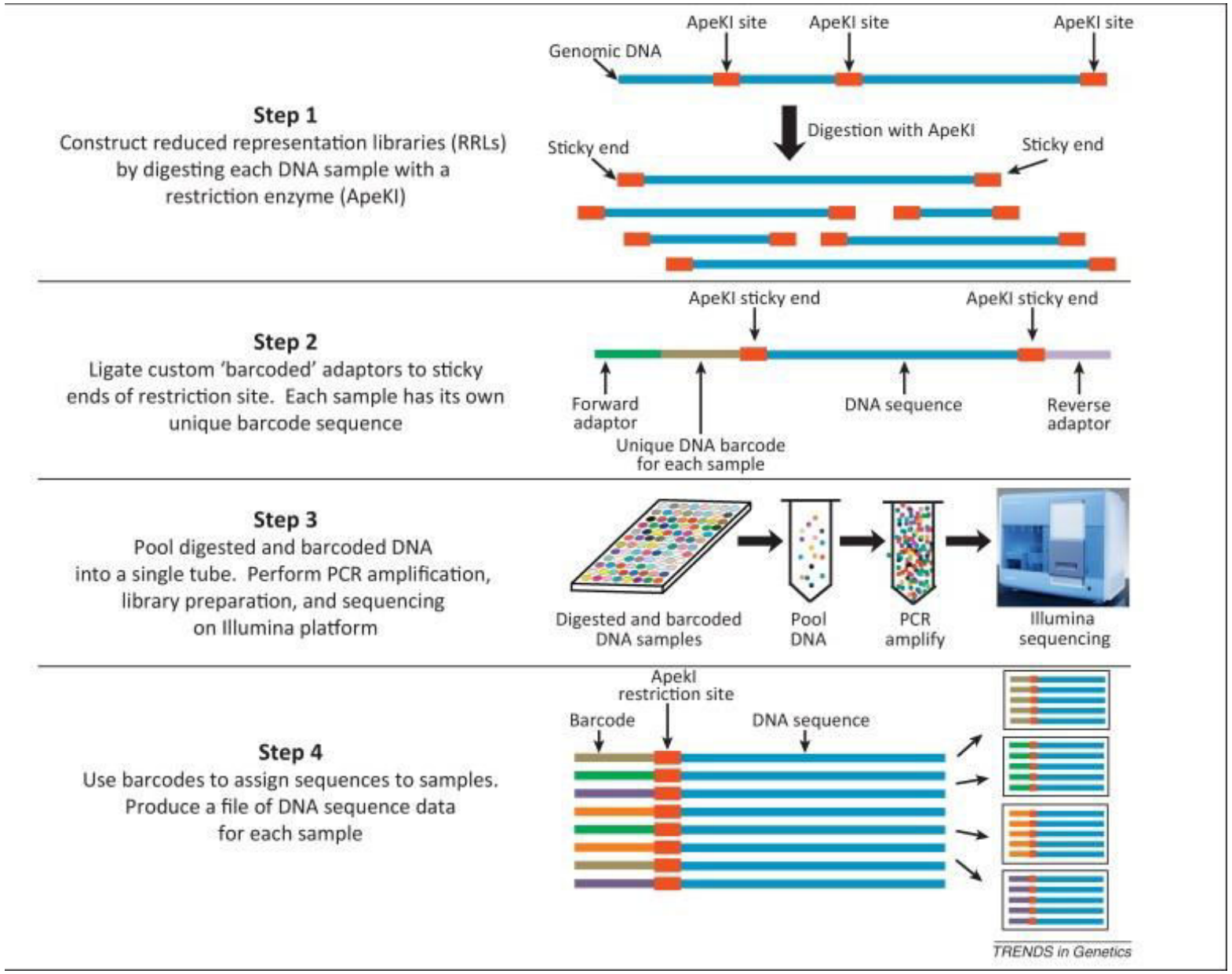
2) Library prep:

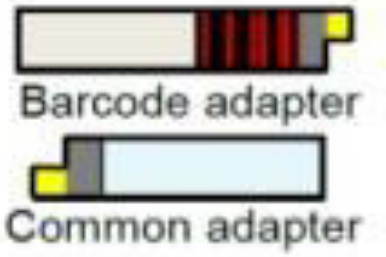
*double digest *adapter/barcode Ligation *size selection
*PCR amplification *Illumina sequencing

3) SNP calling (GATK-based pipeline):

*pre-processing reads (trimming low quality, de-multiplexing)
* Align to reference genome *call SNP genotypes
*filtering for high confidence/quality SNPs

GBS Principles





- De-multiplex pooled samples with barcodes.
- Additional barcode on common/reverse adapter can increase plex-levels
- Double digest more efficient.

Outline

1) Pre-library prep:

*DNA quality check *select enzyme combination
*barcode/adapter design

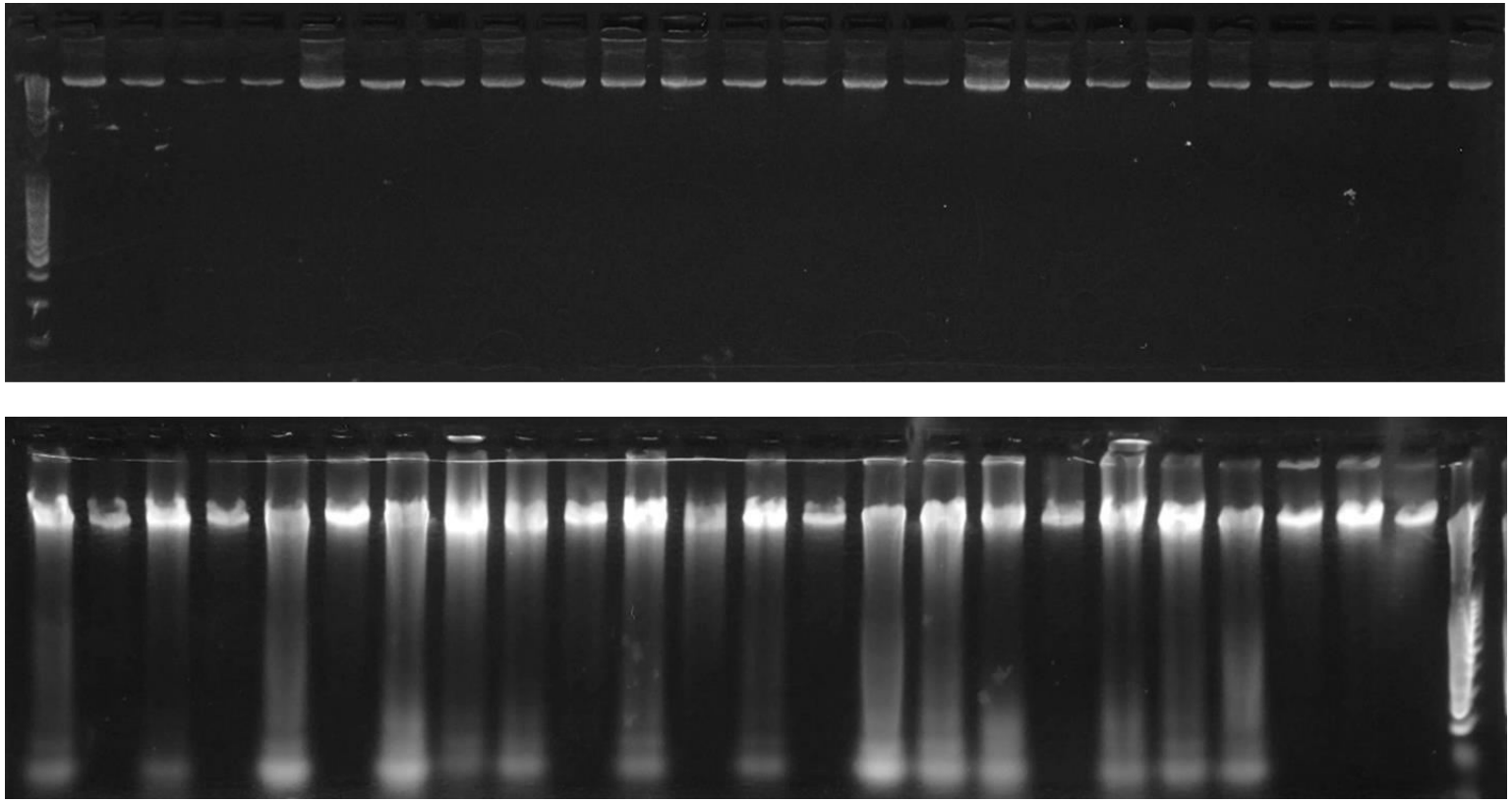
2) Library prep:

*double digest *adapter/barcode Ligation *size selection
*PCR amplification *Illumina sequencing

3) SNP calling (GATK-based pipeline):

*pre-processing reads (trimming low quality, de-multiplexing)
* Align to reference genome *call SNP genotypes
*filtering for high confidence/quality SNPs

Pre-library prep: DNA quality check



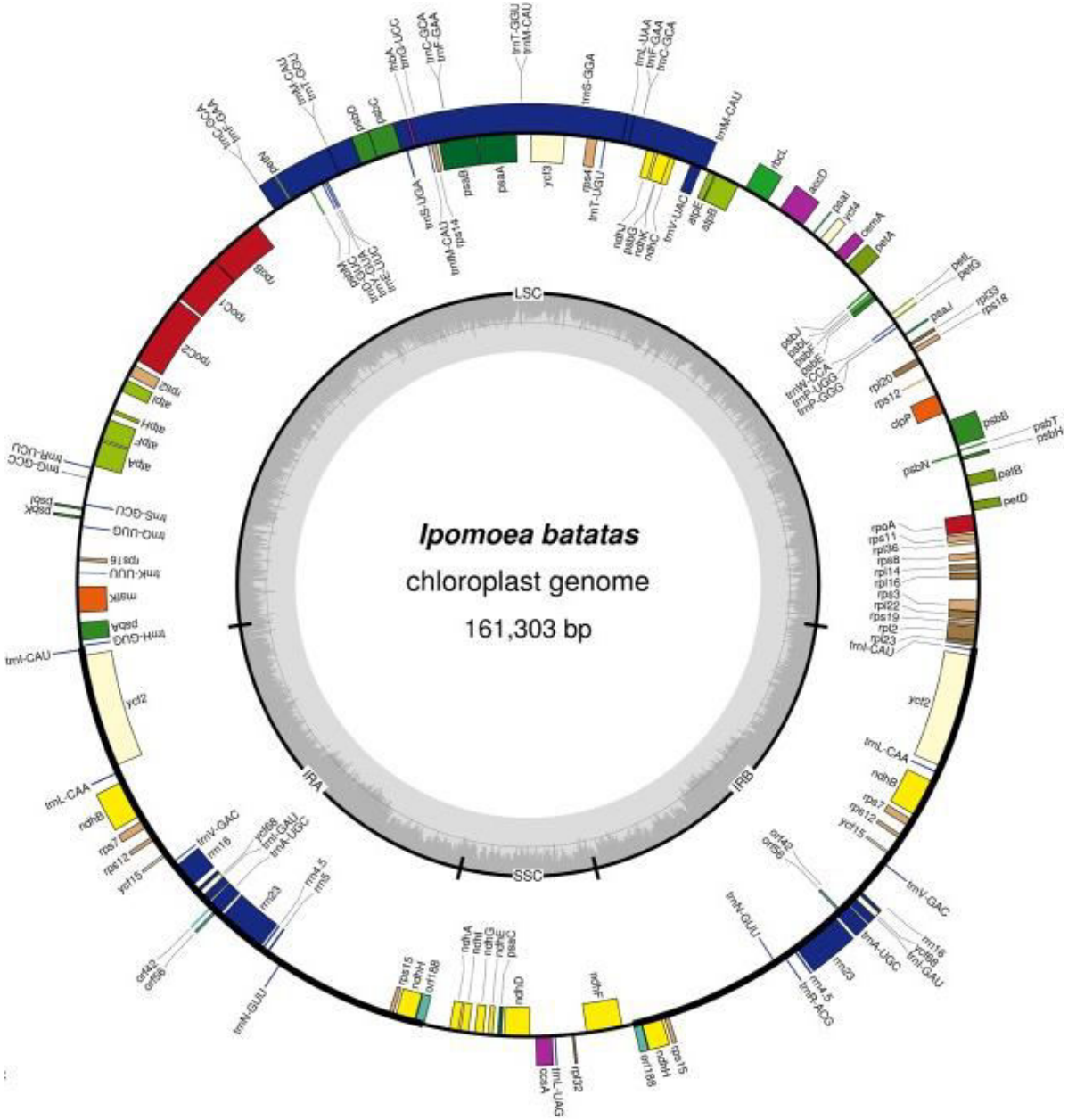
1) Visualize DNA on agarose gel: “no smearing and no RNA is good”
2) Assay like “picogreen” should be used for DNA quantification.
3) Ensure DNA is in low EDTA buffer

Pre-library prep: select enzyme combination

Significance of enzyme combination choice:

- contamination of library with chloroplast DNA.

- leads to low proportion of reads matching reference nuclear genome.



Yan et al. (2015) PLoS One

Pre-library prep: *in silico digest*

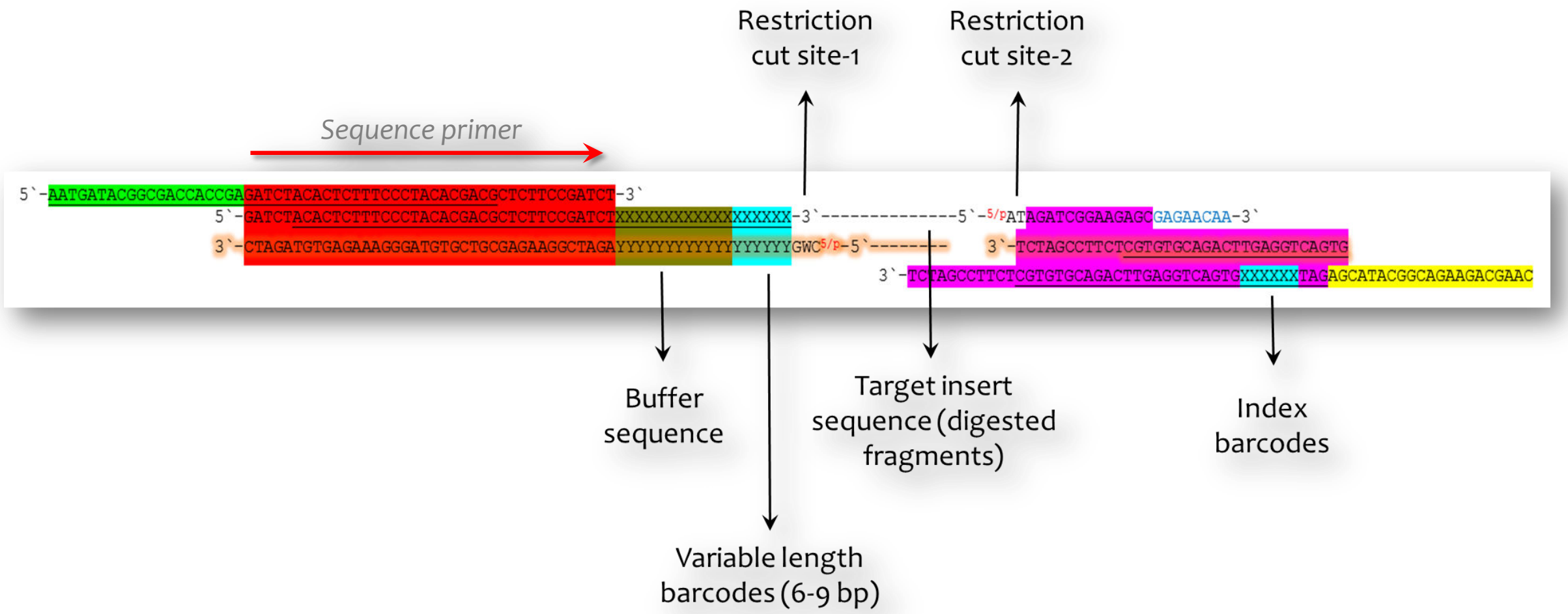
Minimize fragments from chloroplast in library.

1 copy of chloroplast genome

Enzyme combinations	Chloroplast	Trifida	Triloba	Methylation	% of chloroplast fragment
PstI	1	8,086	8,138	Sensitive	0.01
MluI/CviAII	5	18,290	19,588	Sensitive	0.03
Pst1/MspI	15	31,791	29,365	Sensitive	0.05
SphI/MluCI	14	53,456	52,770	Insensitive	0.03
EcoRI/MspI	84	60,381	58,116	Sensitive	0.14
EcoRI/CviAII	106	120,225	116,531	Sensitive	0.09
TseI/CviAII	99	339,660	329,539	Sensitive	0.03

window size (from 160 bp)	Total number of Fragments	
	TseI/CviAII	PstI/MspI
50	61,052	4,753
100	111,367	8,954
150	150,800	12,714
200	182,895	16,126
250	207,110	19,183
300	227,133	22,188
350	243,264	24,883
400	256,128	27,349

Pre-library prep: barcode/adapter design



1) Index barcode to increase plex-level (not very efficient).
2) Barcodes designed to destroy cut site upon ligation.
3) Secondary digest to eliminate chimeric ligation products.

Pre-library prep: buffer sequence

- 12 bp buffer sequence upstream of barcode
- Absorb inflated error at beginning of illumine reads
- Ensures nucleotide diversity crucial for good quality reads

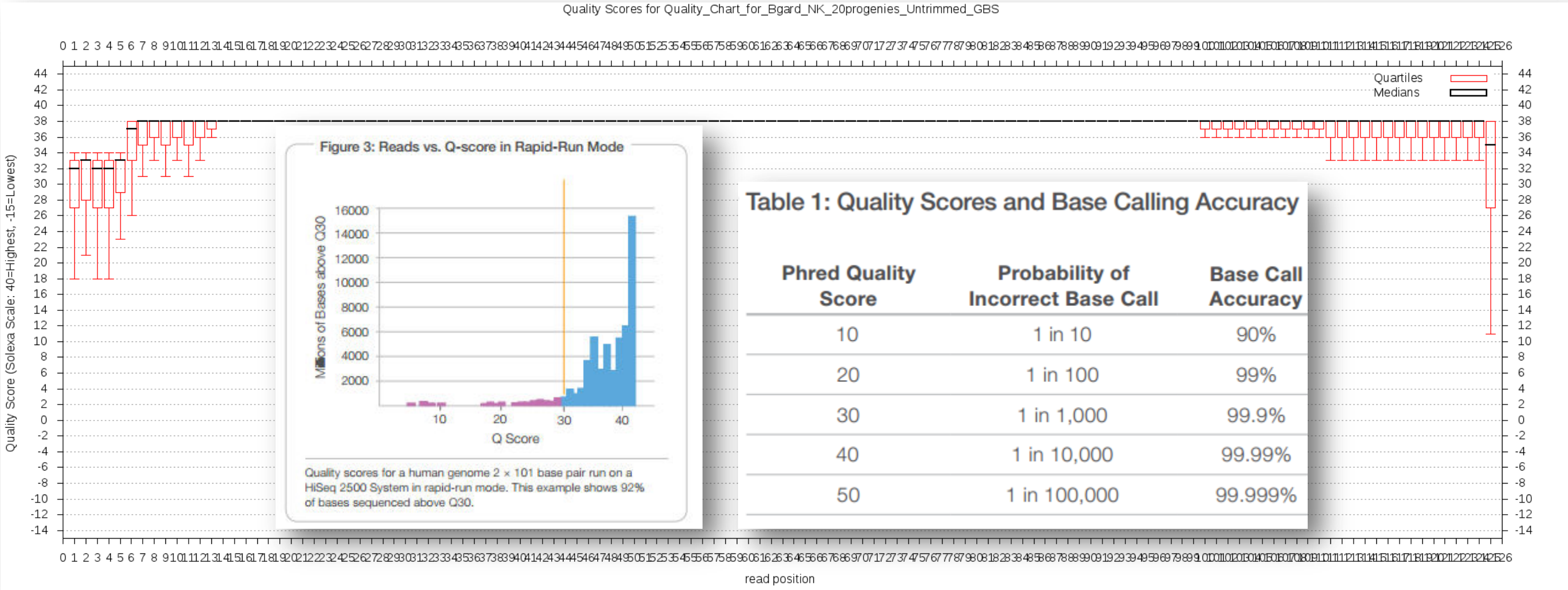


Figure 3: Reads vs. Q-score in Rapid-Run Mode

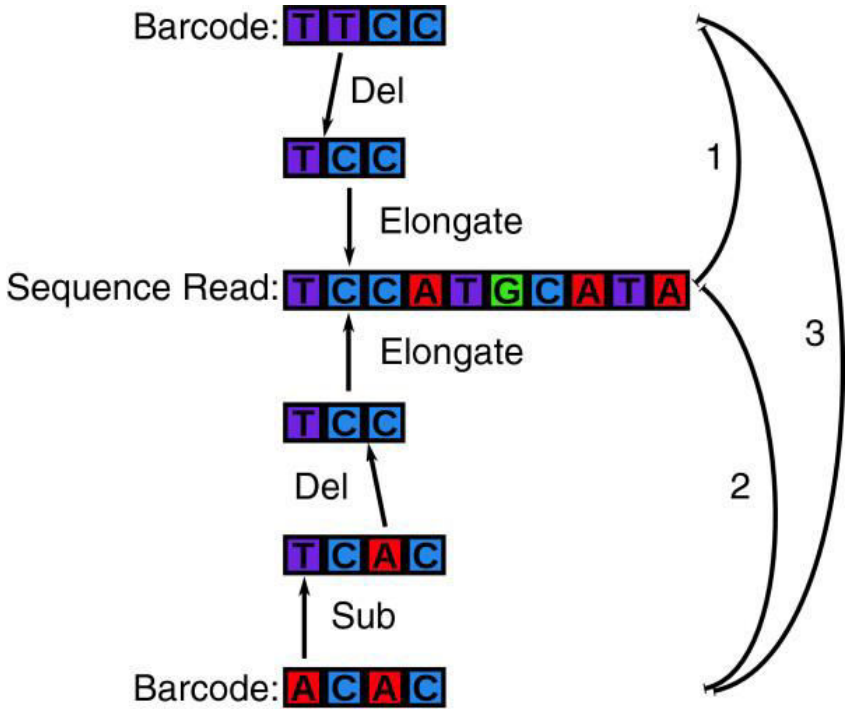
Quality scores for a human genome 2 × 101 base pair run on a HiSeq 2500 System in rapid-run mode. This example shows 92% of bases sequenced above Q30.

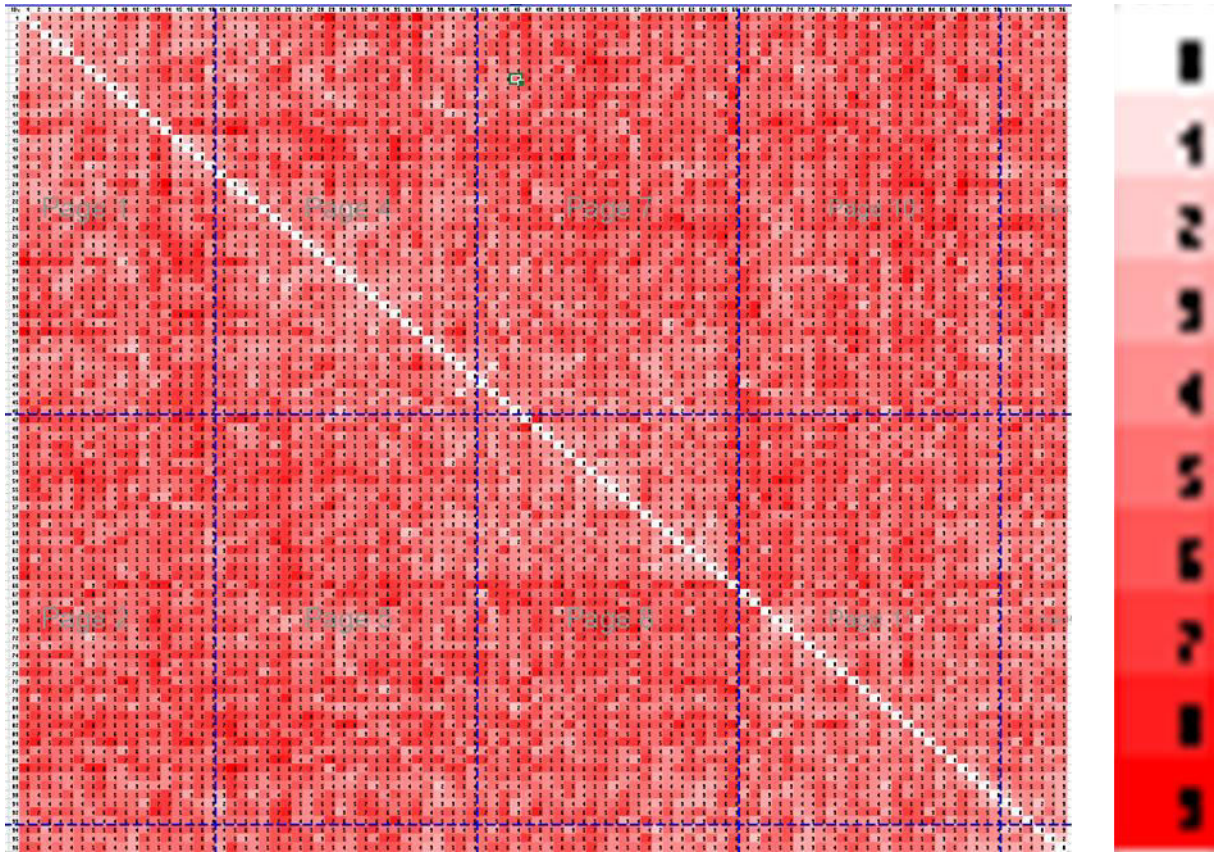
Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Pre-library prep: barcode sequence

- Variable length (6-9 nt) barcode (designed with R-script)
- Accounts for substitution and indel errors (edit/levenstein distance)
- Better than Hamming distance (only substitution error)





Outline

1) Pre-library prep:

*DNA quality check *select enzyme combination
*barcode/adapter design

2) Library prep:

*double digest *adapter/barcode Ligation *size selection
*PCR amplification *Illumina sequencing

3) SNP calling (GATK-based pipeline):

*pre-processing reads (trimming low quality, de-multiplexing)
* Align to reference genome *call SNP genotypes
*filtering for high confidence/quality SNPs

Library prep: Library prep

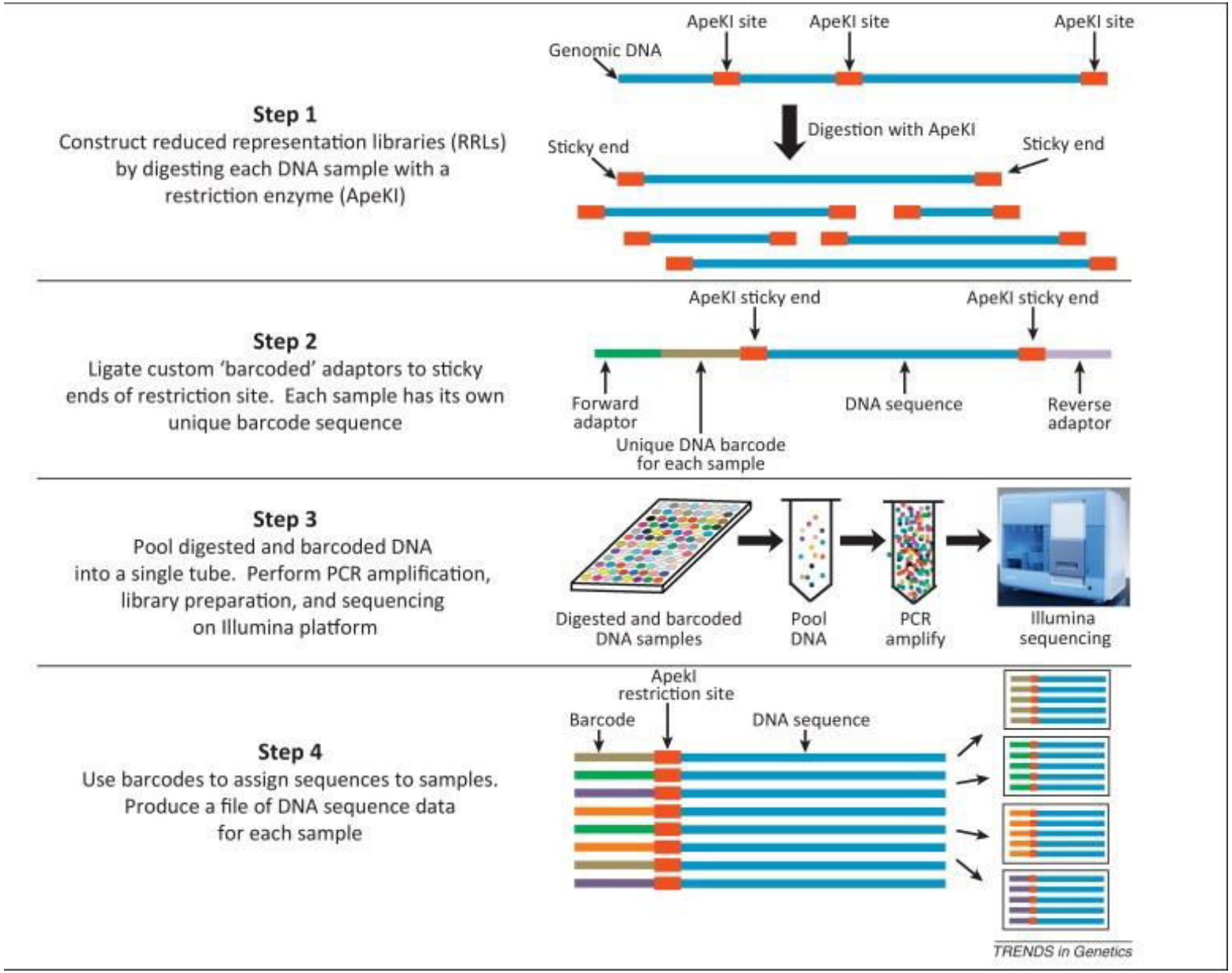
*double digest

*adapter/barcode Ligation

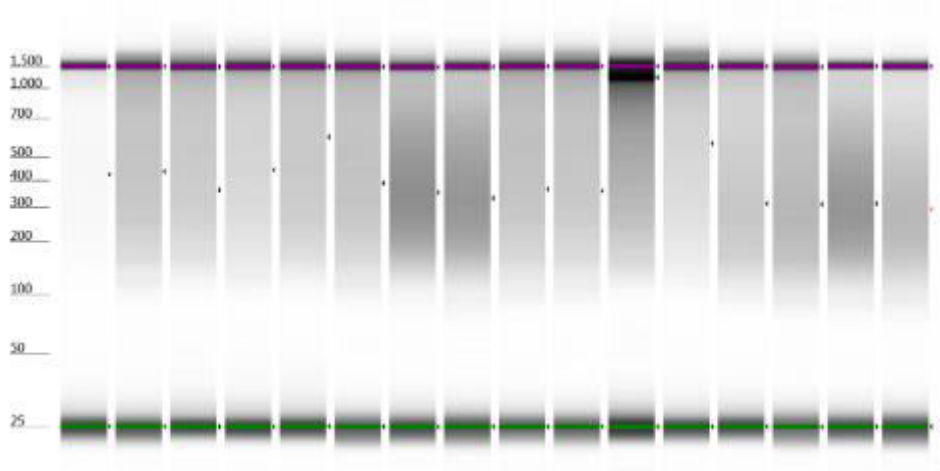
*size selection

*PCR amplification

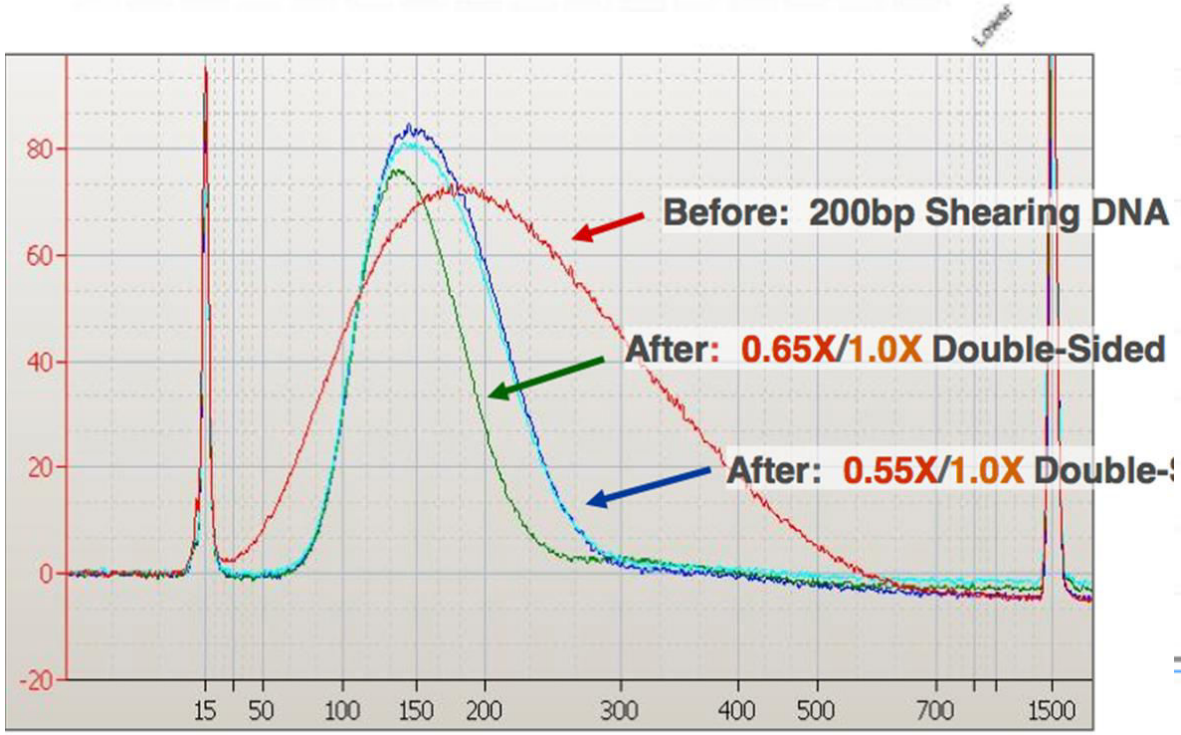
*Illumina sequencing

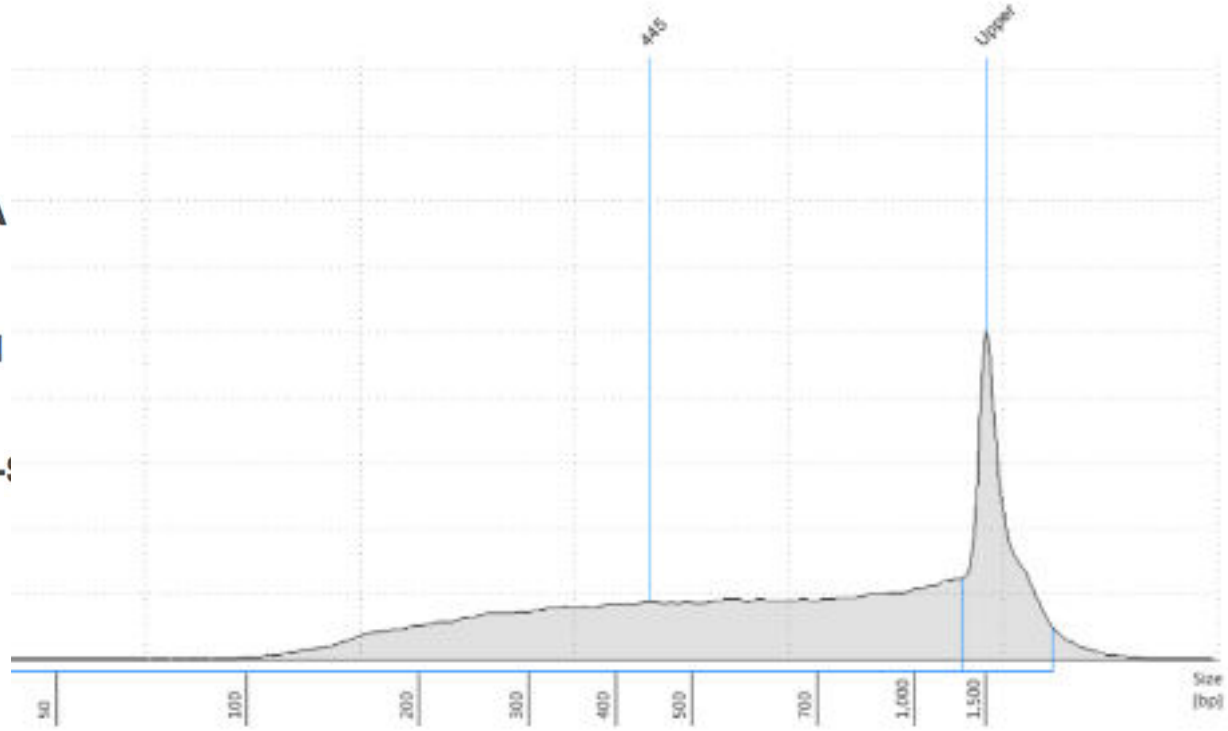


Library prep: double digest



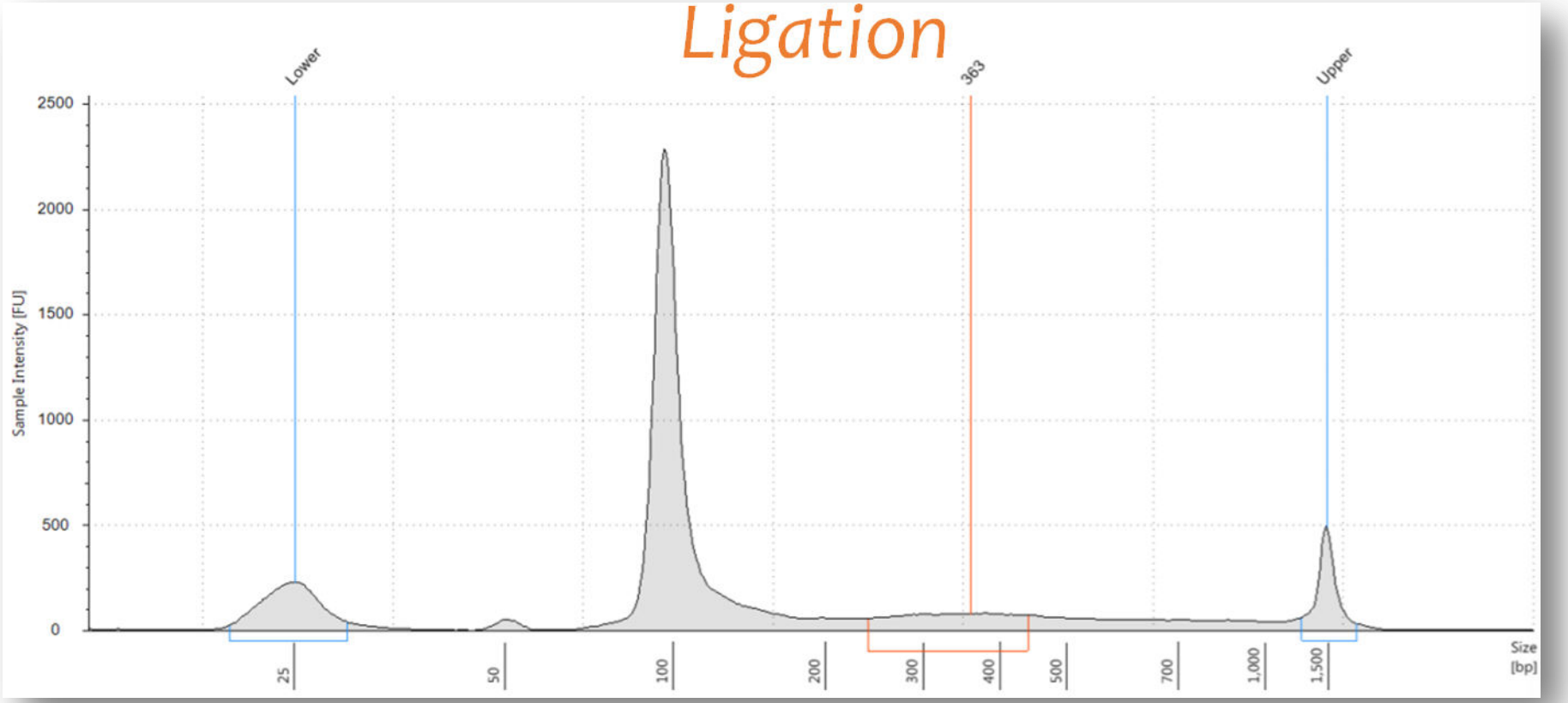
1) Easy absorbance-based DNA quantification
2) Double digest
3) MagBead clean-up
4) Normalize samples by **“florescence-based Picogreen quantification”** before ligation.



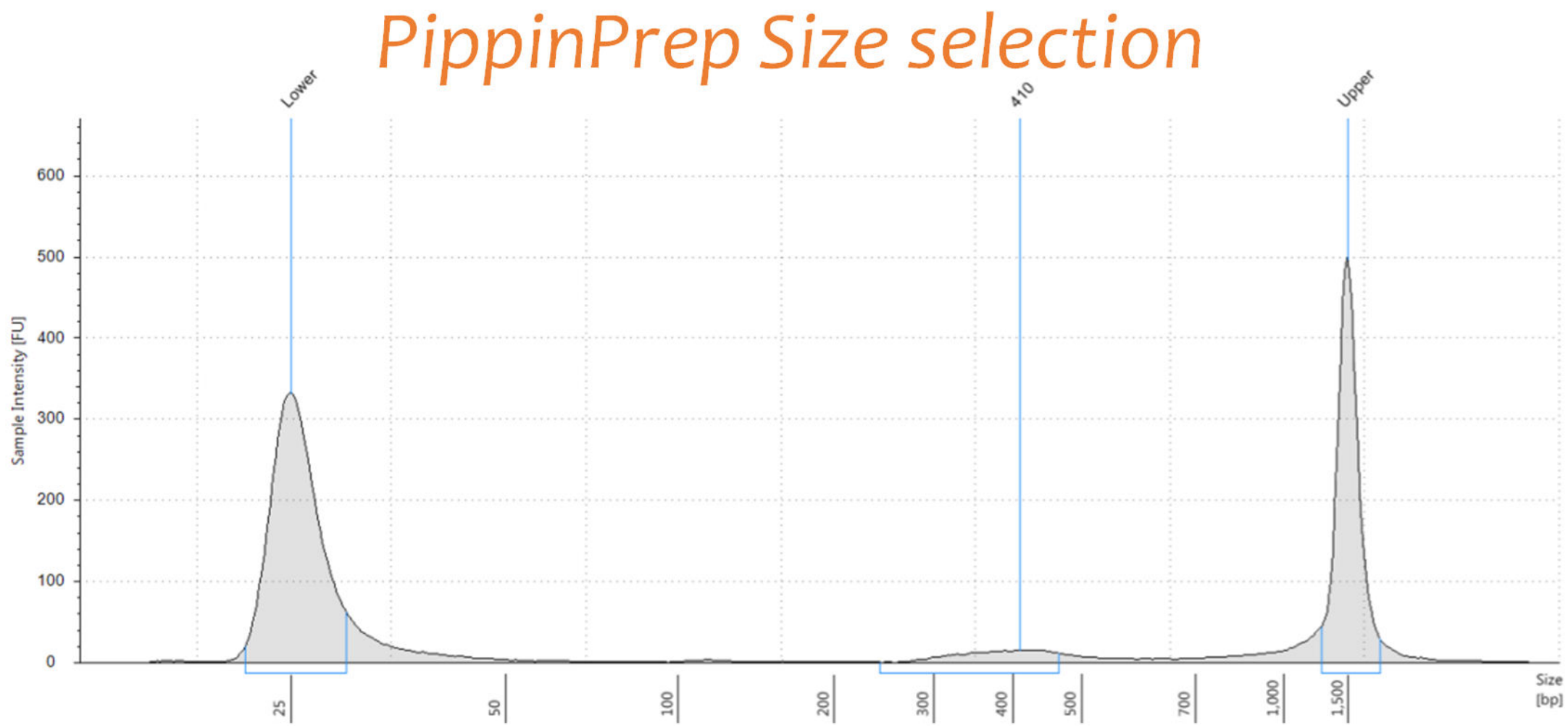


Library prep: Ligation

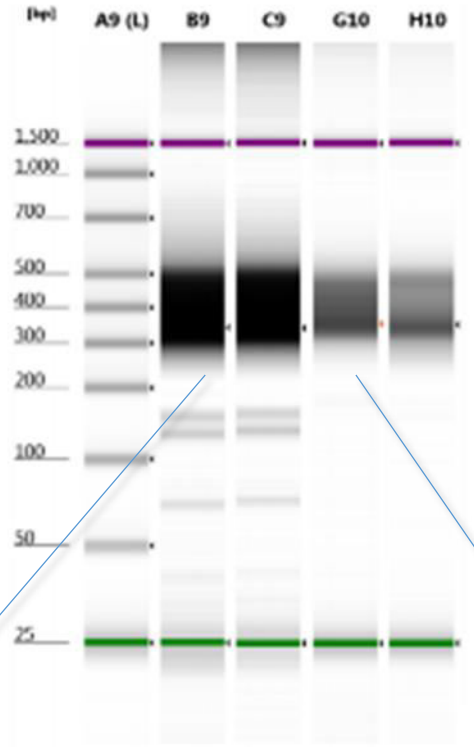
1) Ligate barcoded adapter to each normalized sample
2) Pooled samples on plate-by-plate basis
3) Perform secondary digest on pooled samples to eliminate chimeric fragment, which will not align/match to reference genomic sequence



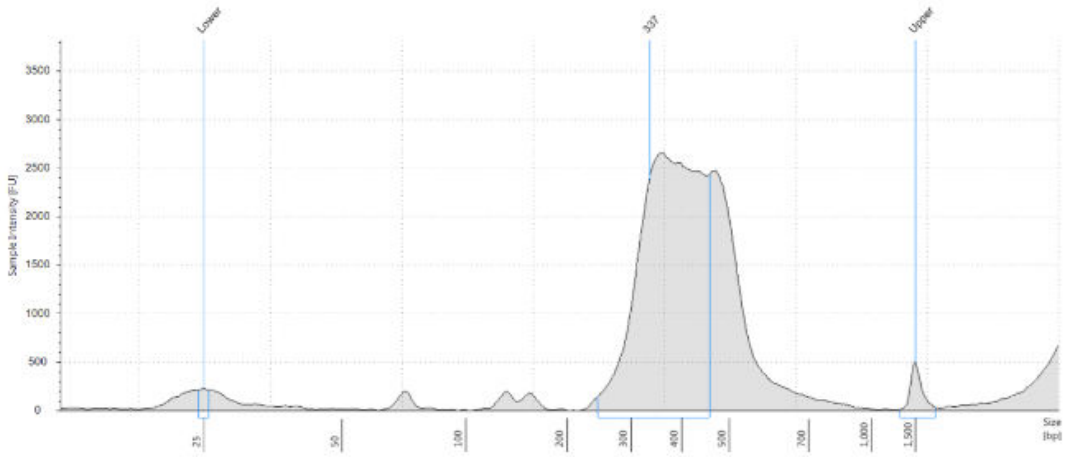
Library prep: Pippin Prep size selection, PCR and cleanup



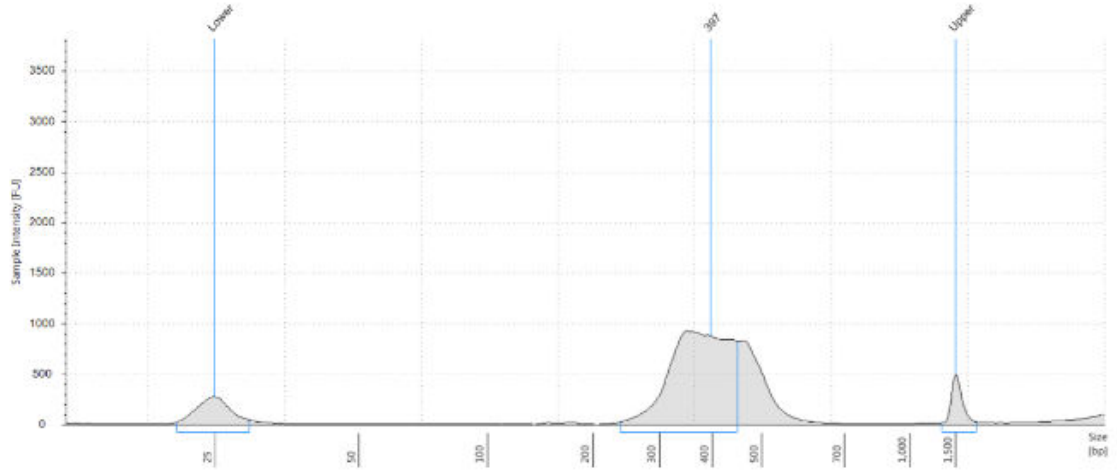
Library prep: Pippin Prep size selection, PCR and cleanup



PCR on ligated fragments



MagBead clean-up



Outline

1) Pre-library prep:

*DNA quality check *select enzyme combination
*barcode/adapter design

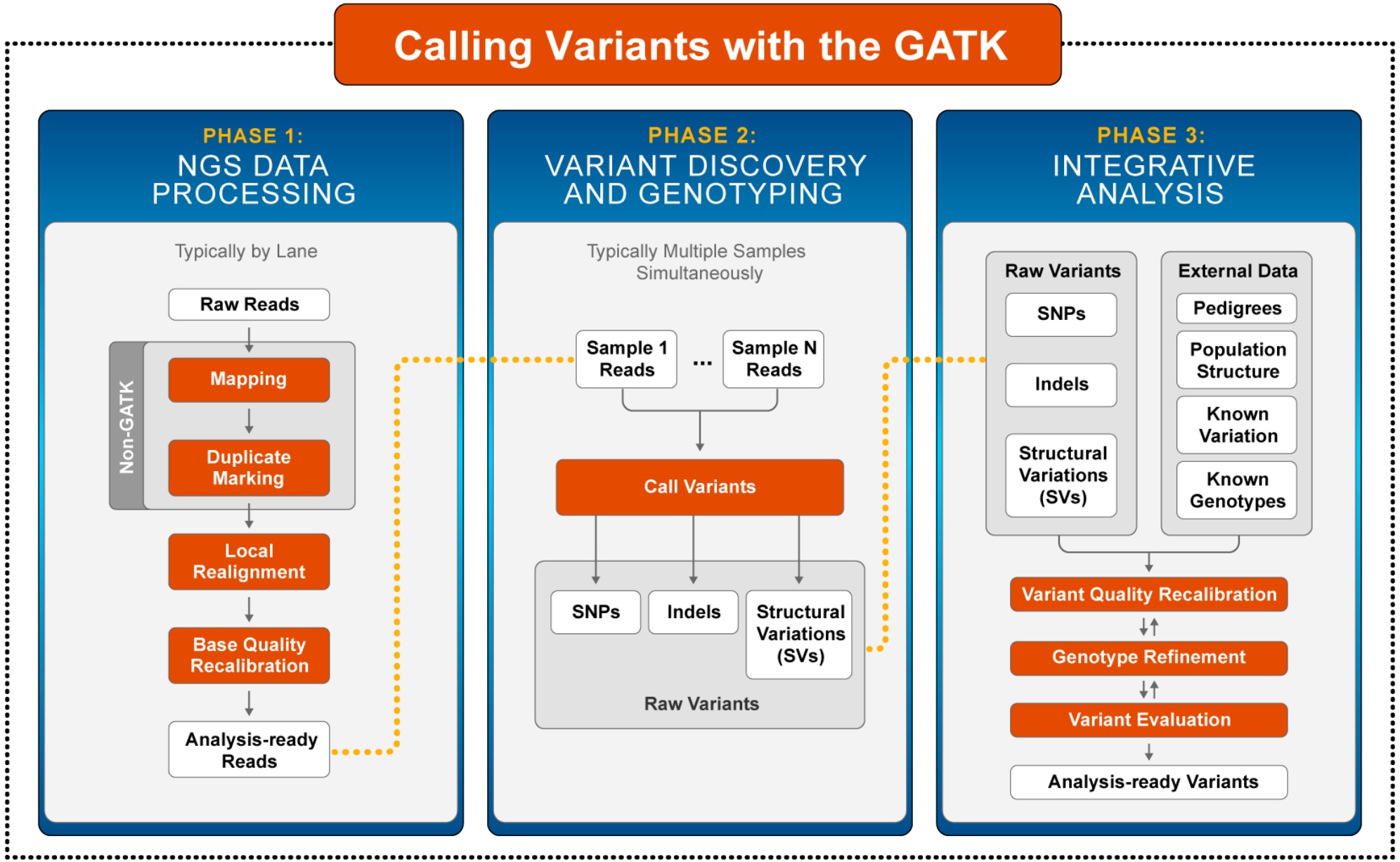
2) Library prep:

*double digest *adapter/barcode Ligation *size selection
*PCR amplification *Illumina sequencing

3) SNP calling (GATK-based pipeline):

*pre-processing reads (trimming low quality, de-multiplexing)
* Align to reference genome *call SNP genotypes
*filtering for high confidence/quality SNPs

SNP calling: GATK



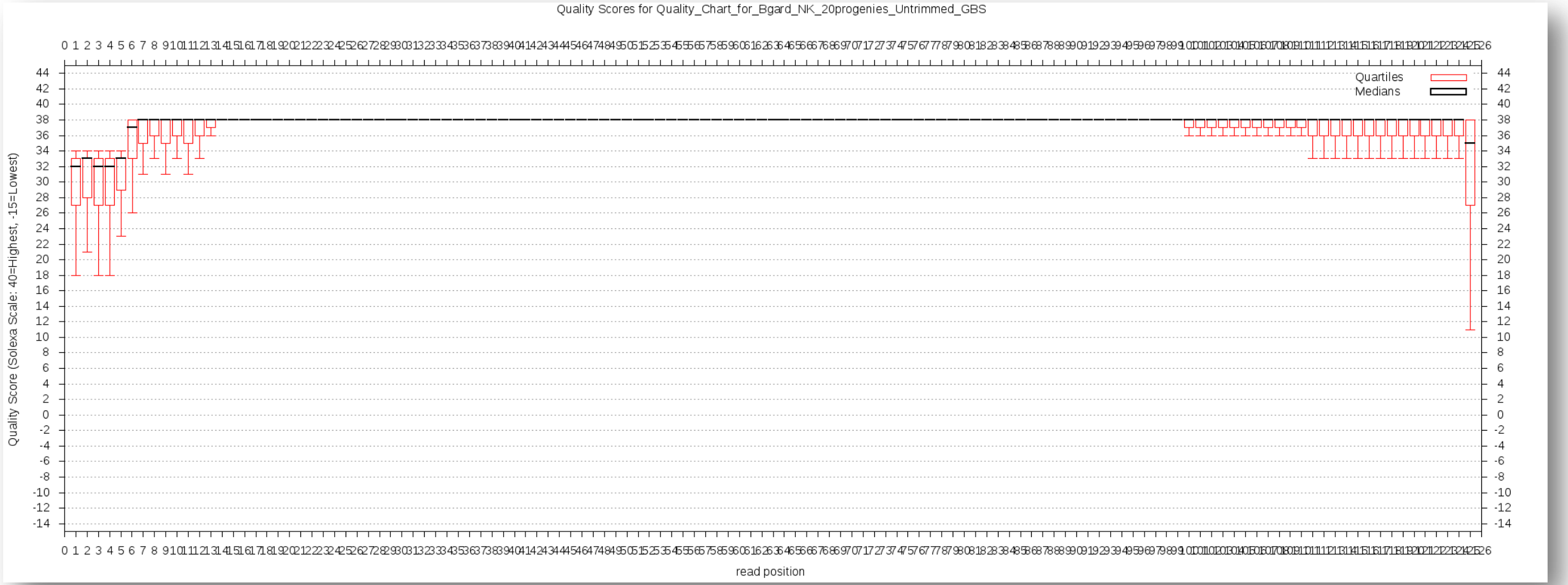
SNP calling: pipeline

1) Quality check on raw reads (FASTX toolkit)
2) De-multiplex (FASTX toolkit)
3) Trim off barcodes and low quality bases (FASTX toolkit)
4) Index reference genome (BWA and SAMtools) and Create sequence dictionary (Picard tools)
5) Align short illumina reads from each sample (BWA)
6) Convert output SAM file to BAM file (Picard tools)
7) Produce statistics on alignment (SAMtools)
8) Mark duplicates (Picard tools)
9) Sort BAM file (Picard tools)
10) Add group header information to each BAM file (Picard tools)
11) Index the resulting BAM file (SAMtools)
12) Re-assign quality scores if scale not matched to GATK scale (GATK)
13) Indel re-alignment (GATK)
14) Call variants using HaplotypeCaller (GATK)
15) Data filtering (VCF tools)
16) Summary boxplot/beanplot (R)

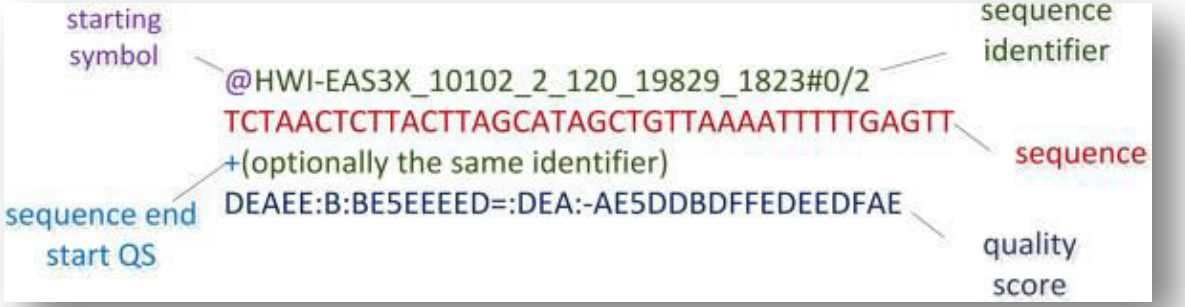
Tools:

- FASTX toolkit
- BWA
- SAMtools
- Picard tools
- GATK
- VCF tools
- Packes in R software

SNP calling: quality check (QC)



SNP calling: De-multiplexing



De-multiplex :

Mismatch ≤ 0 86.9 % of reads recovered

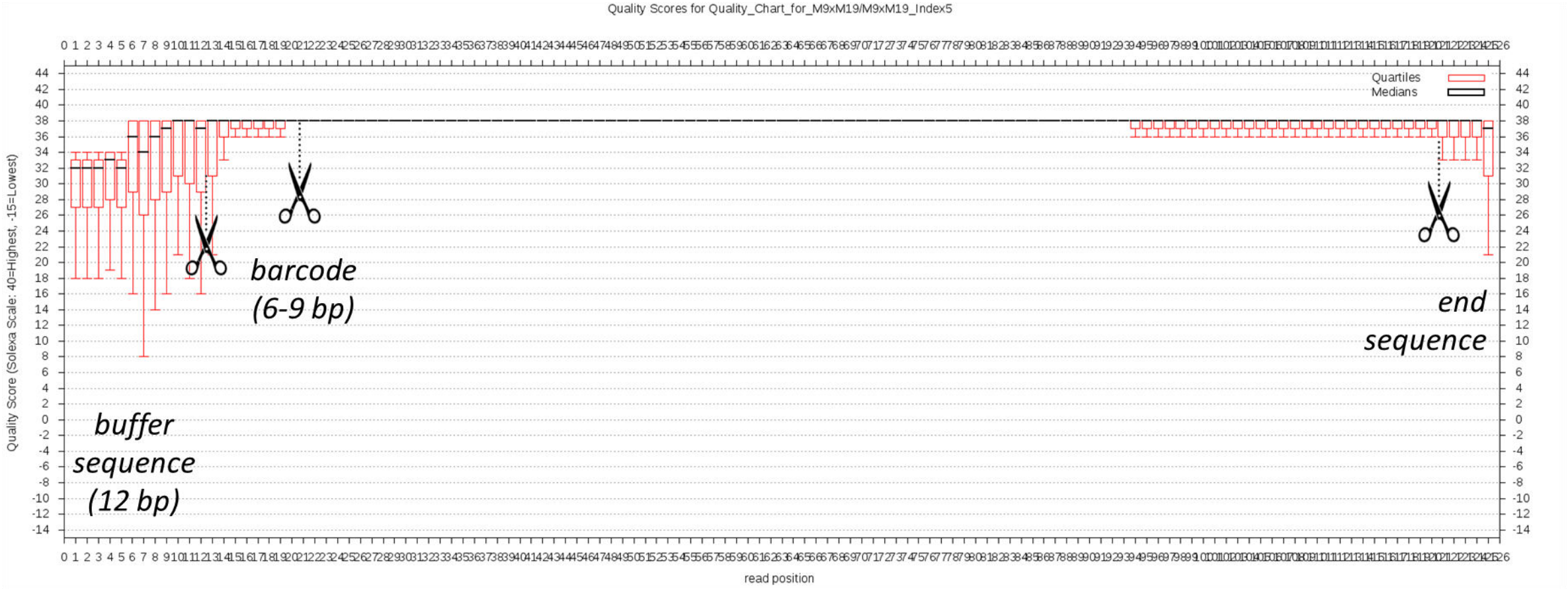
Mismatch ≤ 1 97.5 % of reads recovered

Demultiplex based on 12 bp buffer sequence

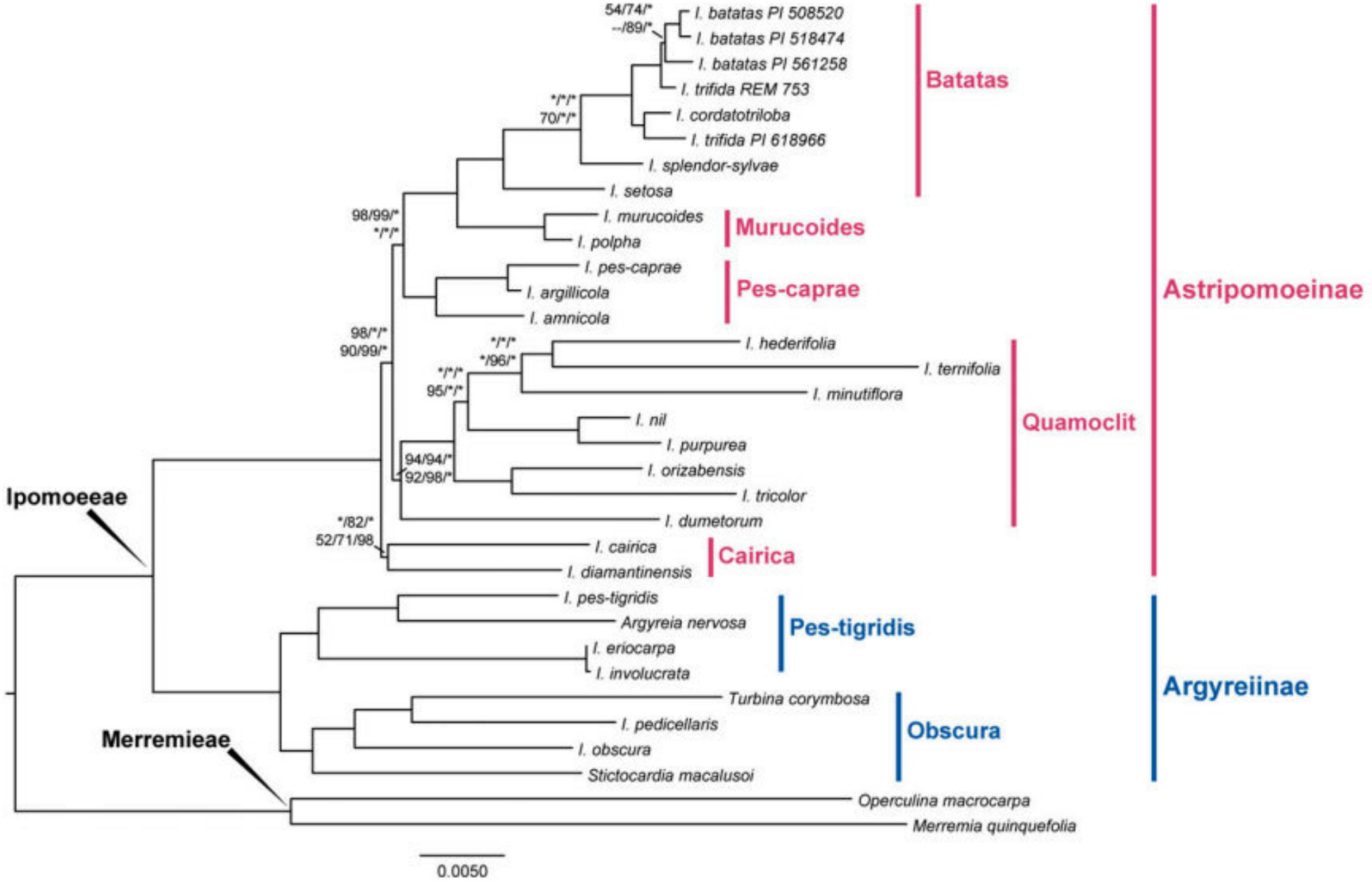
Mismatch ≤ 0 48 % of reads recovered

Mismatch ≤ 1 67 % of reads recovered

SNP calling: Trim reads



SNP calling: Index reference genomes and alignment



Eserman et al. (2014) American Journal of Botany.

- Whole chloroplast genome sequencing
- 29 morning glory species

SNP calling: reference genomes and alignments

- Alignment of reads to nuclear and chloroplast genome:
  - Average match to nuclear genome -> 90.8 %
  - Average match to chloroplast genome -> 11.2 %

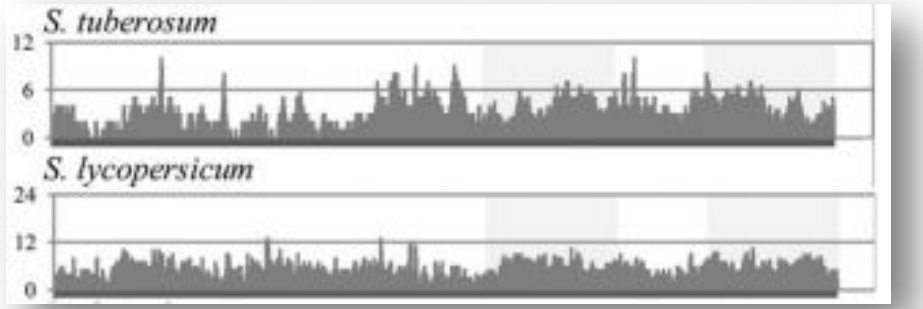
  - nuclear plastid DNA-like sequences (NUPTs) probably account for overlap

Samples	**M9; M19**	**M9; M19**
Reference genome	Trifida	Triloba
% of reads matching reference	90.98; 90.87	93.71; 93.60
Number of Sites	68,411	66,563
Proportion Missing	0	0
Proportion Heterozygous	0.719	0.727
Genetic distance	0.5142	0.5140

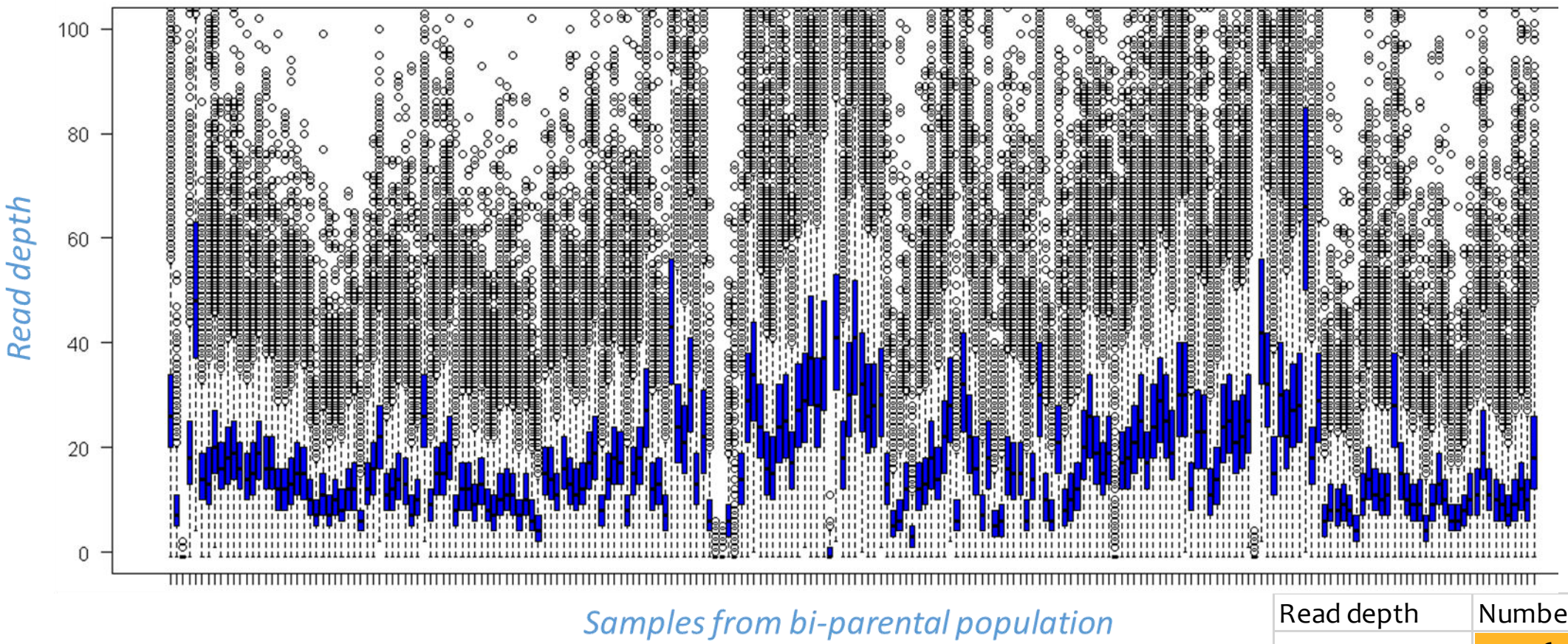
SNP calling: reference genomes and alignments

Species	Genome size (Mb)[a]		NUPTs		Proportion to nuclear genome (%)
	Nuclear	Chloroplast	Number	Length (kb)	
A. thaliana	119	0.15	38 (31)	17.7 (14.5)	0.015 (0.012)
C. papaya	343	0.16	613 (486)	269.8 (216.2)	0.079 (0.063)
V. vinifera	486	0.16	900 (497)	337.7 (209.0)	0.069 (0.043)
L. japonicus	301	0.15	394 (392)	147.3 (147.0)	0.049 (0.049)
M. truncatula	567	0.12	361	477.8	0.084
G. max	974	0.15	1435	406.3	0.042
M. esculenta	533	0.16	199	54.4	0.010
R. communis	107	0.16	632	264.2	0.247
P. trichocarpa	481	0.16	293	241.8	0.050
C. sativus	203	0.15	169	49.0	0.054
F. vesca	195	0.16	218	58.2	0.030
S. tuberosum	727	0.16	563	429.6	0.059
S. lycopersicum	782	0.16	1513	674.4	0.084
B. distachyon	271	0.14	863	531.5	0.196
O. sativa	382	0.13	611 (495)	846.6 (804.3)	0.222 (0.210)
S.bicolor	697	0.14	515 (417)	169.4 (142.3)	0.024 (0.020)
Z. mays	2066	0.14	1459 (1099)	1041.3 (880.4)	0.050 (0.043)

In parentheses, the estimated values are shown when mitochondria–chloroplast transferred DNAs were excluded.
[a]Accumulative length of determined whole-genome sequences in database.

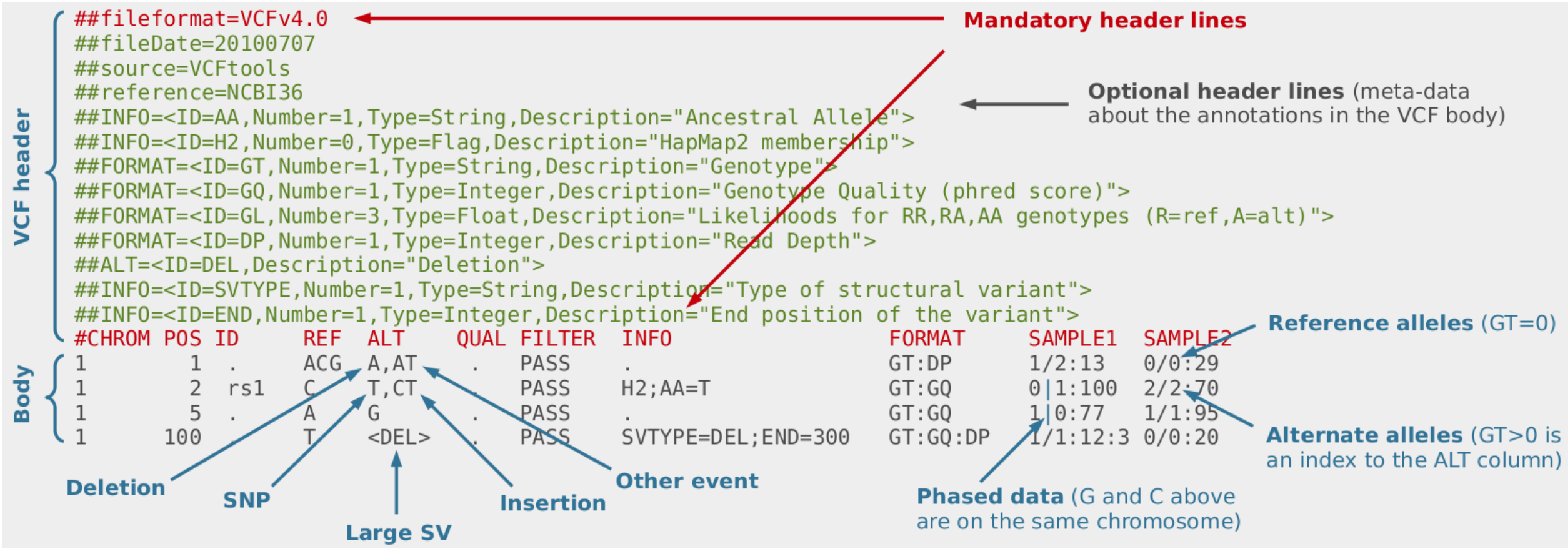


Uniform read depth across SNPs and samples



Read depth	Number of SNPs
1	190,369
2	139,990
3	113,492
4	94,611
5	80,074
After filtering	27,761

VCF file format



Processing VCF files with VCFtools

1) Recode SNPs as missing if read depth is below threshold. Also, if SNP has too many reads, recode as missing (probably a paralog):
   - Diploid = 5 reads
   - Hexaploid = 30 reads

2) Filter based on missing data: No more than 20% missing

3) Decide if markers should be strictly bi-allelic. Also decide if you want to retain indels.

4) Extract genotype calls, read depth and alleles from VCF file.

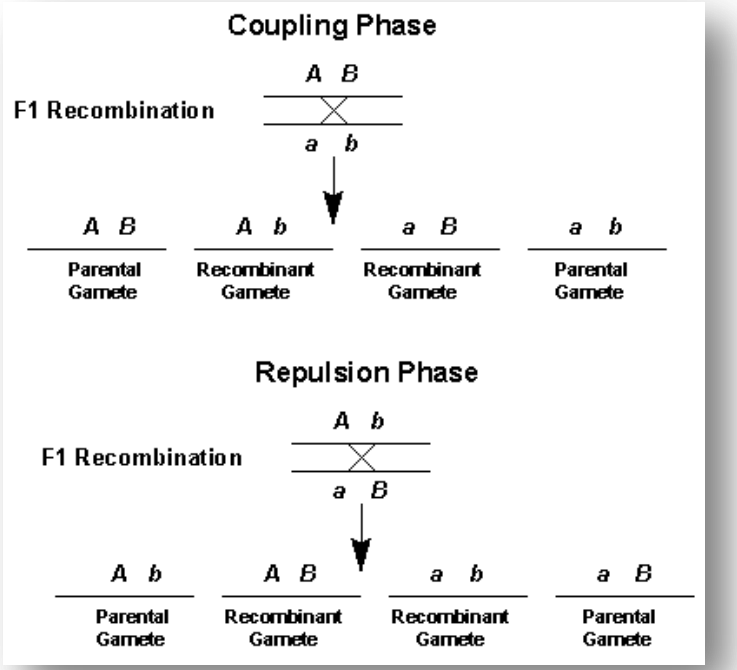
5) Determine segregation ratio for SNPs and use this parameter to clean up data for SNPs that have segregation distortion (polySegratio: R-package). Data is ready for statistical analysis.

Utility of GBS SNP markers

1) Diversity study
2) Linkage disequilibrium and re-constructing haplotypes
3) Constructing genetic linkage map
4) QTL analysis
5) Association mapping:
6) Genomic selection

Linkage Maps Construction

1) Recode data to match coding nomenclature of software (JoinMap)
2) Create dummy loci to capture all possible linkage phase
3) Group markers into 15 groups matching Trifida chromosomes
4) Group markers with right matching linkage phase
5) Order SNP markers
6) Evaluate map with plot of pairwise “recombination frequency” and “LOD” to detect problematic markers
7) Correct linkage map



M9 xM19 Linkage Maps

Linkage Group	M9		M19	
	# markers	**map length (cM)**	**# markers**	**map length (cM)**
1	246	151	228	118
2	200	132	158	92
3	195	119	174	95
4	131	111	117	80
5	144	104	166	84
6	161	98	212	90
7	170	95	183	79
8	218	94	225	100
9	142	86	124	70
10	169	86	191	78
11	117	79	158	90
12	53	72	81	69
13	175	69	187	81
14	47	48	36	38
15	24	26	20	20
Total	**2192**	**1,370 cM**	**2260**	**1,185 cM**

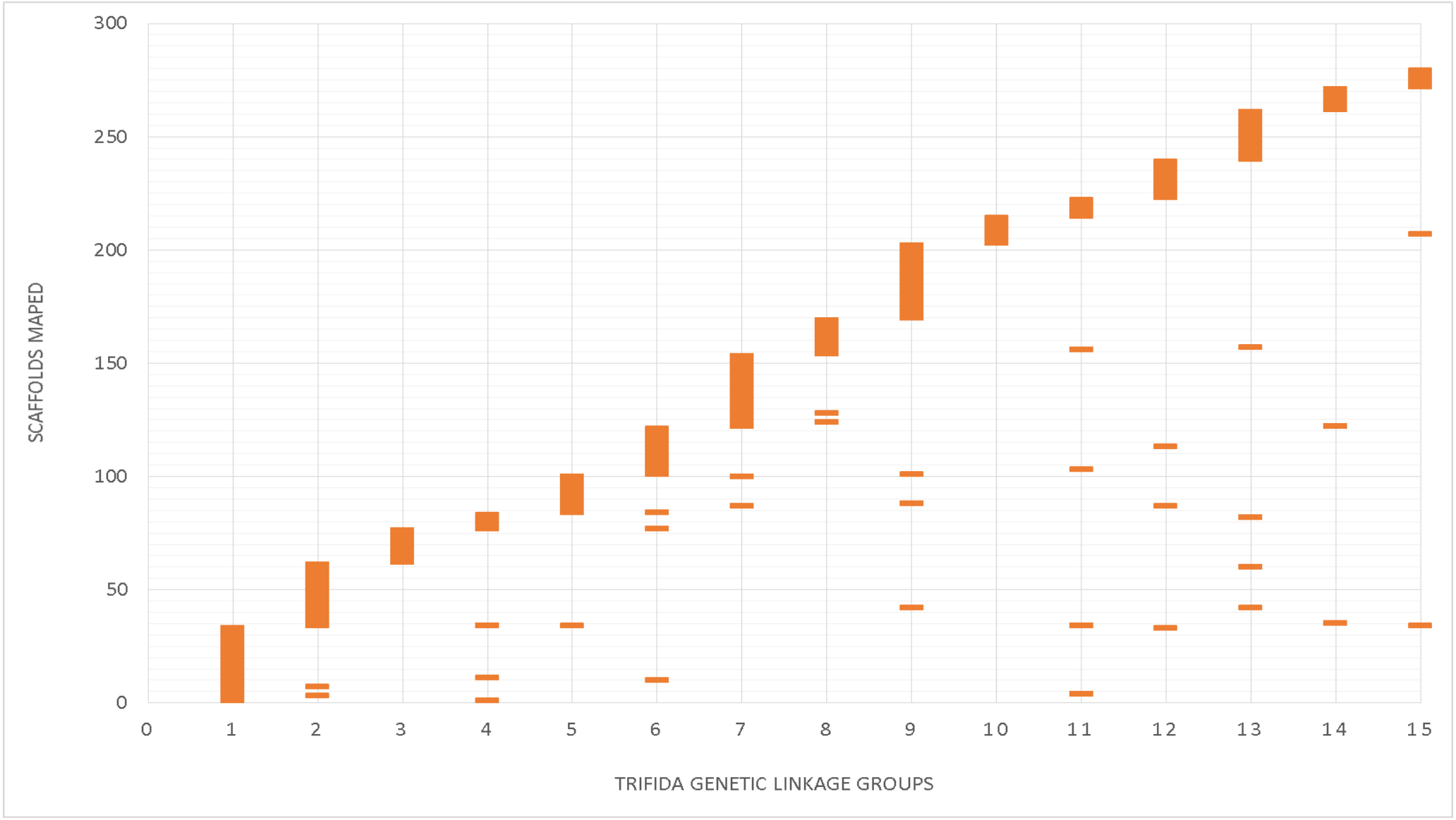
Genome coverage	322,659,957 bp (322.7 MB)
% of genome coverage	60.65
# of Scaffolds	279
# of SNPs	3221

M9 xM19 Linkage Maps

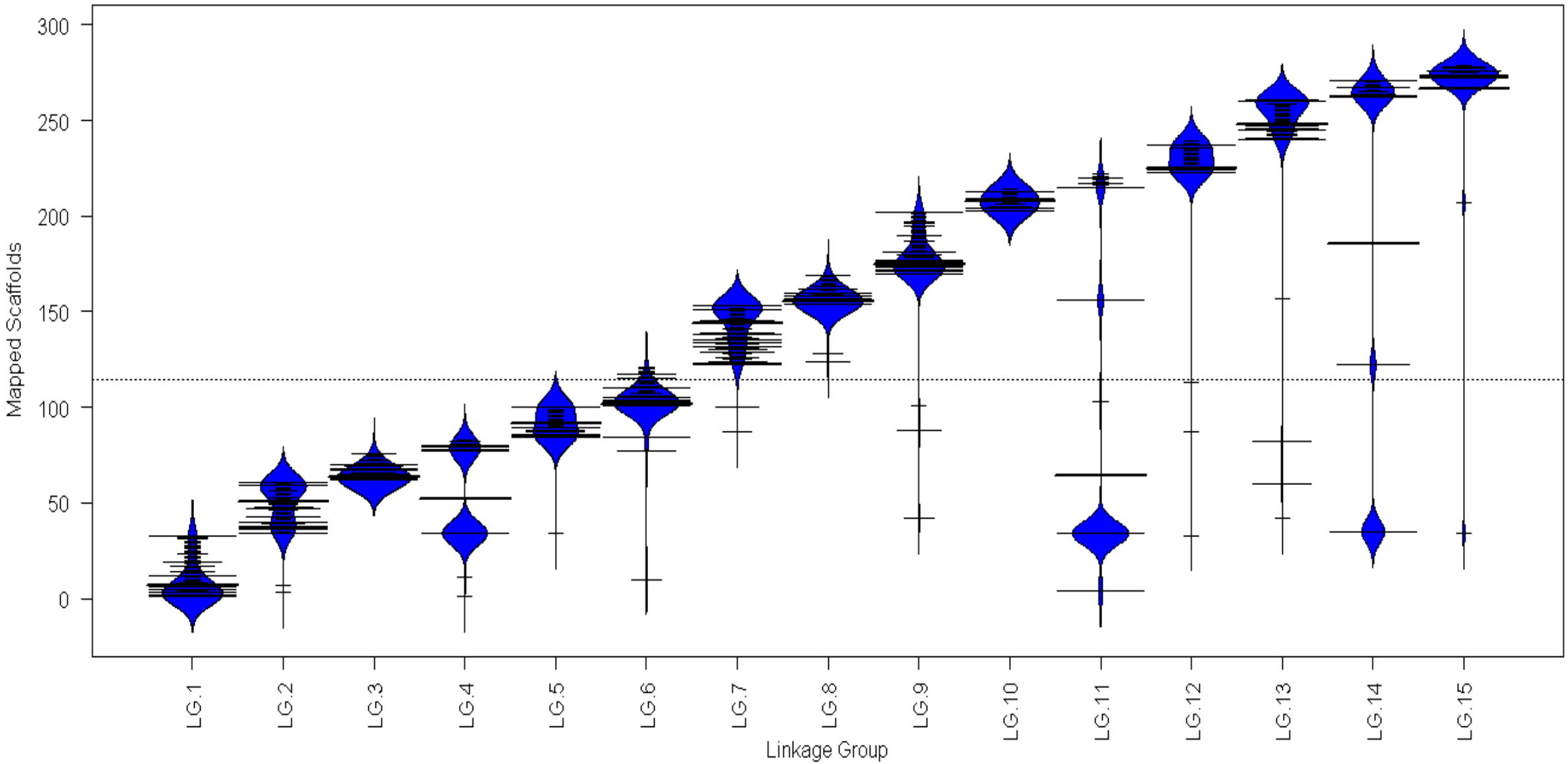
Linkage Group	M9		M19	
	# markers	**map length (cM)**	**# markers**	**map length (cM)**
1	246	151	228	118
2	200	132	158	92
3	195	119	174	95
4	131	111	117	80
5	144	104	166	84
6	161	98	212	90
7	170	95	183	79
8	218	94	225	100
9	142	86	124	70
10	169	86	191	78
11	117	79	158	90
12	53	72	81	69
13	175	69	187	81
14	47	48	36	38
15	24	26	20	20
Total	**2192**	**1,370 cM**	**2260**	**1,185 cM**

Genome coverage	322,659,957 bp (322.7 MB)
% of genome coverage	60.65
# of Scaffolds	279
# of SNPs	3221

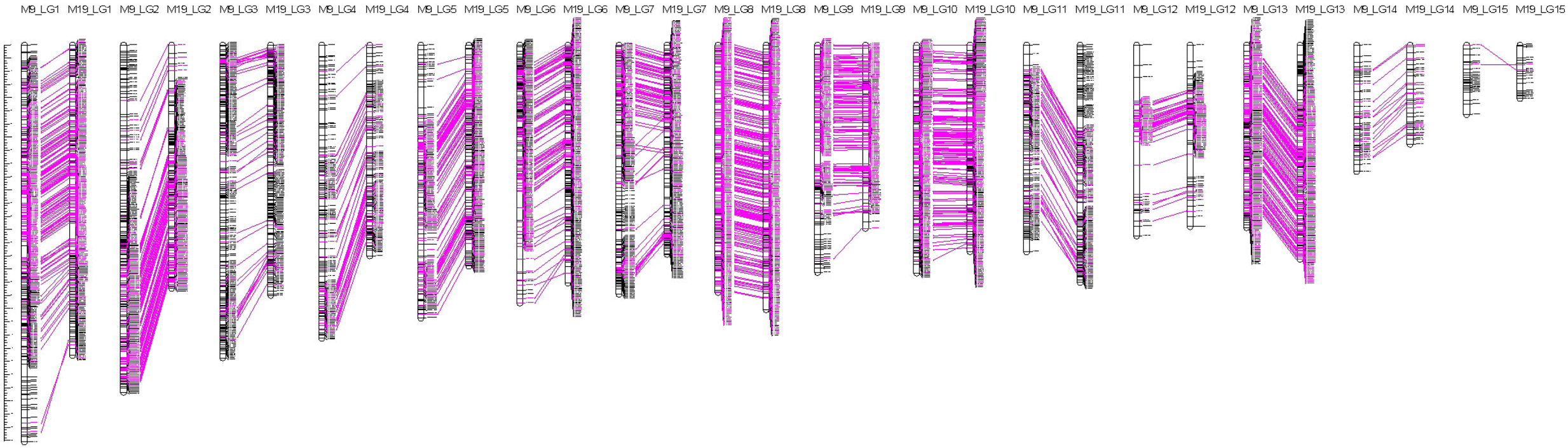
M9 xM19 Linkage Maps

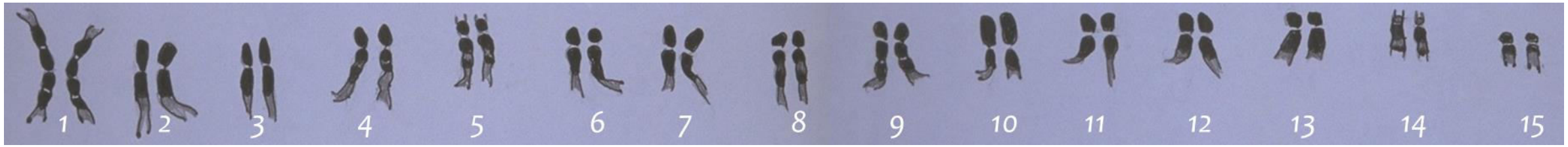


M9 xM19 Linkage Maps

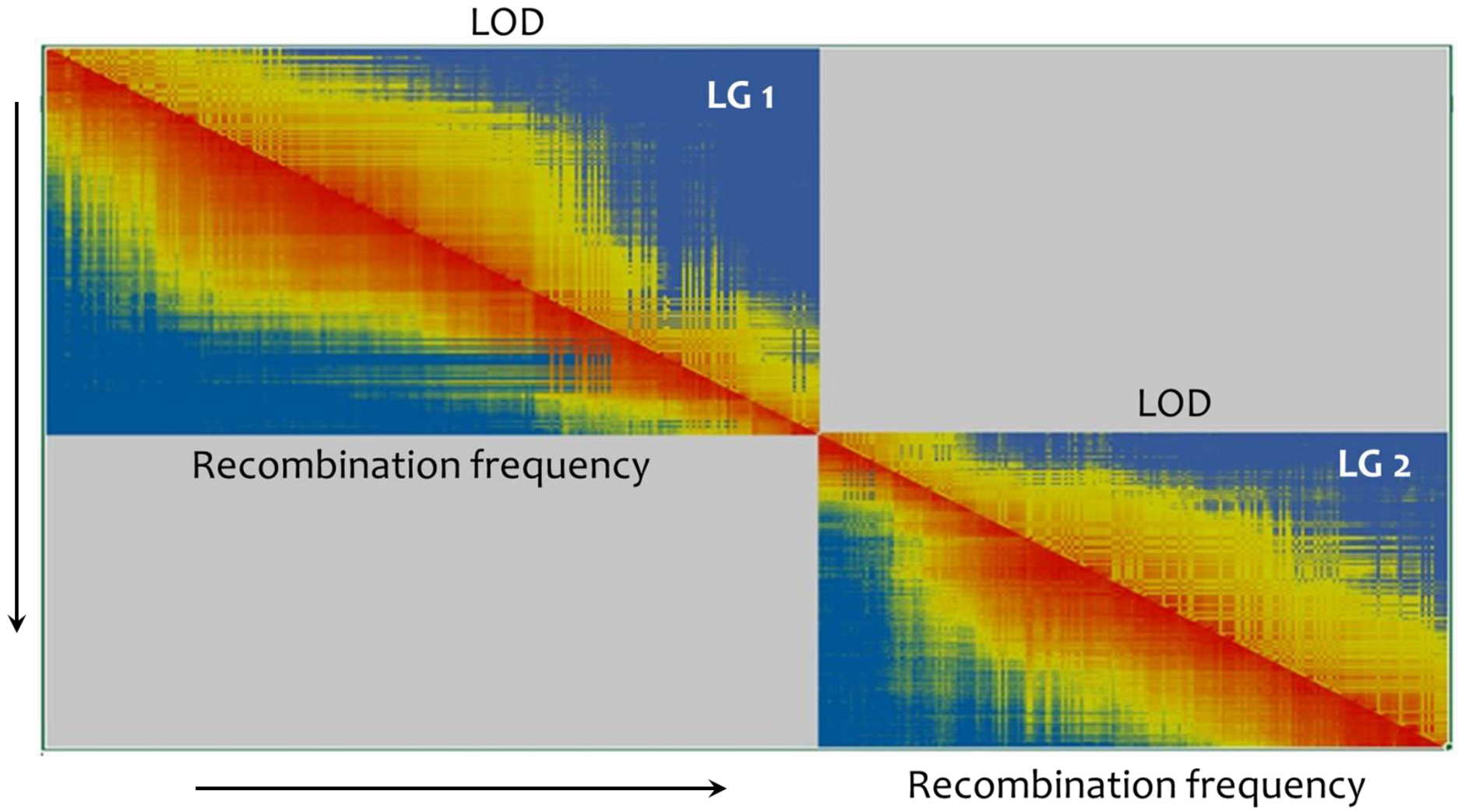


M9 xM19 Linkage Maps





M9 xM19 Linkage Maps



Rec. Freq	LOD
0	100
1	90
2	80
15	70
20	60
25	50
30	40
35	30
40	20
45	10
50	0