GENOMIC TOOLS FOR SWEETPOTATO IMPROVEMENT

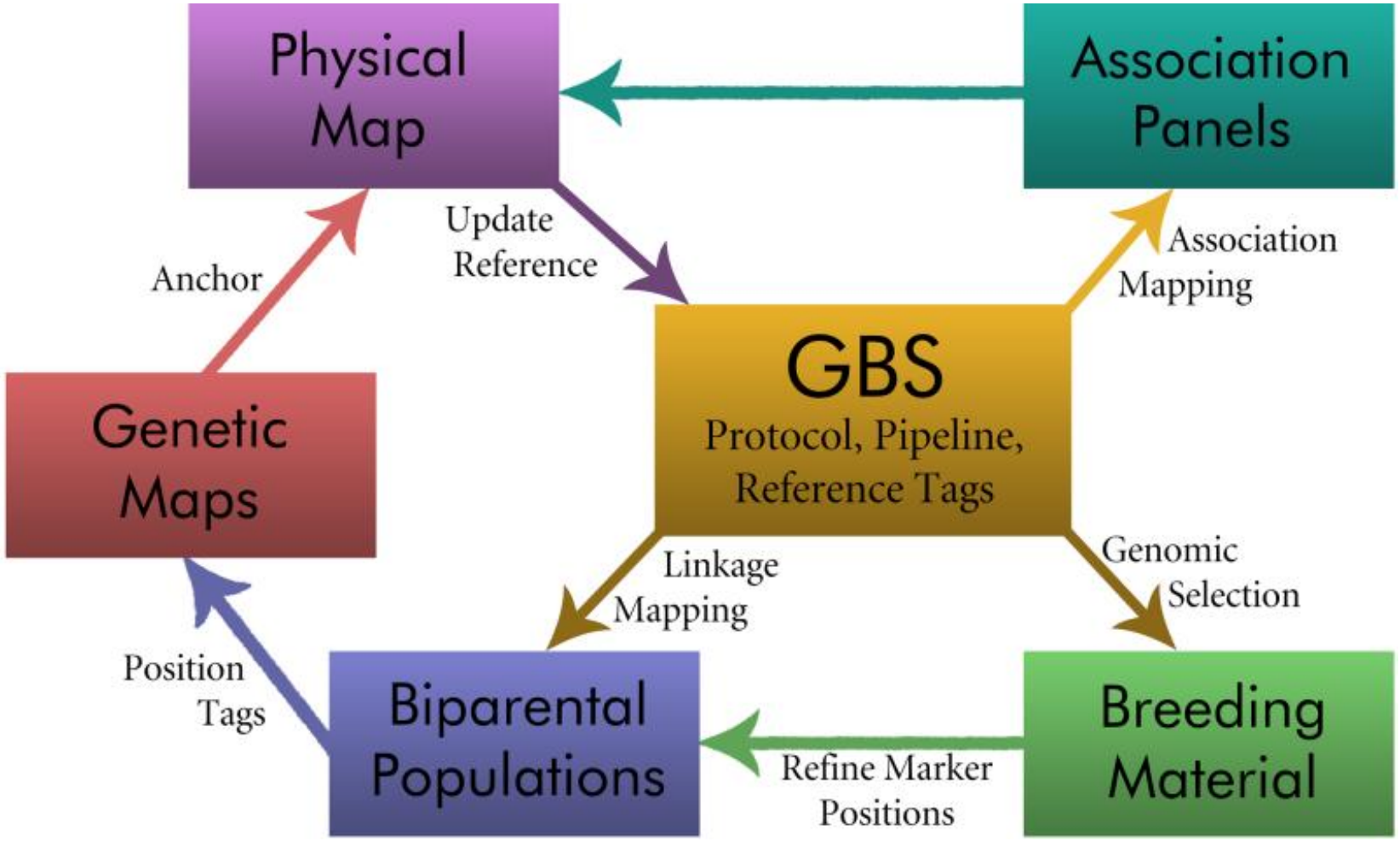
Genotyping by sequencing in Hexaploid Sweetpotato:

Current Status and Future Directions

Bode Olukolu & Craig Yencho



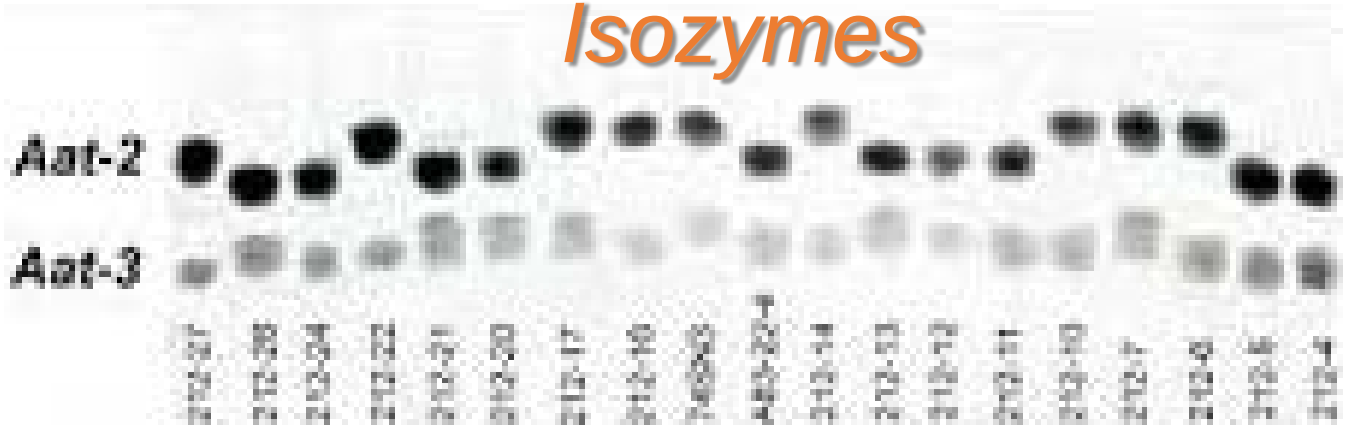

Molecular markers:
central to “genomic resources” and “genetic analyses”



Poland and Rife (2012) The Plant Genome

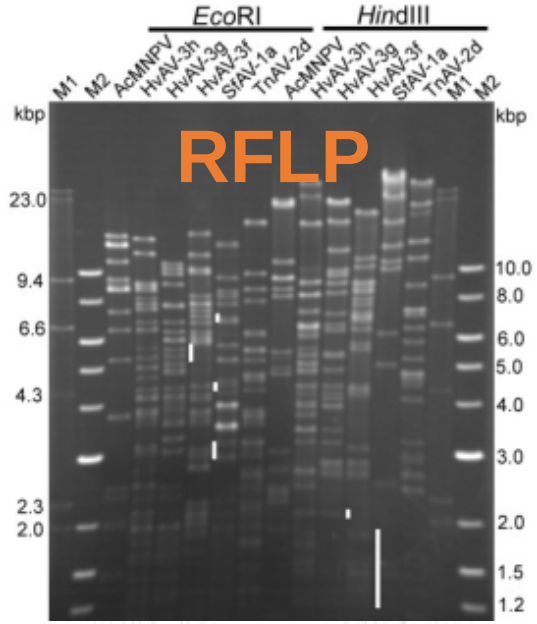
Genetic Markers

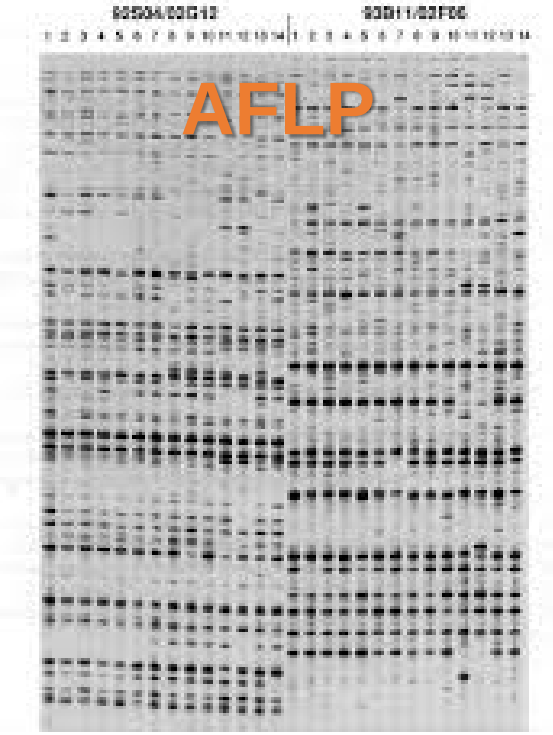
Type	Benefit	Drawback	Example
Morphological markers	- Easy to assay - Low cost	- Highly dependent on environmental factors - Difficult to analyze for quantitative traits - Difficult to determine heterozygosity	Color and shape.
Protein markers	- Low cost - Co-dominant - Less dependent on environmental factors	- Assay samples must be in good condition - Limited availability - Unstable materials (protein)	Isozymes

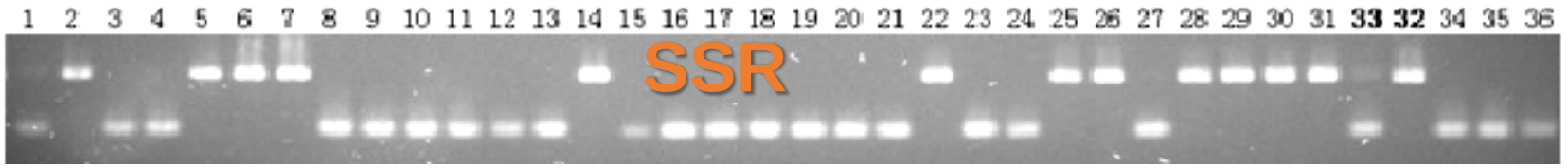


Genetic Markers

Type	Benefit	Drawback	Example
DNA markers (hybridization -based)	- Do not require sequence information of the target - Co-dominant - Unaffected by environmental factors	- Costly and time consuming - Use isotopes - Require large quantities of high molecular weight DNA - Difficult to automate	RFLP
DNA markers (PCR-based)	- Require low quantities of DNA - Quick and easy to assay - High accuracy - Unaffected by environmental factors	- Require expensive equipment - Sometimes requires sequence information	RAPD, AFLP, SSR, and SNP



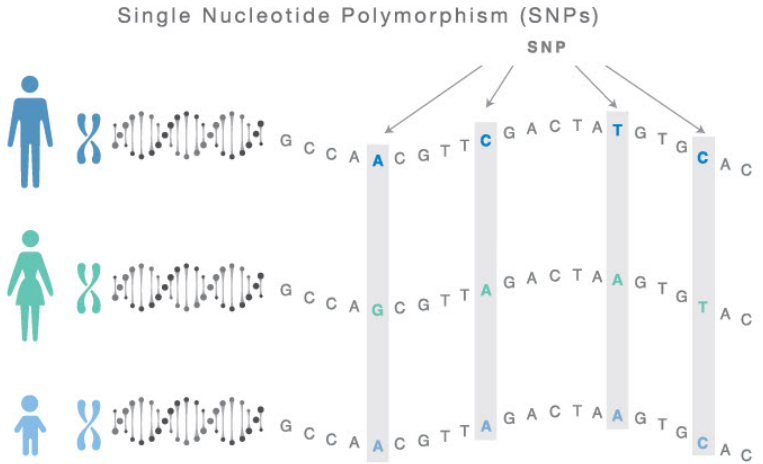




Next-Generation Genotyping: SNP Arrays

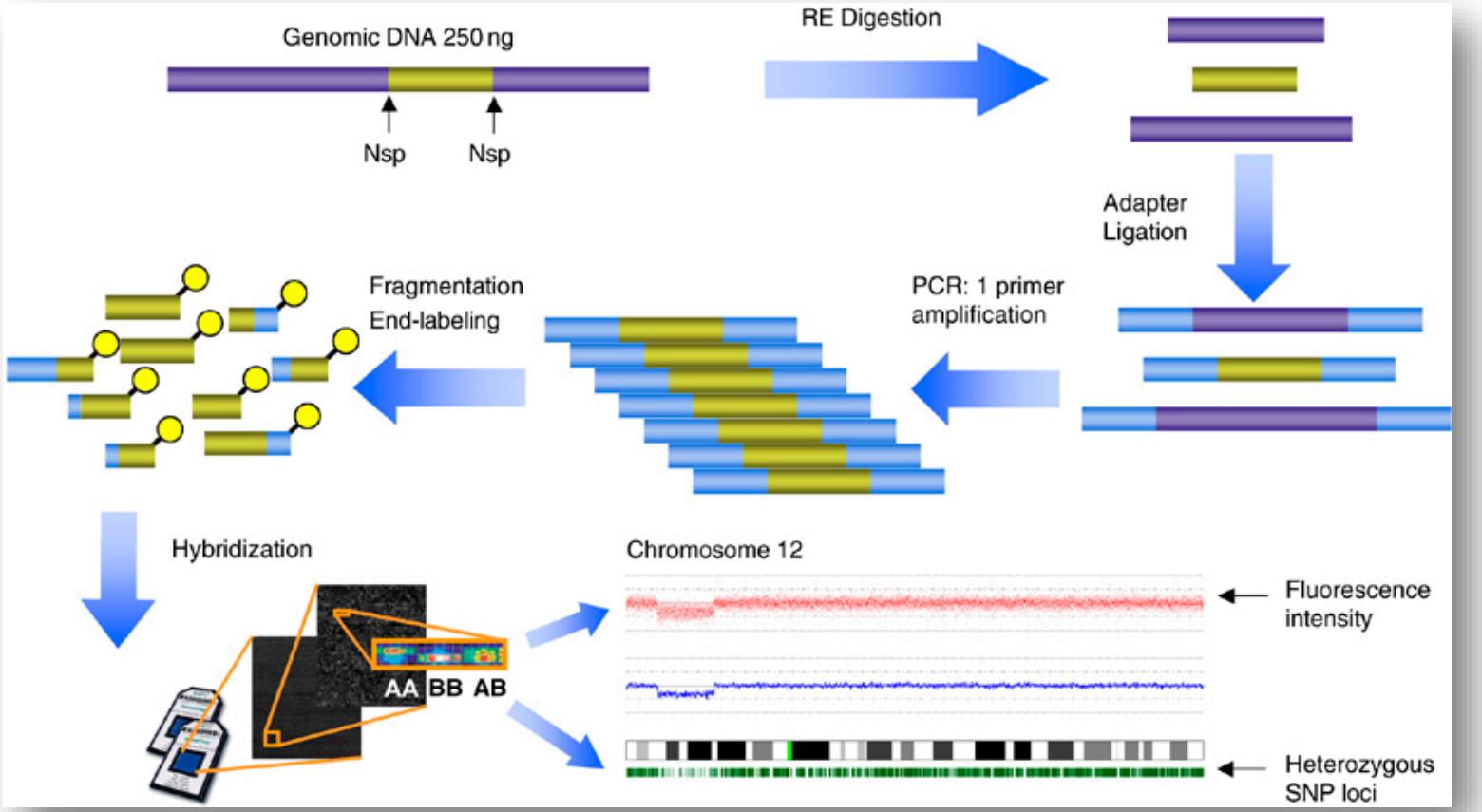
Affymetrix chip

Illumina chip



- Single Nucleotide Polymorphism (SNP)
- SNPs are much more common in genome
- Usually 2 alleles and maximum of 4 alleles

SNP chip/array



o Contains thousands to hundreds of thousands of unique DNA sequences.

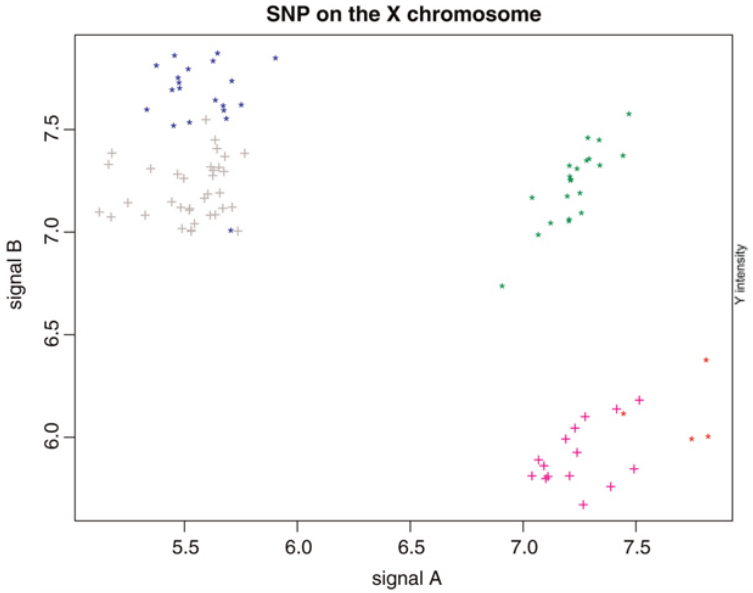
o Single intensity depends on amount of target DNA in sample.

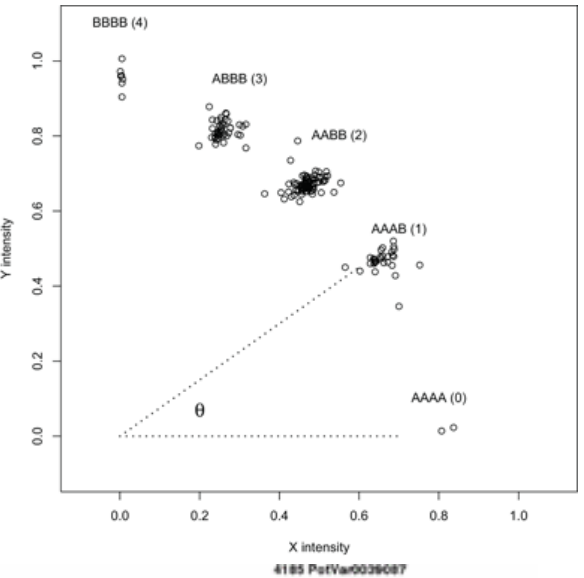
o Manufacturers report genotyping accuracy of 99.5% in diploid genomes.

SNP Calling Algorithms

Affymetrix GeneChip array.

- Normalised & summarized allele intensity



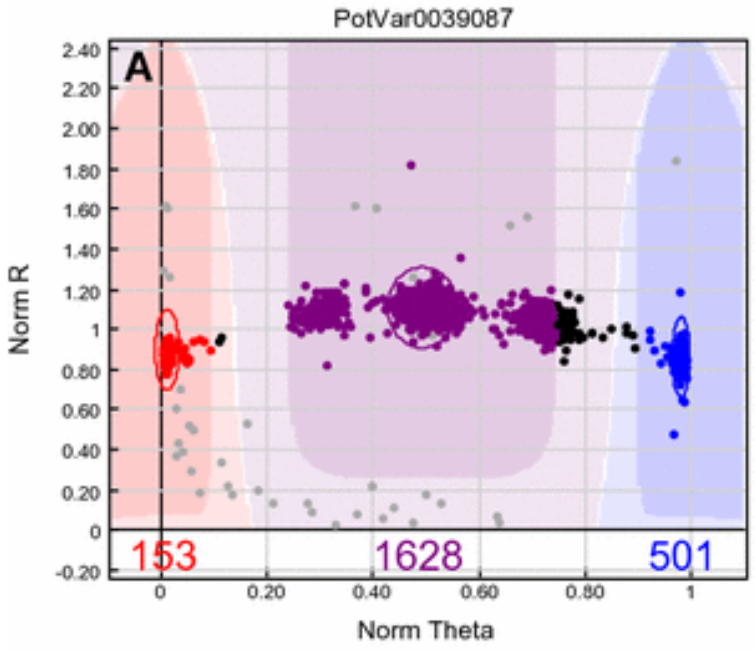


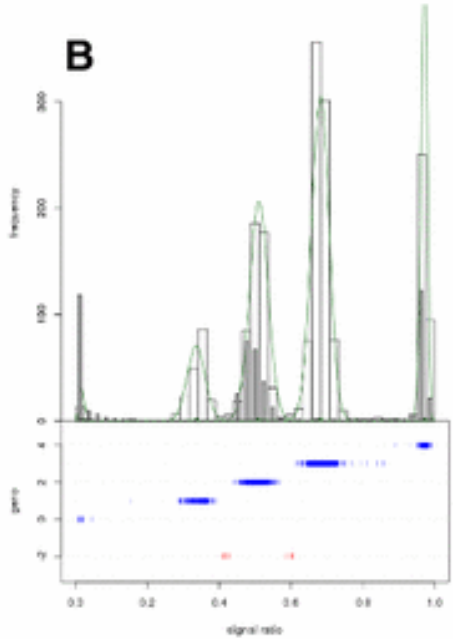
Clustercall

(R-package)

Illumina GenCall

- Bead studio application





FitTetra

(R-package)

SNP Chip/Array: Pros and Cons

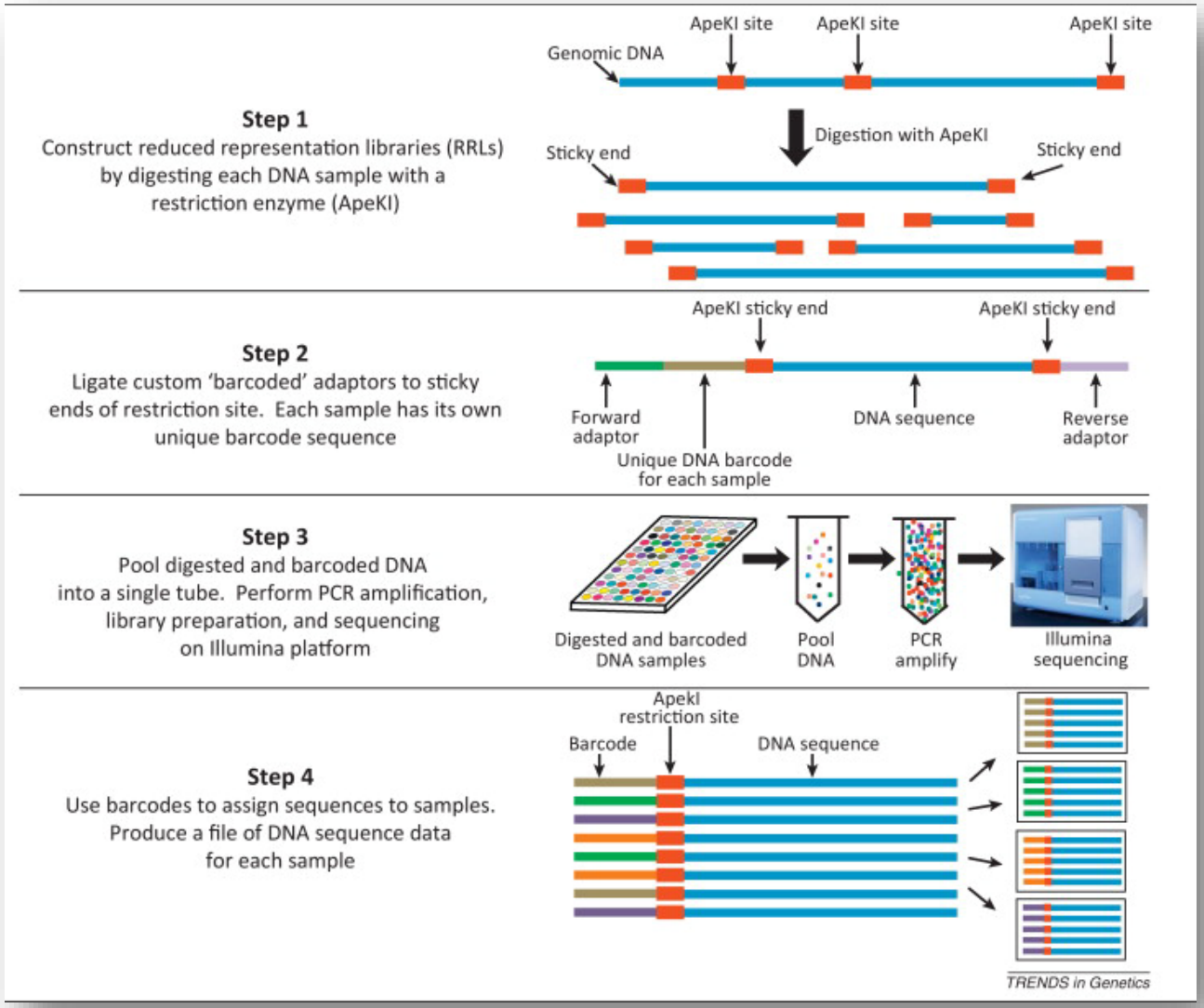
Pros

- ✓Less bioinformatic analysis (user-friendly GBSapp under development)
- ✓Few missing data (GBSpoly optimized to achieve this)
- ✓Inexpensive after chip design (GBSpoly now cheaper)

Cons

- o Initial cost of chip design is expensive (GBSpoly is inexpensive)
- o Ascertainment bias: not all SNP probes are use-able/informative (minimal ascertainment bias with GBS)
- o Might perform poorly for allele dosage calling in polyploids (GBSpoly is optimal for allele dosage calling)

Sequencing-Based Genotyping: **GBS**



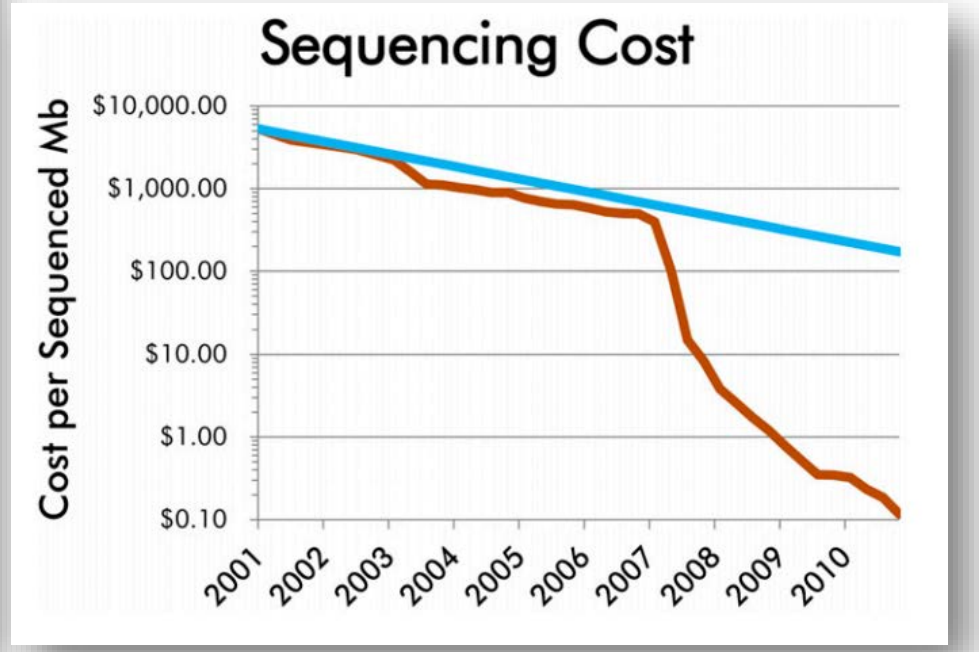
RAD-seq: Restriction-site Associated DNA

ddRAD-seq: double digest RAD-seq

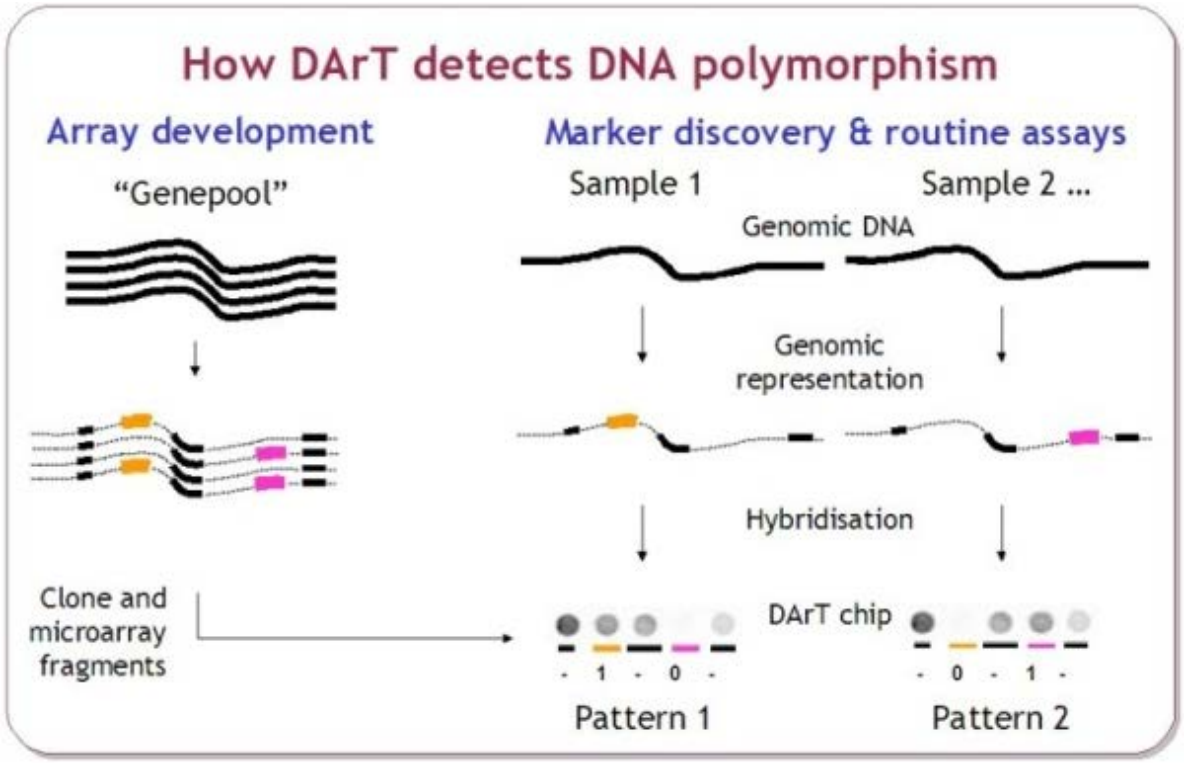
GBS: Genotyping-By-Sequencing

DArTseq: Diversity Arrays Technology Seq

GBSpoly/GBSapp: GBS for all ploidy levels



Sequencing-Based Genotyping: **DArTseq**



- o ***DArT:*** *variant of "SNP array" technologies*
- o ***DArTseq****: Variant of "sequencing-based genotyping/GBS" technologies*

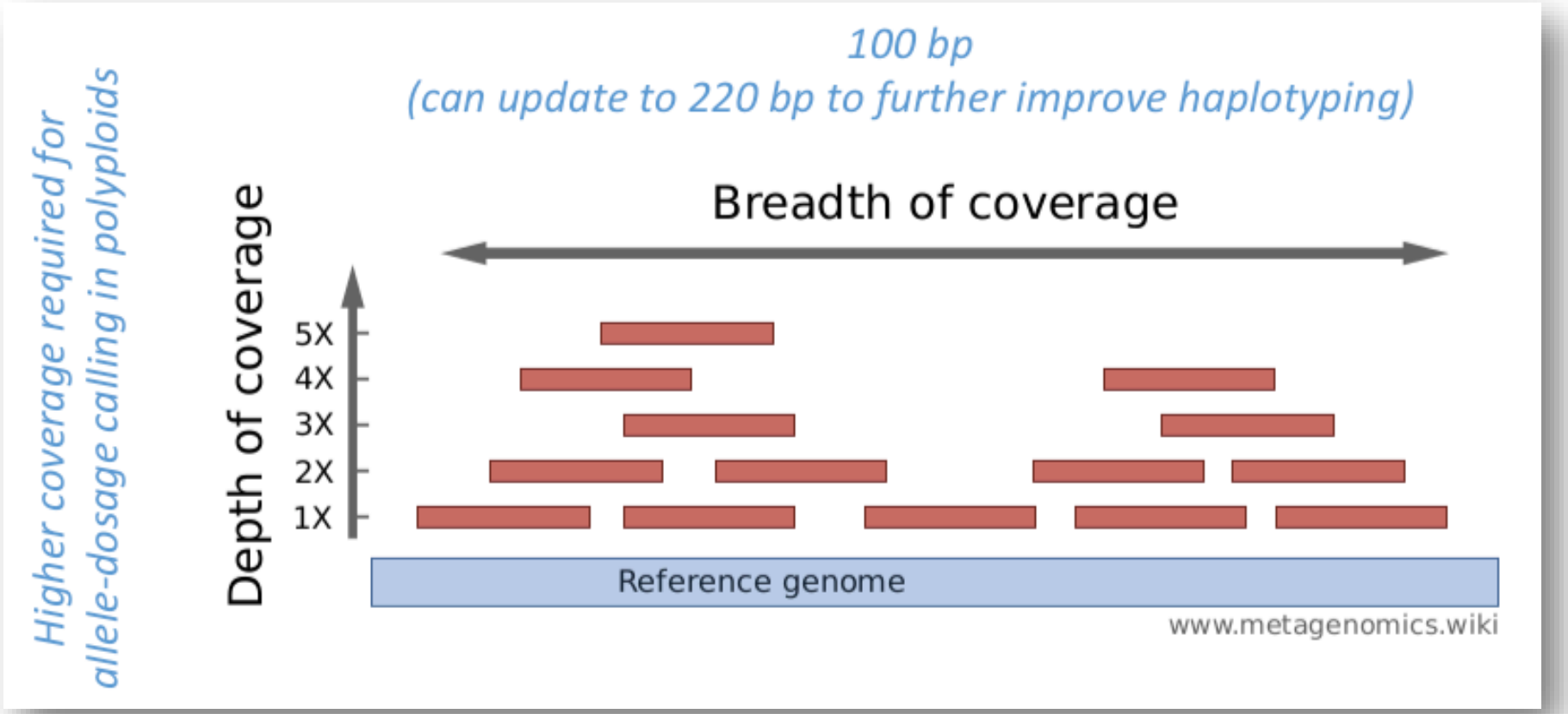
- o Sequence reads (DArTseq) vs. hybridization on chips (DArT)
- o To the best of our knowledge, DArTseq does not capture allele dosage in polyploids.
- o Attempts to use DArTseq for polyploids produces diploidized genotypes:

- DArTseq might be useful for phylogeny, diversity and pedigree studies.
- Absence of "**filtering"** is problematic for all types of analysis.
- Can be problematic for genetic studies such as linkage/QTL analyses, GWAS, Genomic selection.

Sequencing-Based Genotyping: **Why GBSpoly?**

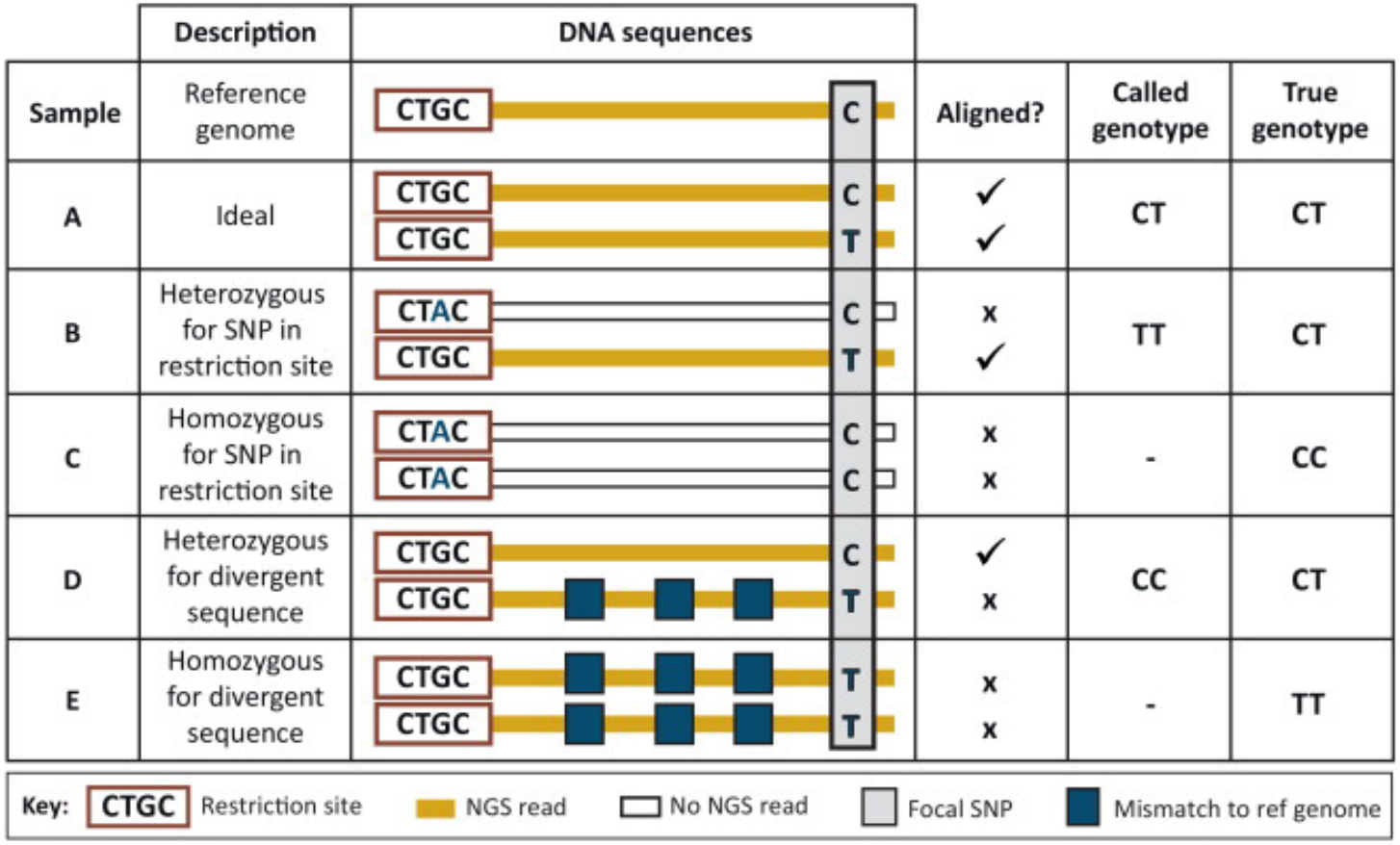
1) Optimal double digest (low chloroplast contamination)
2) Minimal sequencing error and accurate de-multiplexing
3) Eliminates chimeric reads (fragments joined together from different parts of the genome to create non-contiguous sequences)
4) Minimal “missing data” and “no ascertainment bias”
5) Minimal PCR bias results in uniform representation of loci and samples
6) Accurate “genotype” and “allele dosage” calls: includes ability to call sub-genome specific genotypes (2x, 4x and 6x for auto-allo-hexaploidy sweetpotato)
7) Identify and remove bad SNPs (especially derived from paralogs) that creates noise during downstream genetic analysis.
8) Various marker types: SNP, indel, restriction-site polymorphism, and epiSNP*

Sequencing Terminologies



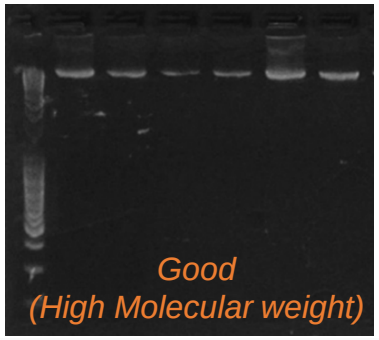

```
ATGCGTAGCGCGGTCAGCGATTGCGCTAGGCCGTAAAAGAT   -Haplotype 1 (CCTGA)
ATGCGTAGAGCGGTTAGCGATGGCGCTAAGCCGTTAAAGAT   -Haplotype 2 (ATGAT)
ATGCGTAGCGCGGTTAGCGATTGCGCTAAGCCGTTAAAGAT   -Haplotype 3 (CTTAT)
       SNP1 SNP2    SNP3    SNP4   SNP5
```

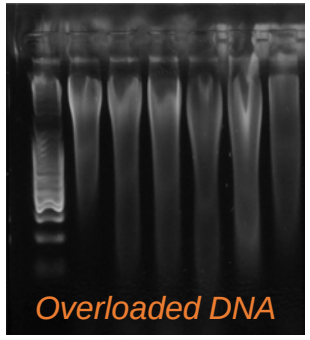
Considerations: Filtering is crucial for all platforms

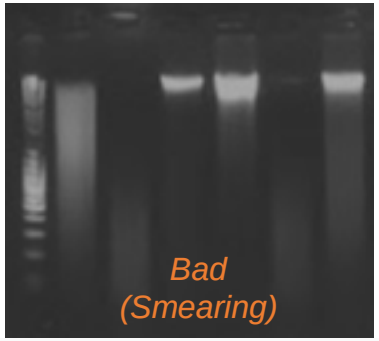


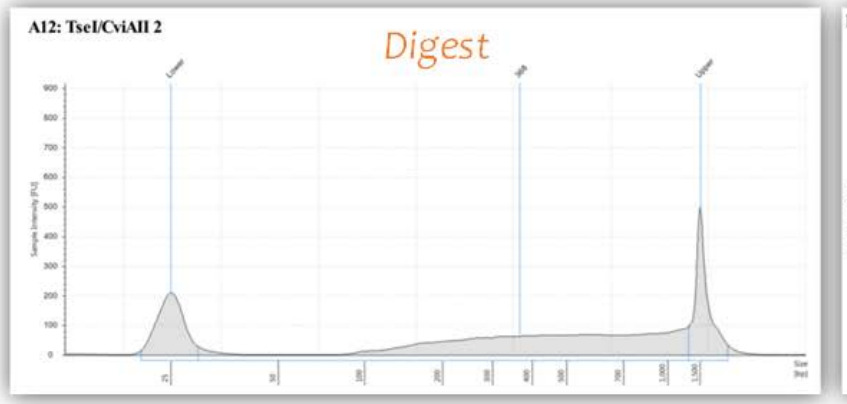
- *Methylated DNA can produce lots of noise (most of SNPs in this regions)*
- *Polymorphism in restriction site is another significant source of error.*
- *~ 96.5 - 99.5 % (depending of read depth filtering) of the SNP data are noise.*

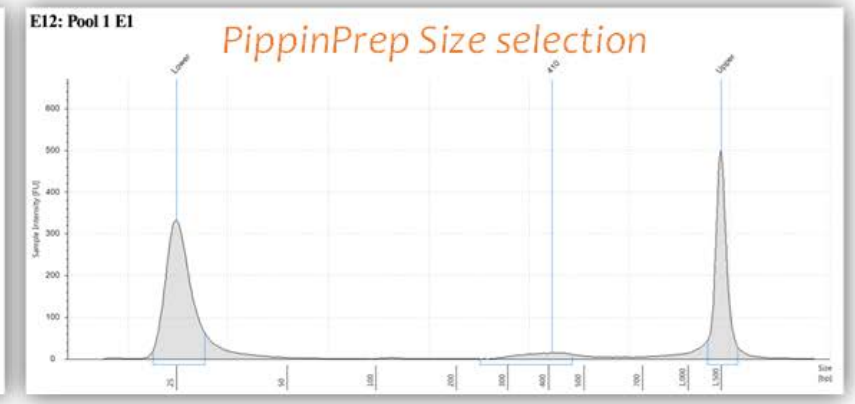
Library Preparation Quality Controls

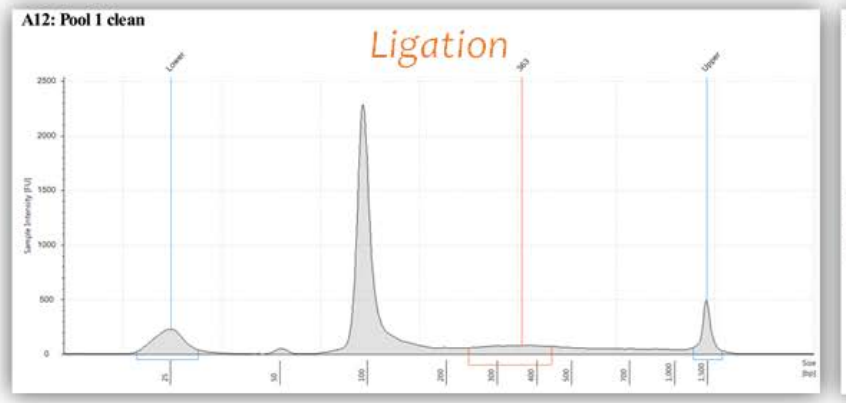


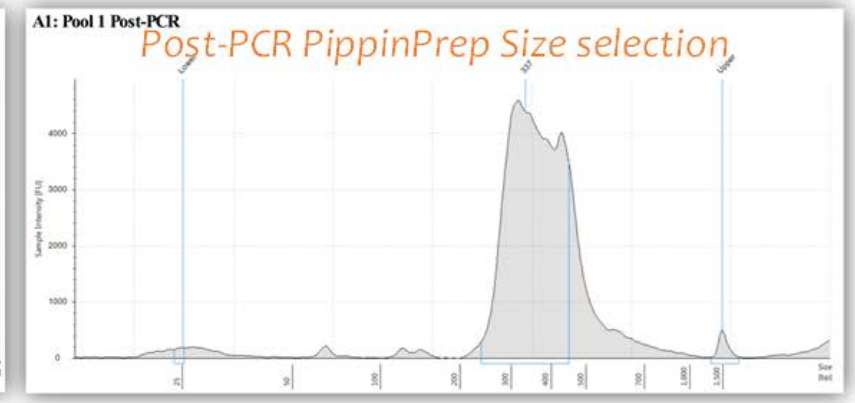












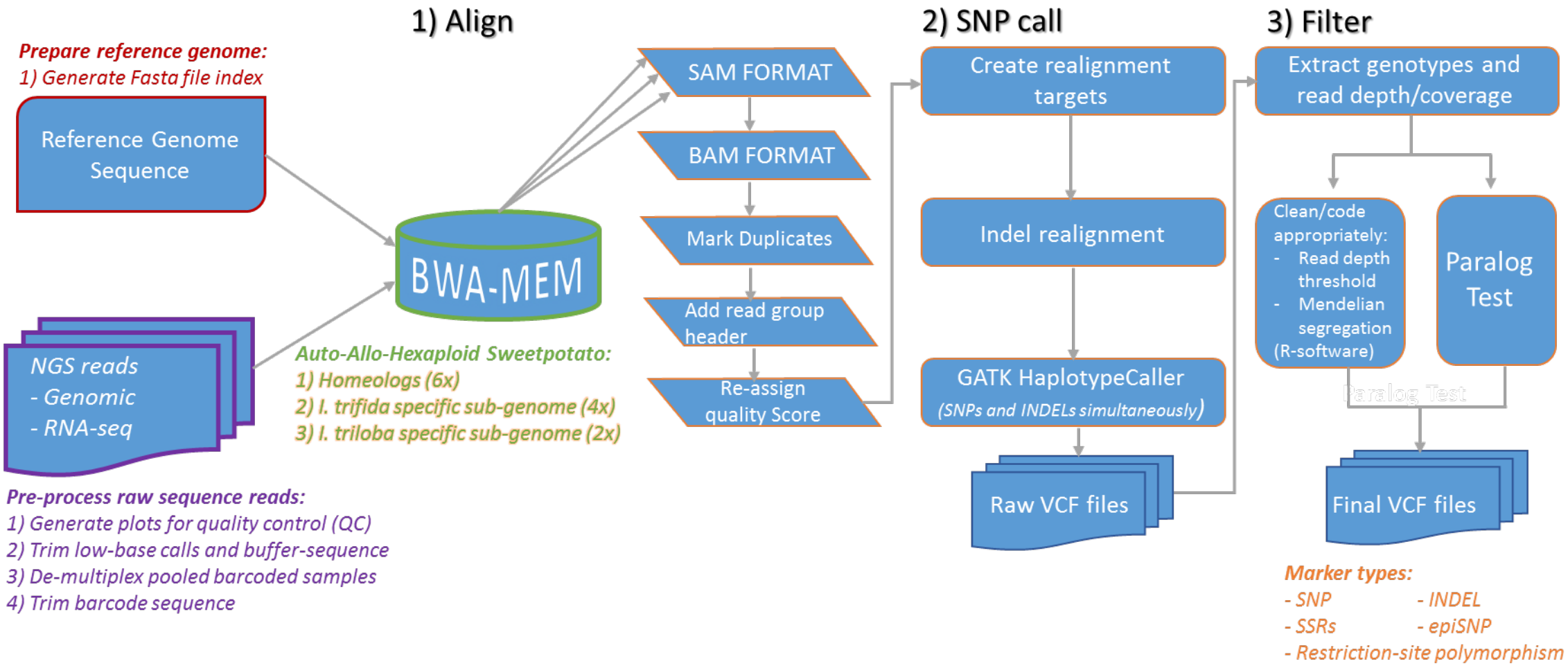
1) Pre-library prep:
- o DNA quality check
- o Enzyme combination
- o Barcode/adapter design

- Bonny Oloka provides case studies

2) Library prep:
- o Double digest
- o Adapter/barcode Ligation
- o Size selection
- o PCR amplification
- o Illumina sequencing

GBSapp: User-friendly software

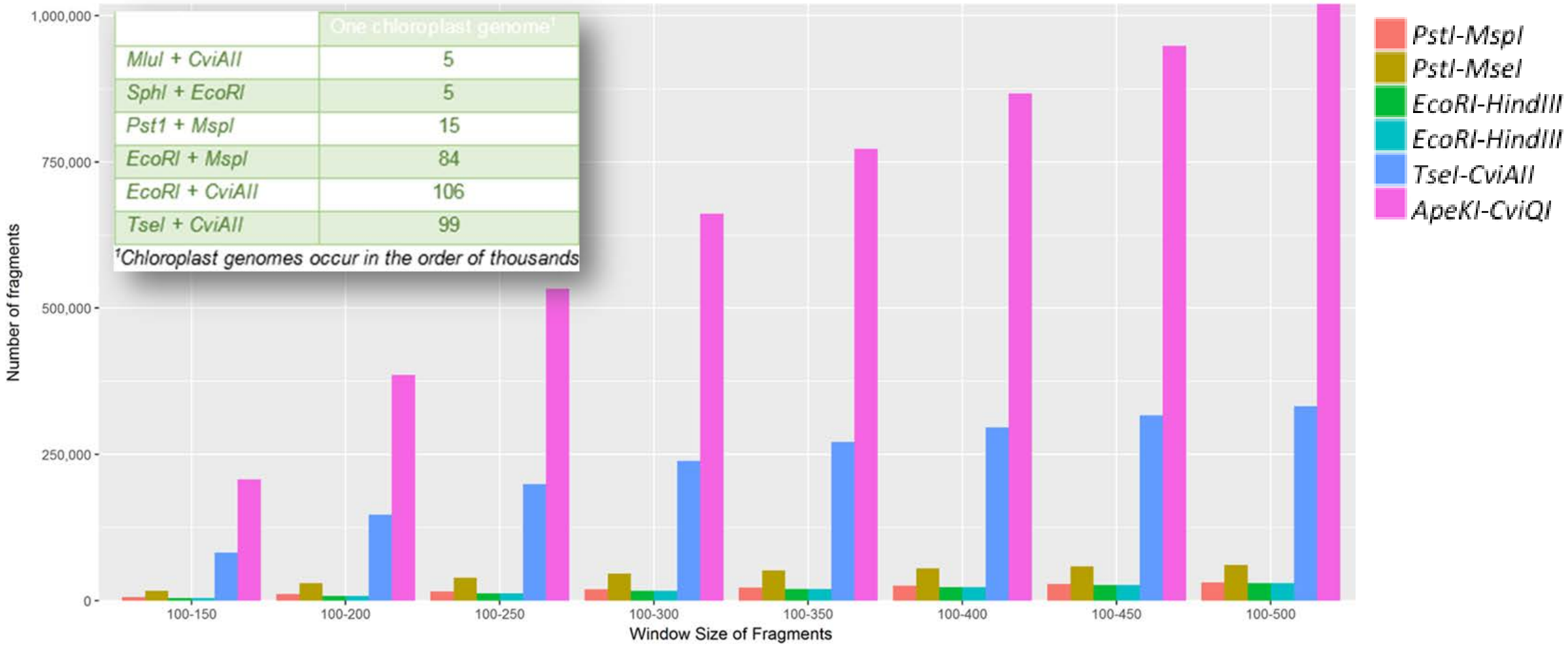
Pipeline Overview



Selection of Restriction enzyme

	One chloroplast genome[1]
MluI + CviAII	5
SphI + EcoRI	5
Pst1 + MspI	15
EcoRI + MspI	84
EcoRI + CviAII	106
TseI + CviAII	99

[1]*Chloroplast genomes occur in the order of thousands*



- ✓ Optimal double digest produces lots of fragments
- ✓ Lower chloroplast contamination

Quality Control

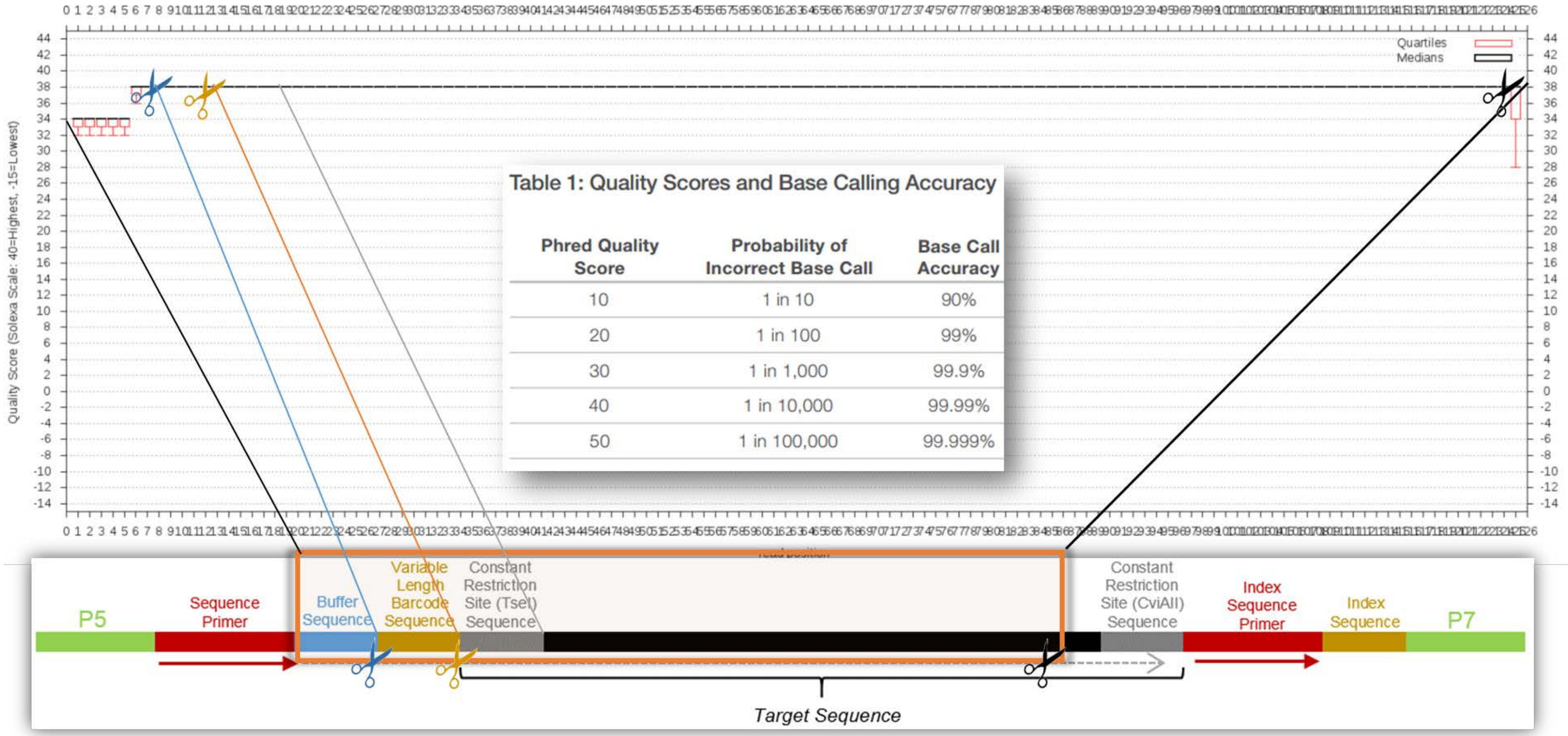


Table 1: Quality Scores and Base Calling Accuracy

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%

- ✓ High quality scores for base calling
- ✓ High quality score of base calling in barcode, hence, accurate de-multiplexing
- ✓ Elimination of chimeric sequences

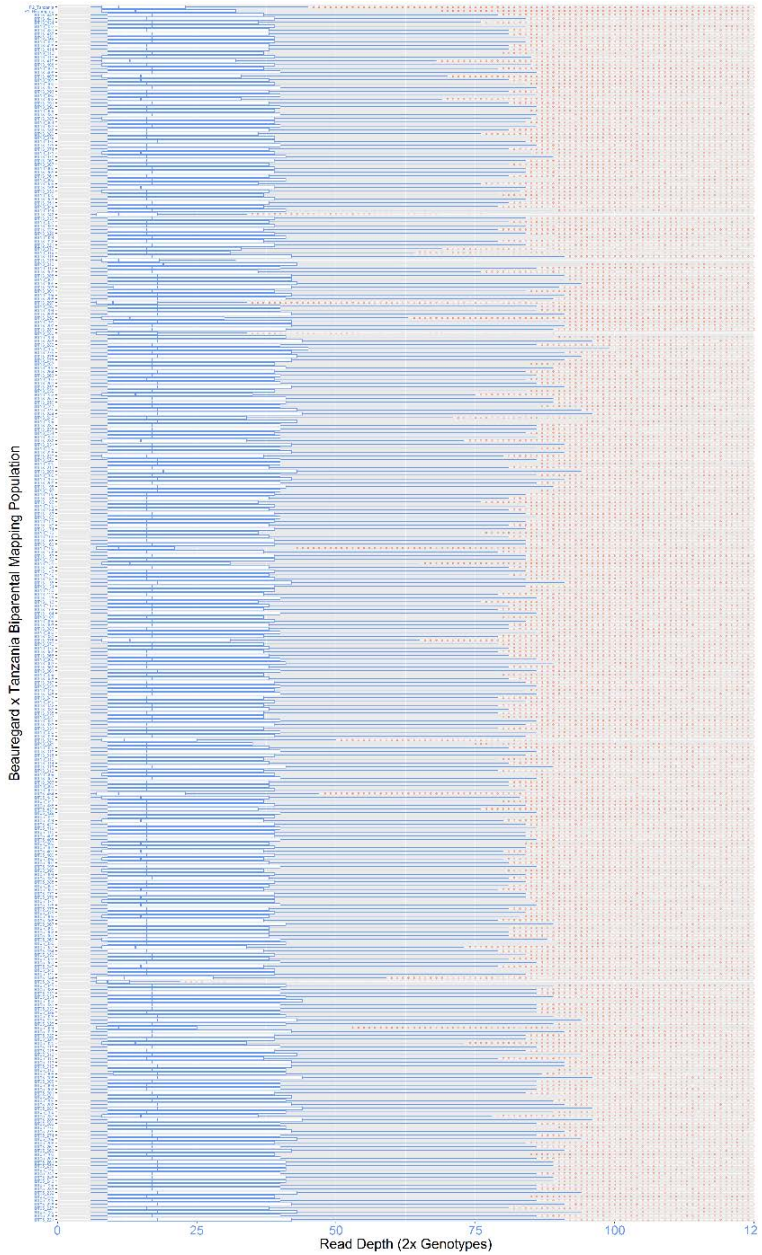
```
@D00555:183:CA8TJANXX:8:2210:1247:2210 1:N:0:29
GTCAGAACACCTCTTACCAGCGAGCTTTGTAGTTCCAAGAAGATTTCAAGACTGAAAAGATGTCTGCTTTCATAATCTTTCTGTGAGTTTGACTGAAAACTAAAATATCATTTGAATAGGATTGC
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG<
@D00555:183:CA8TJANXX:8:2210:1167:2221 1:N:0:29
GTTCCAAGCGATTGCCAGCTAGCGCGCTTGACTAAAGAAAGGGCTTCATTTACTGCTTTGTTTCTTGTTTAGTAAAGTCAAGTTTTTGGCCTTATCTTGCAGGTGACGACGACGTCGAGTTGACG
+
3:<AAFEGGBGFEEFEGGGGGFGGGGGGGGGGGDGDGGDGGGGGGGEGGGGGGGGFGG@FGGGGGGGGFGC1FGCDFFGEGGGGGFGDGGG@08GGGCF00CFGEG0C..CA.CC<C.C8CEDED
```

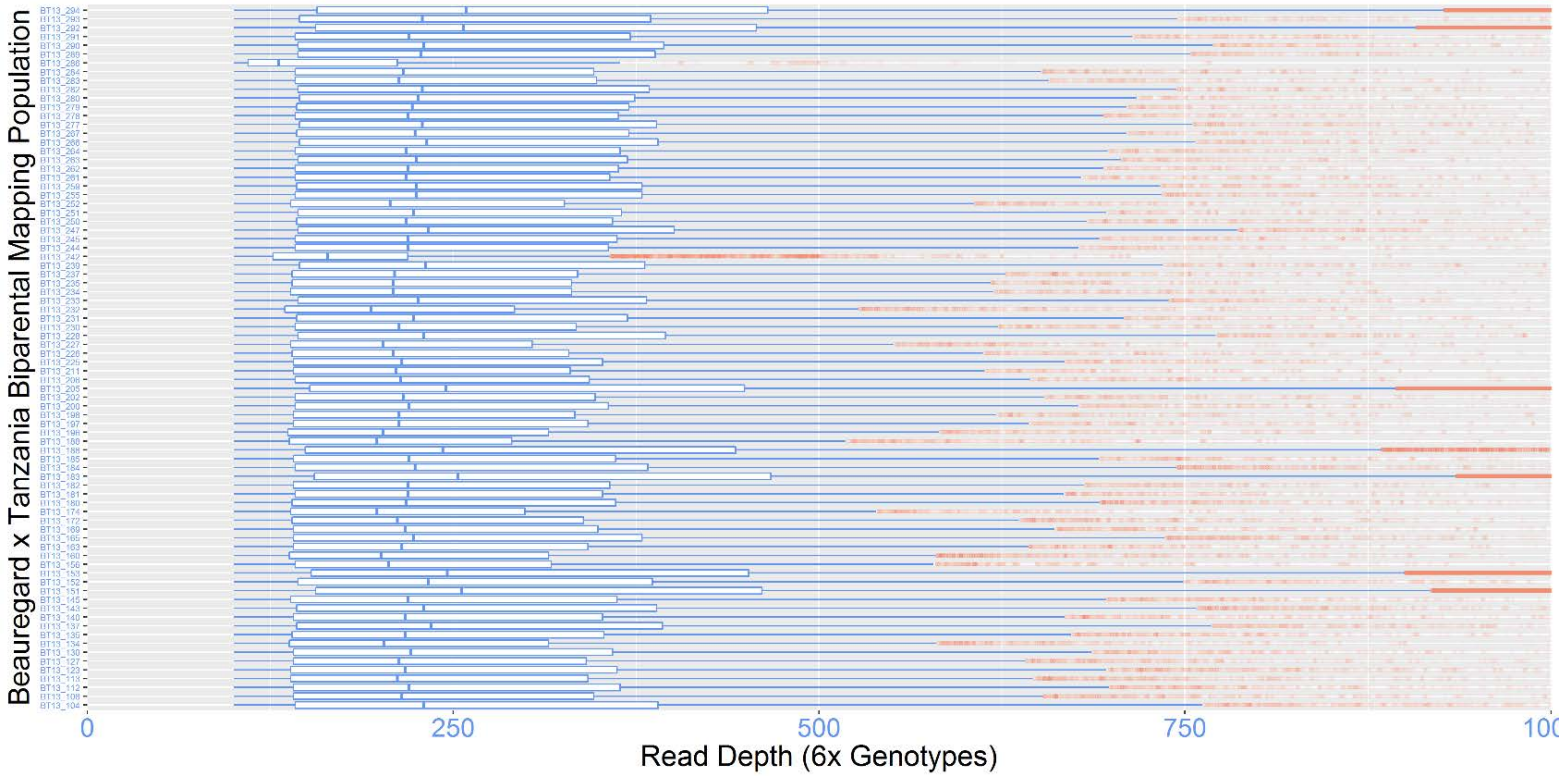
BT05_285

BT05_263

CWGC (W=A or T)
TseI-cut site

Uniform Coverage across clones and loci

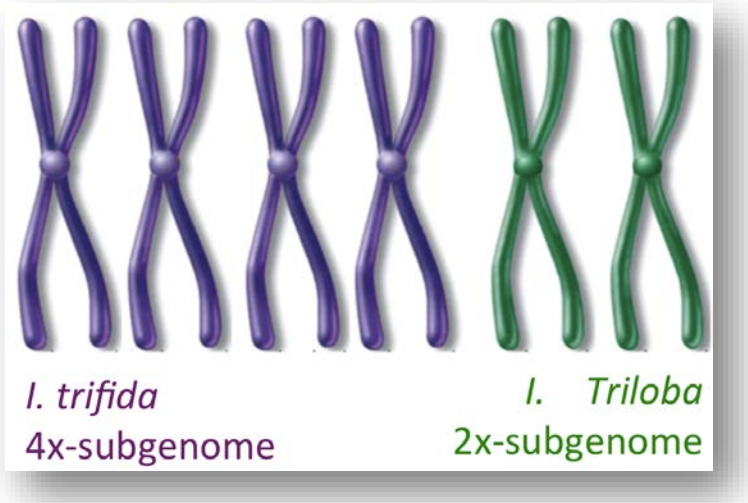




- ✓ Outliers (orange dots) indicate fragments derived from highly repetitive sequences (e.g. transposons)
- ✓ Average read depth significantly higher than threshold required for SNP-calling

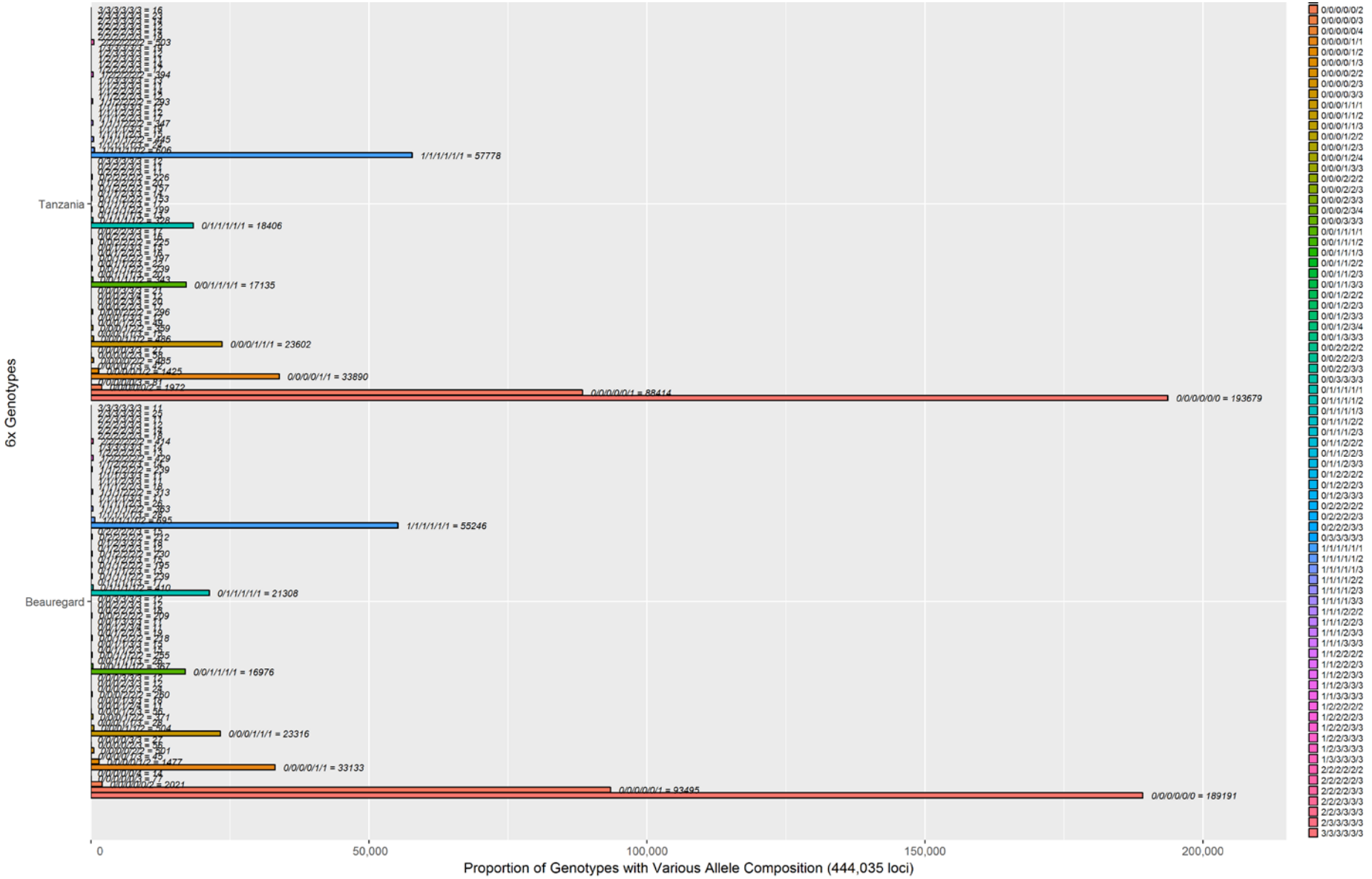
SNP calling pipeline: alignment to reference genomes

	Hexaploid Sweetpotato (%)
De-multiplexing Accuracy (spiked with 5% Phix)	94.12
Reads common to *I. trifida* and *I. triloba* (6x)	89.94
Reads Specific to *I. trifida* (4x)	3.71
Reads Specific to *I. triloba* (2x)	3.78
Total reads aligned to reference genomes	97.43



Illumina HiSeq 2500
Number reads/Lane: ~ 250 millions reads

Frequency of 6x bi-allelic & multi-allelic genotypes



Bi-allelic examples:

Diploid: AB

Tetraploid: ABBB

AABB

Hexaploid: ABBBBB

AABBBB

Multi-allelic examples:

Tetraploid: AABC

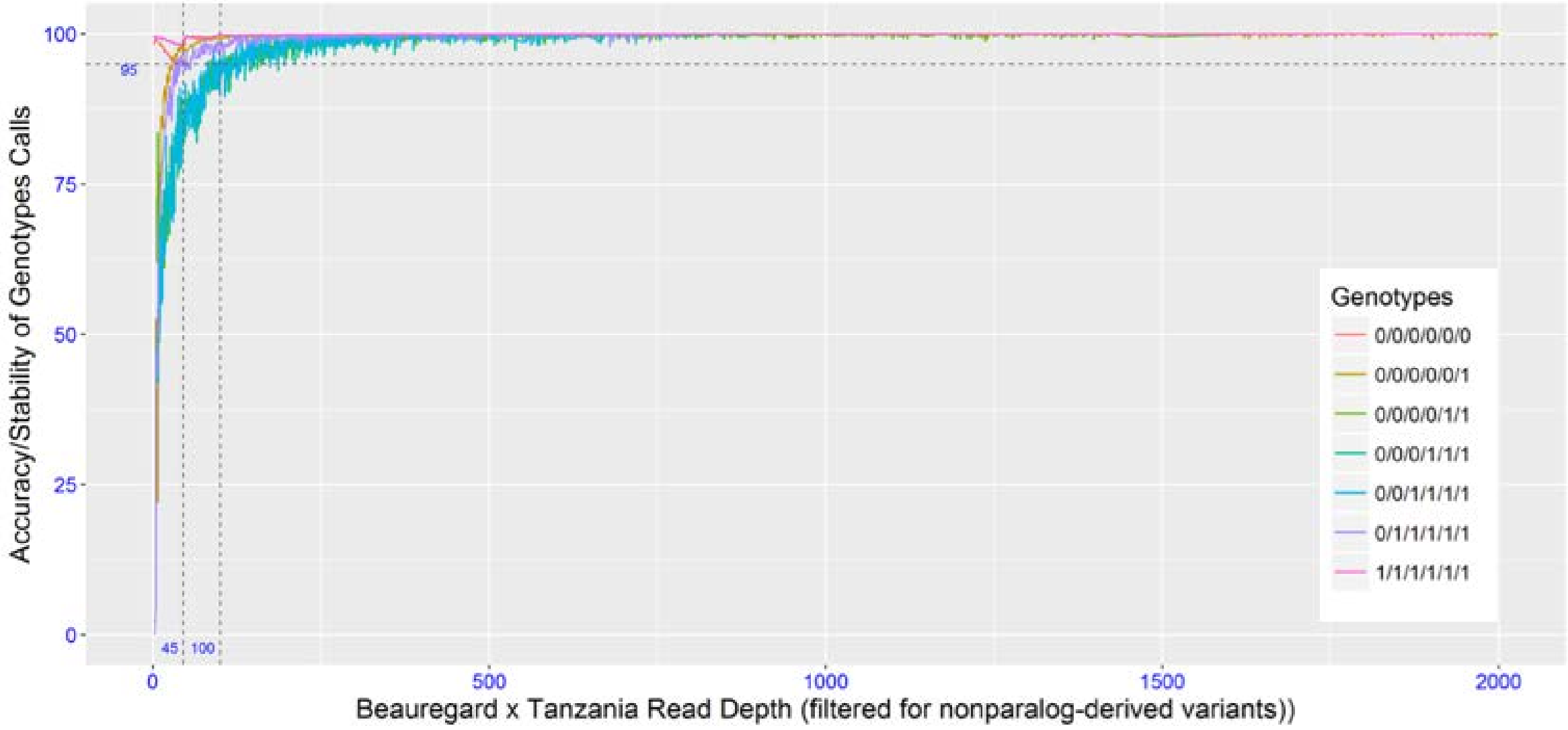
Hexaploid: AABBCC

AABBCD

AABCDE

ABCDEF

6x SNP-calling: After filtering



Frequency of 6x bi-allelic genotype calls

Nulliplex:

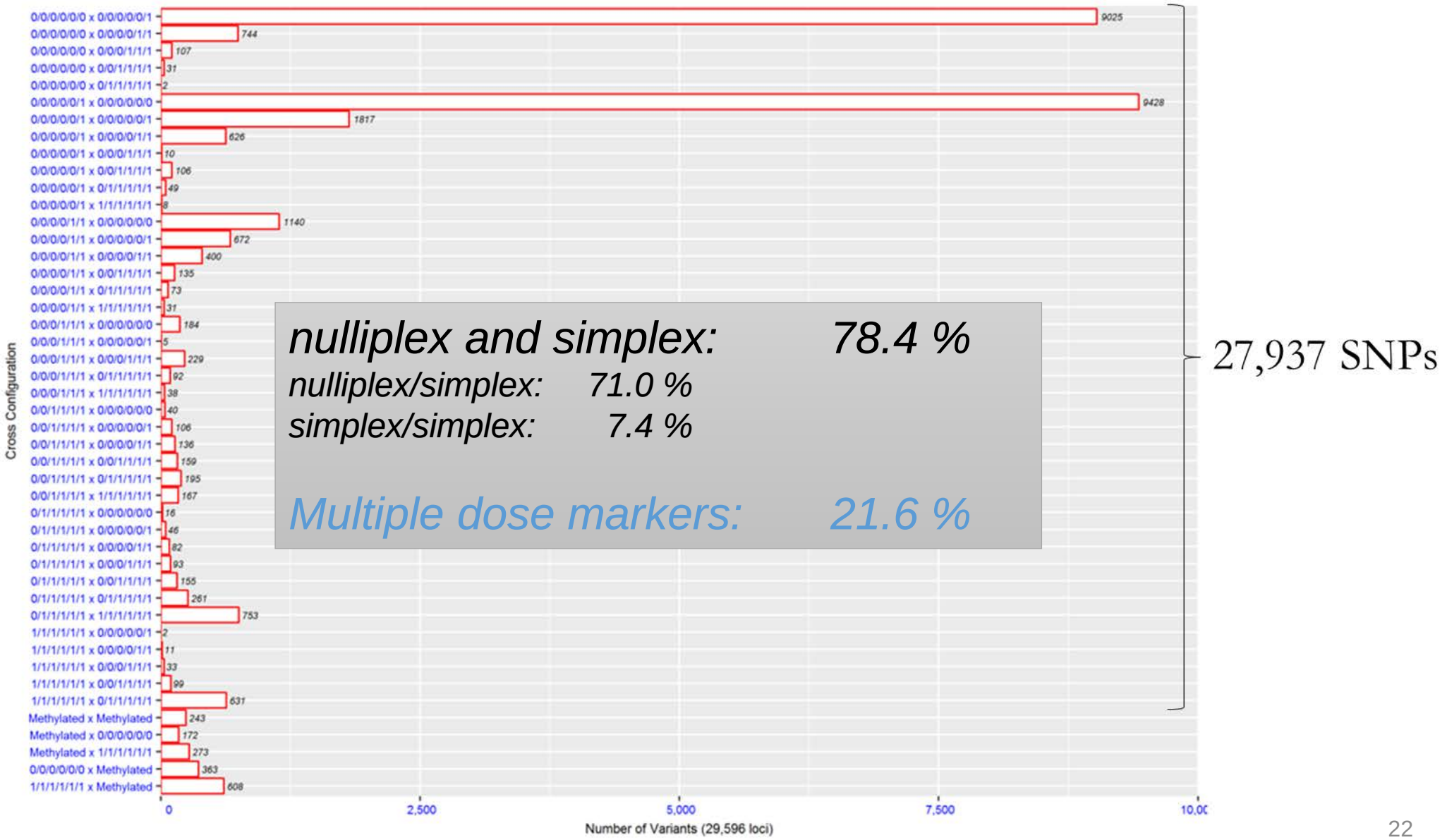
000000
111111

Simplex:

000001
011111

Multiple dose:

000011
000111
001111



SNP Validation

1) Filtered SNPs physically map to a single/unique genomic region.

2) Genotypes and allele dose calls are very stable at different coverage/read depth (above threshold).

3) Independent marker ordering (genetic linkage map) shows high collinearity (and conservation of synteny) with physical reference genome.
(Presentations: Marcelo and Guilherme)

4) SNP data accurately predicts parents in multi-family crossing blocks. Similarly, it predicts pedigree with high accuracy in globally diverse clones.

Validation results support high-fidelity in genotypes and allele dosage.

Deliverables

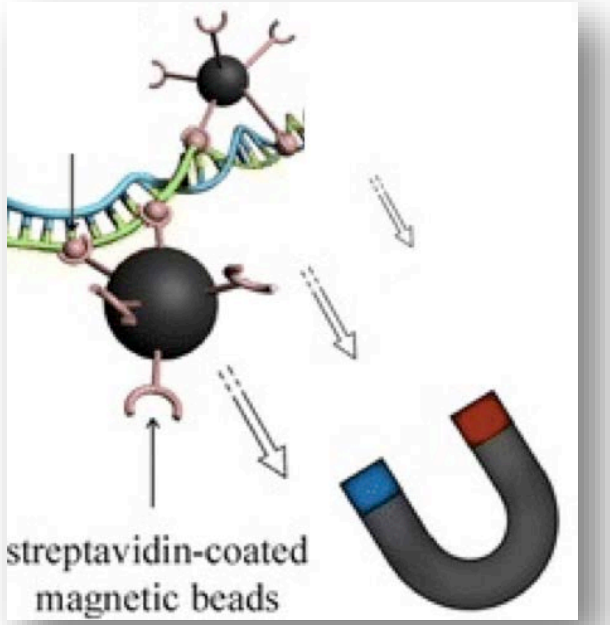
1. Inexpensive/Scalable Genotyping platforms: *GBSpoly*

2. User-friendly GBSapp standalone and cloud-based software (under development): *GBSapp beta-tested by independent groups.*

3. GBSarray: Oligo-array (probes) several orders cheaper than conventional arrays

4. Genotyped mapping population:
   - Beauregard x Tanzania (BxT) - TxB
   - New Kawogo x Beauregard (NKB) - MDP 8x8 parents
   - M9xM19 diploid population
   - *non-GT4SP:*

**USDA SP germplasm* **DC SP population* **Tomato population*
**Strawberry population* **Blackberry population*



Genotyping costs

Platform	Plex-level	Cost/sample	# of SNPS (Average)
GBSpoly *Essential for discovery phase*	96	$20.04 ($17.54)	5,000[a]
		$33.58 ($31.08)	10,000[b]
		$87.75 ($85.25)	30,000[c]
GBSarray *Leverages strengths of GBSpoly and SNP chip/array technology*	96	$20.04 ($17.54)	30,000[c,d]
	192	$13.27 ($10.77)	
	384	$9.89 ($7.39)	
	768	$8.19 ($5.69)	
	1,536	$7.35 ($4.85)	
	2,304	$7.06 ($4.56)	

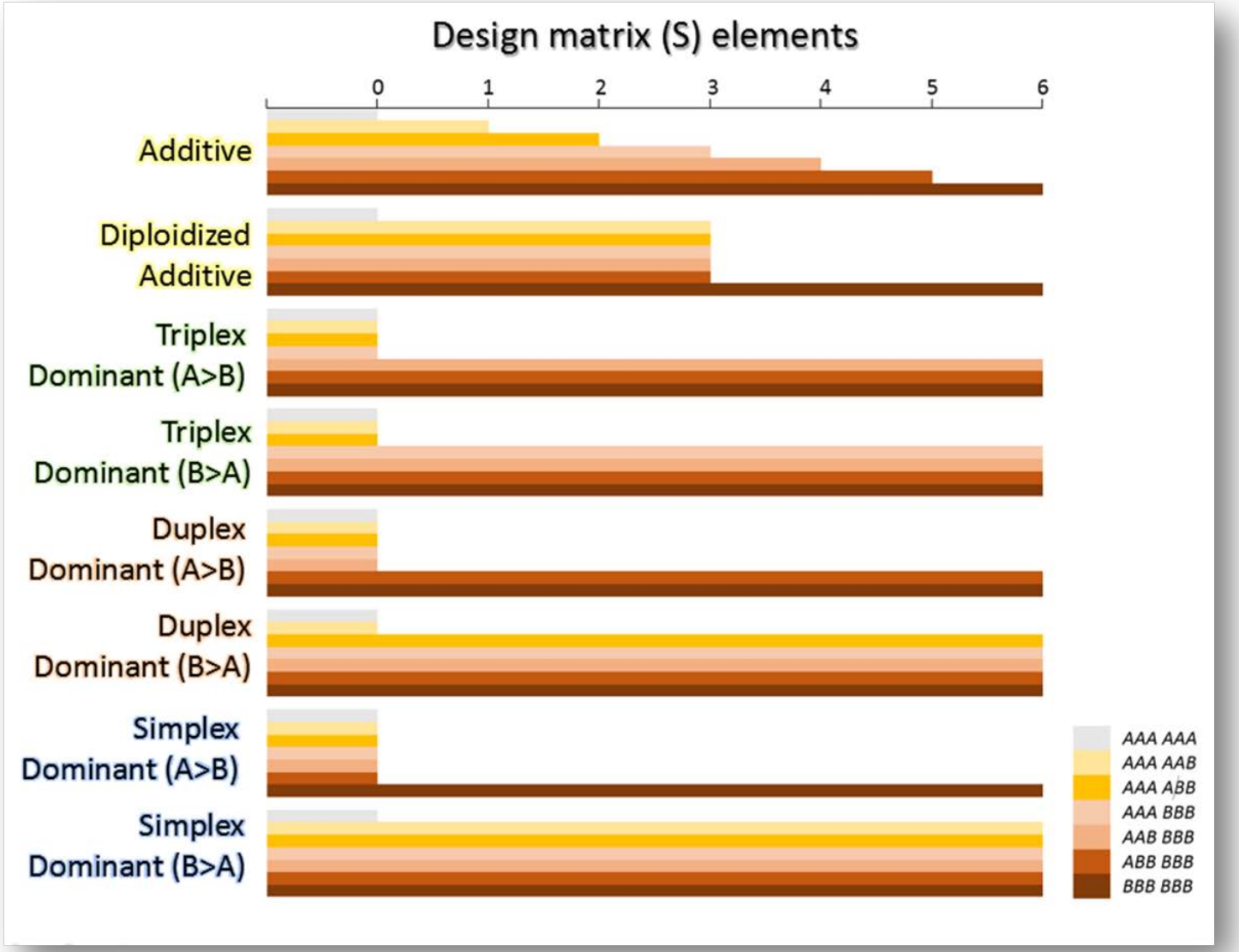
Library prep cost per sample : $6.50
Updated library cost per sample: $4.00
HiSeq2500 S4 (250 million per lane) : $1,300
*Update to NovaSeq S4: **4.27 times more** yield at comparable cost.*

[a]100 bp window
[b]200 bp window
[c]300 bp window
[C]0.5 - 3.5 % of GBSpoly raw SNPs

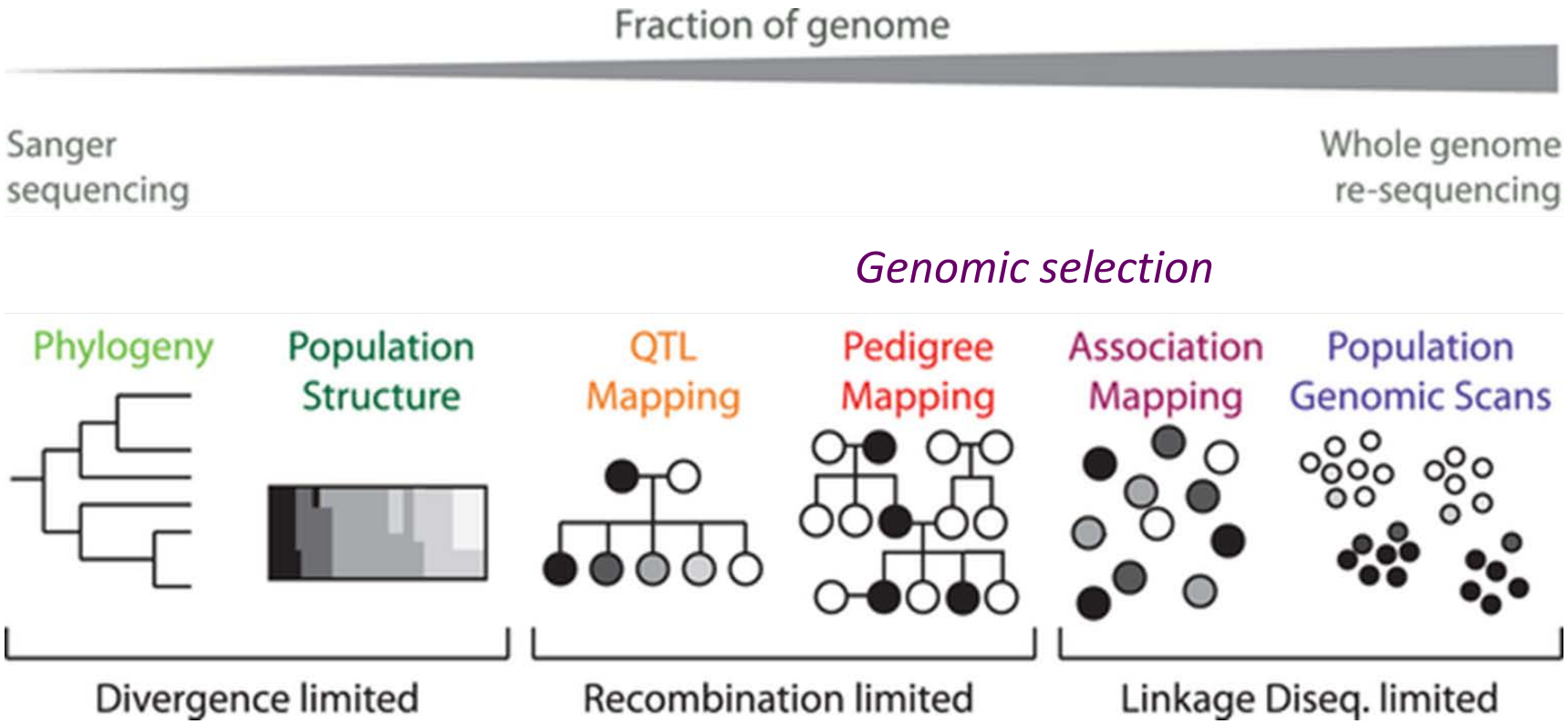
Information on Sweetpotato Populations

1. M9xM19 diploid population: 210 F1 progenies; 9,500 SNPs *(100 bp window)*
2. Beauregard x Tanzania (BxT): 315 F1 progenies; 27,937 SNPs *(300 bp window)*
3. Global diversity panel: 417 clones; 32,784 SNPs *(100 bp window)*
4. TxB: 245 F1 progenies *(300 bp window)*
5. New Kawogo x Beauregard (NKB): 287 F1 progenies *(300 bp window)*
6. DM04-001 x Covington (DC): 450 F1 progenies *(300 bp window)*
7. MDP 8x8 parents: 16 parents; 2,000 F1 progenies
8. Global diversity panel: 700-800 clones; expecting >100,000 SNPs (300 bp window)

Dosage: utility for marker/genomics-assisted breeding



Dosage: utility for marker/genomics-assisted breeding



Conclusion

1. Advancement in genotyping technology from low-throughput to high-throughput platforms. *Low-throughput assays are still important (open for discussion)*
2. GBSpoly leverages the strengths of other technologies, new innovative ideas, and multiple QC steps to resolve typical problems (i.e. high error rates, biased libraries, low efficiency, and high cost).
3. GBSpoly is cheaper and delivers allele-dosage at all ploidy-levels. Accuracy is high and confirmed by several empirical validations.
4. Using GBSpoly as a SNP discovery phase, we are developing a GBSarray platform, which will further drive cost down.
5. GBSpoly is now routinely used in both diploid and polyploid crops.

Acknowledgements

- Craig Yencho (NCSU): Sharon Williamson
- Zhao-Zang Zeng (NCSU): Guilherme Da Silva Pereira, Marcelo Mollinari
- David Baltzegar (Genomics Sequencing Lab, NCSU): Hannah Huntley, Erin Young
- Zhangjun Fei (BTI, Cornell): Honghe Sun, Chen Jiao
- Robin Buell (Michigan State University)
- Lachlan Coin (University of Queensland, Australia)