



Discovery & Development of Genetic Markers for use in High Throughput Labs



**Bayer Russia Molecular
Marker Training**

25 July 2023



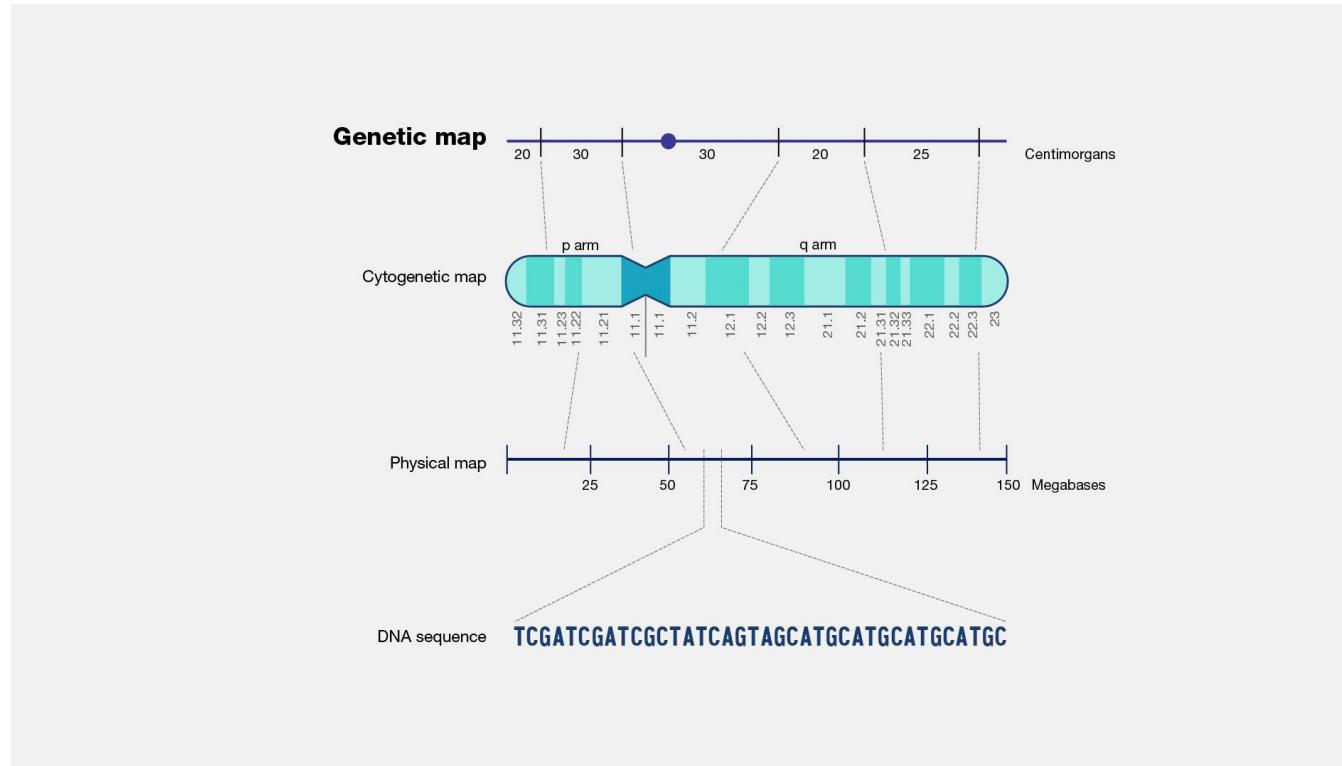
Bayer Russia Molecular Marker Training : Genetic Markers and Technologies



- 1 Genetic Markers for Discovery**
- 2 Genetic Mapping**
- 3 Trait-linked Marker Development**
- 4 Markers for Genome Wide Selection**
- 5 Bayer Genotyping Technologies**

What is genetic marker?

A genetic marker is a DNA sequence with a known physical location on a chromosome.



<https://www.genome.gov/genetics-glossary/Genetic-Map>



Genetic markers available

Genetic marker can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed.

What is an ideal genetic marker?

- co-dominant
- evenly distributed throughout genome
- highly reproducible
- easy to detect higher level of polymorphism

Most common genetic markers

- single nucleotide polymorphisms (SNPs)
- simple sequence length polymorphisms (SSLPs)
- insertions/deletions

SNP

TCGATCCAAATAATTAAATTA
TCGATCCAAA**CA**ATTAAATTA

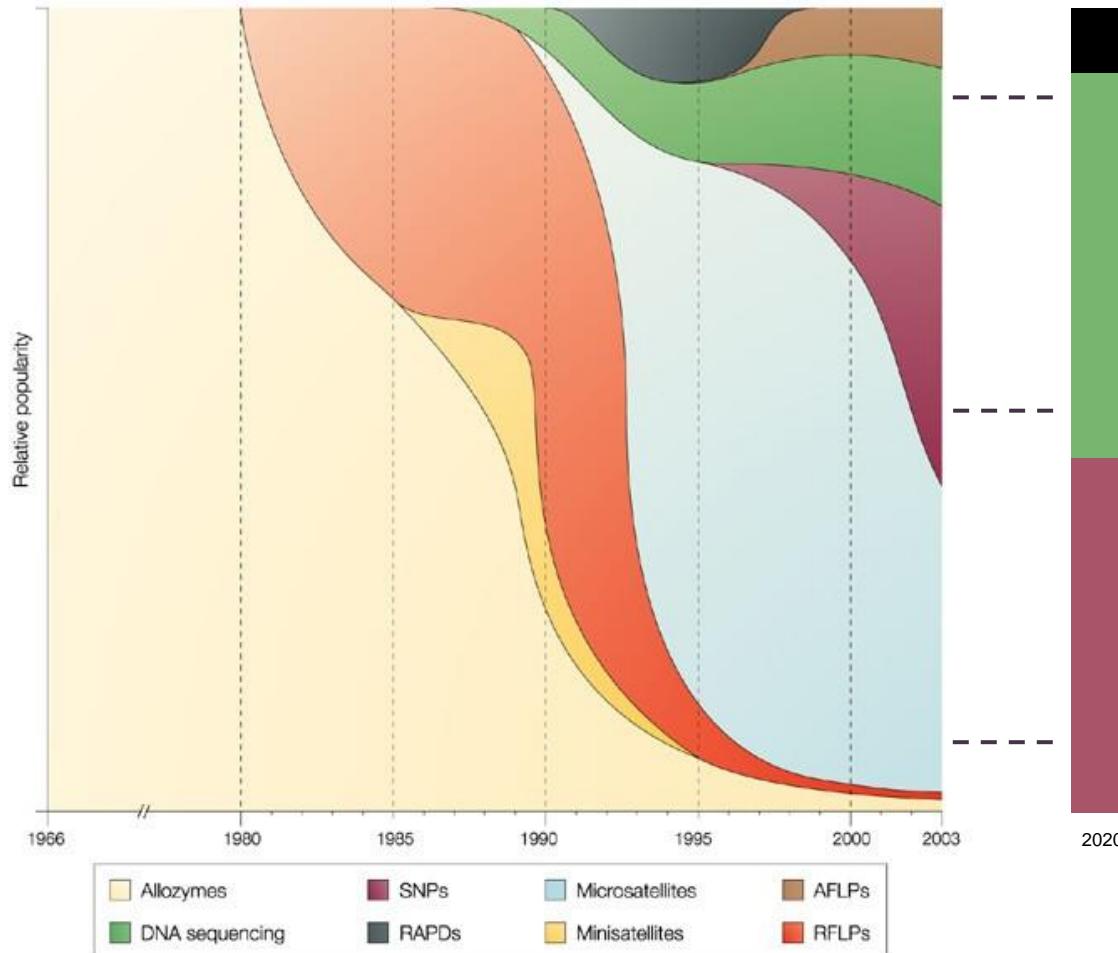
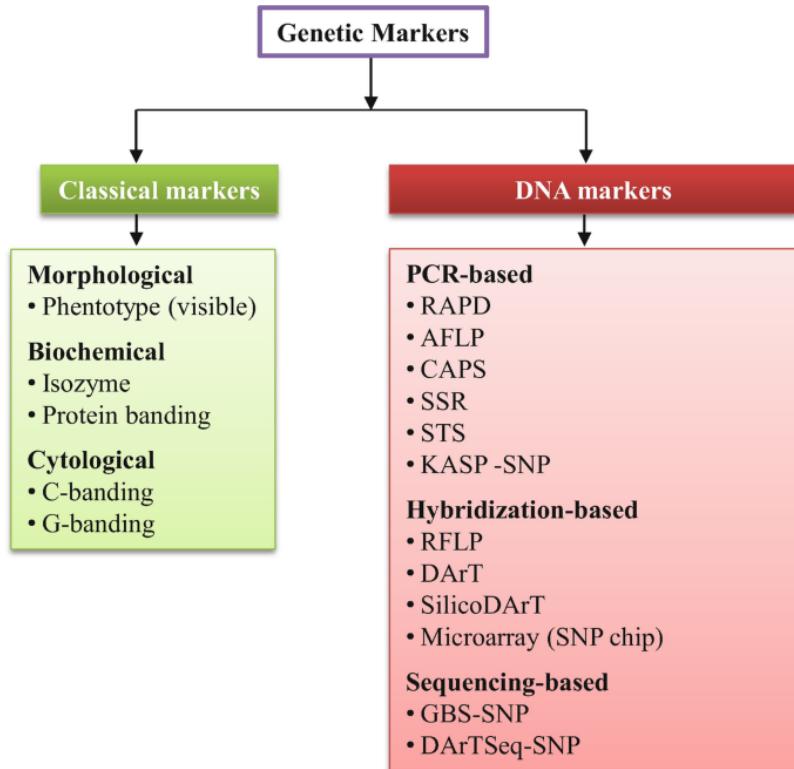
Insertions/deletions

TAAGATCCTTCAGGAAAA**TTCCAA**AAGGGATGCAACTG
TAAGATCCTTCAGGAAAA-----AAGGGATGCAACTG

SSLP

TACAATAATTCT**TATATATA**AGTCACGAACAGAAAATG
TACAATAATTCT**TATATATATATATATATA**AGTCACGAACAGAAAATG

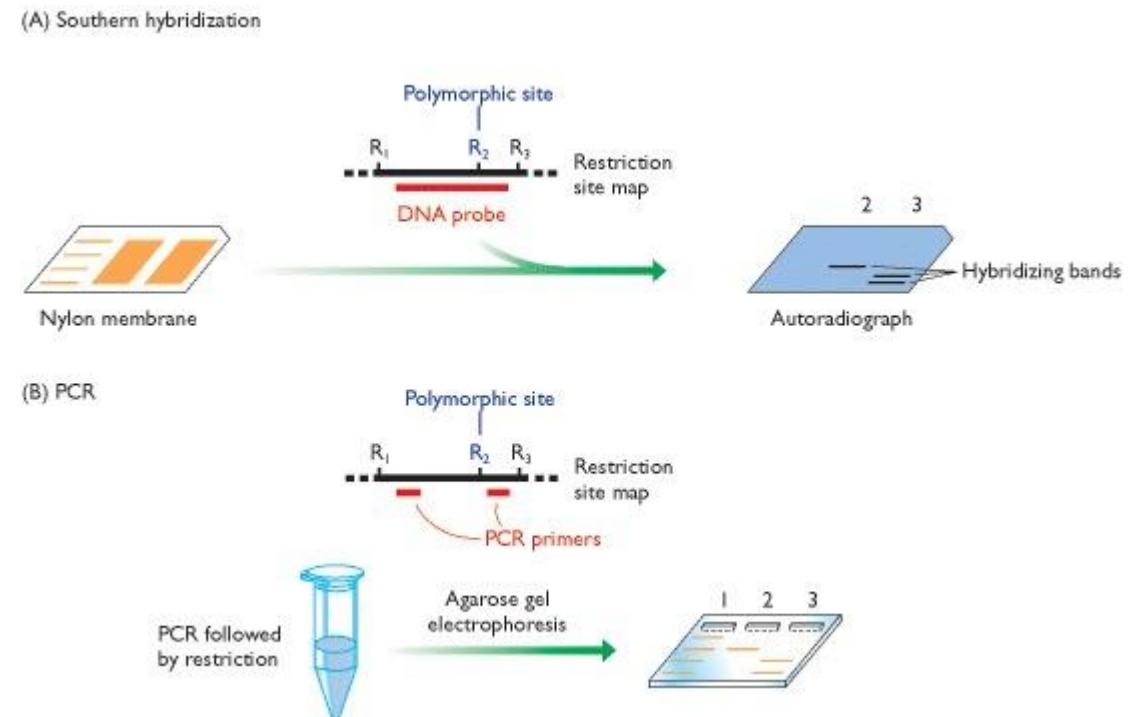
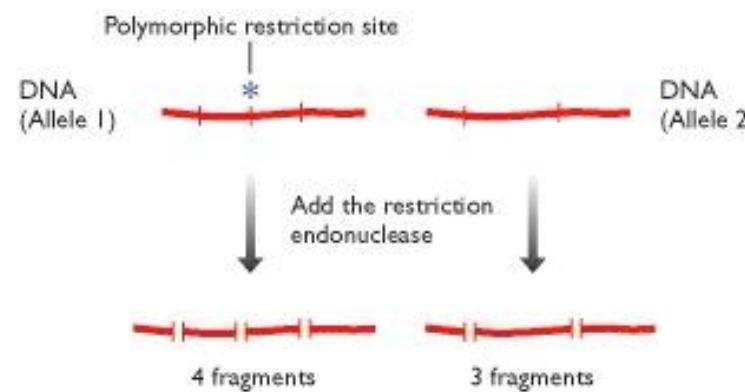
Genetic markers available



<https://doi.org/10.1038/nrg1249>

Nature Reviews | Genetics

Restriction Fragment Length Polymorphism (RFLP)

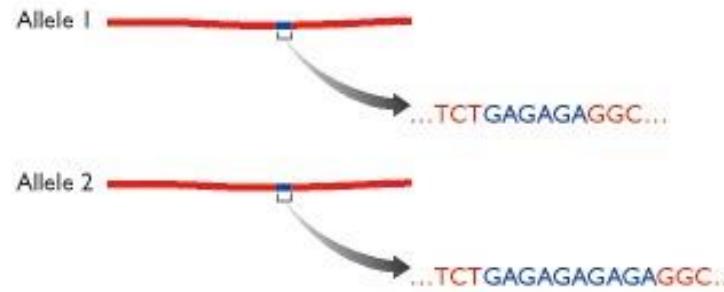


<https://www.ncbi.nlm.nih.gov/books/NBK21116/>

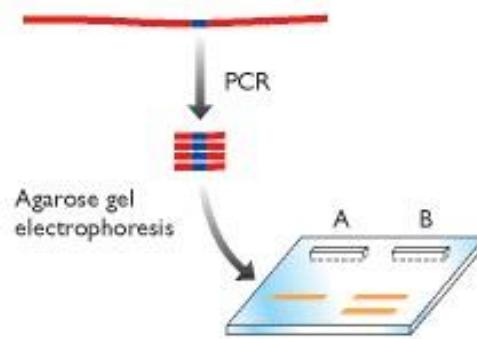
- co-dominant
- highly locus-specific
- time consuming and labor intensive

Simple Sequence Repeat (SSR) assay

(A) Two variants of an SSLP



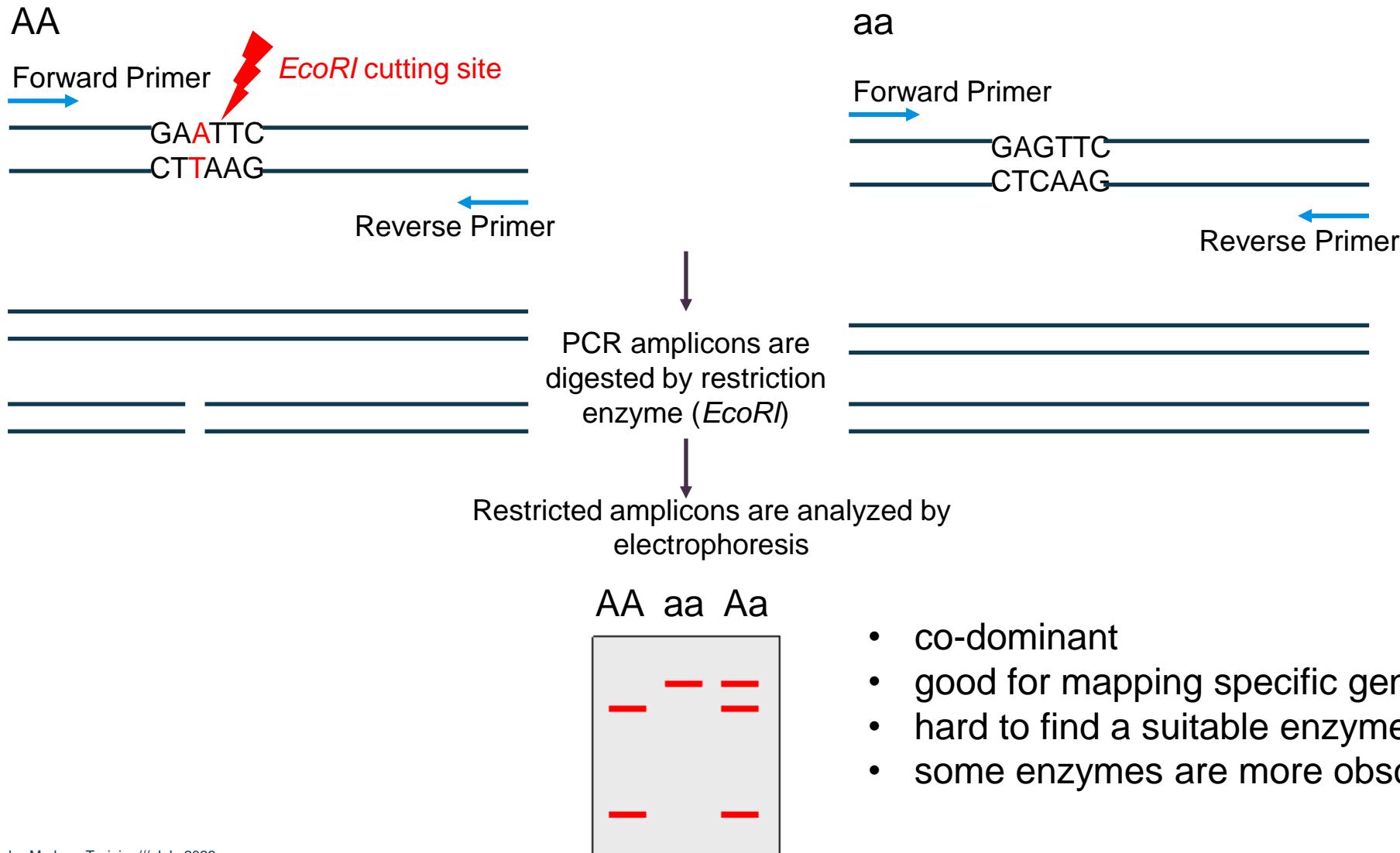
(B) Typing an SSLP by PCR



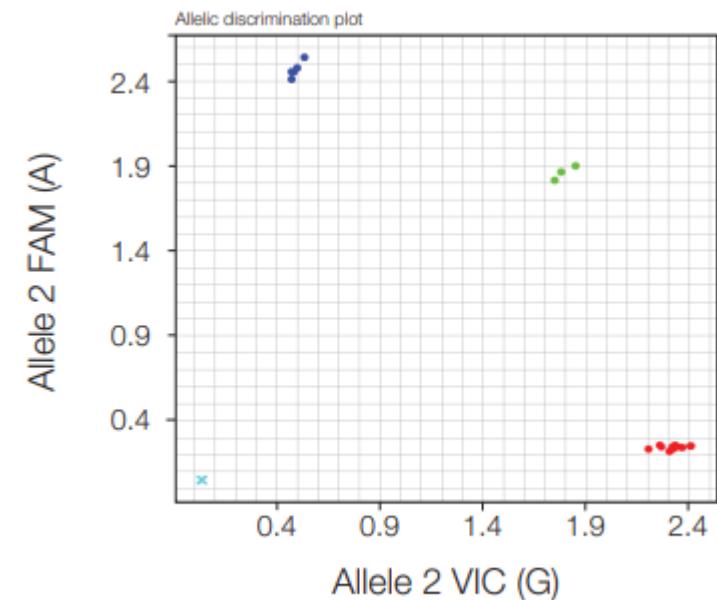
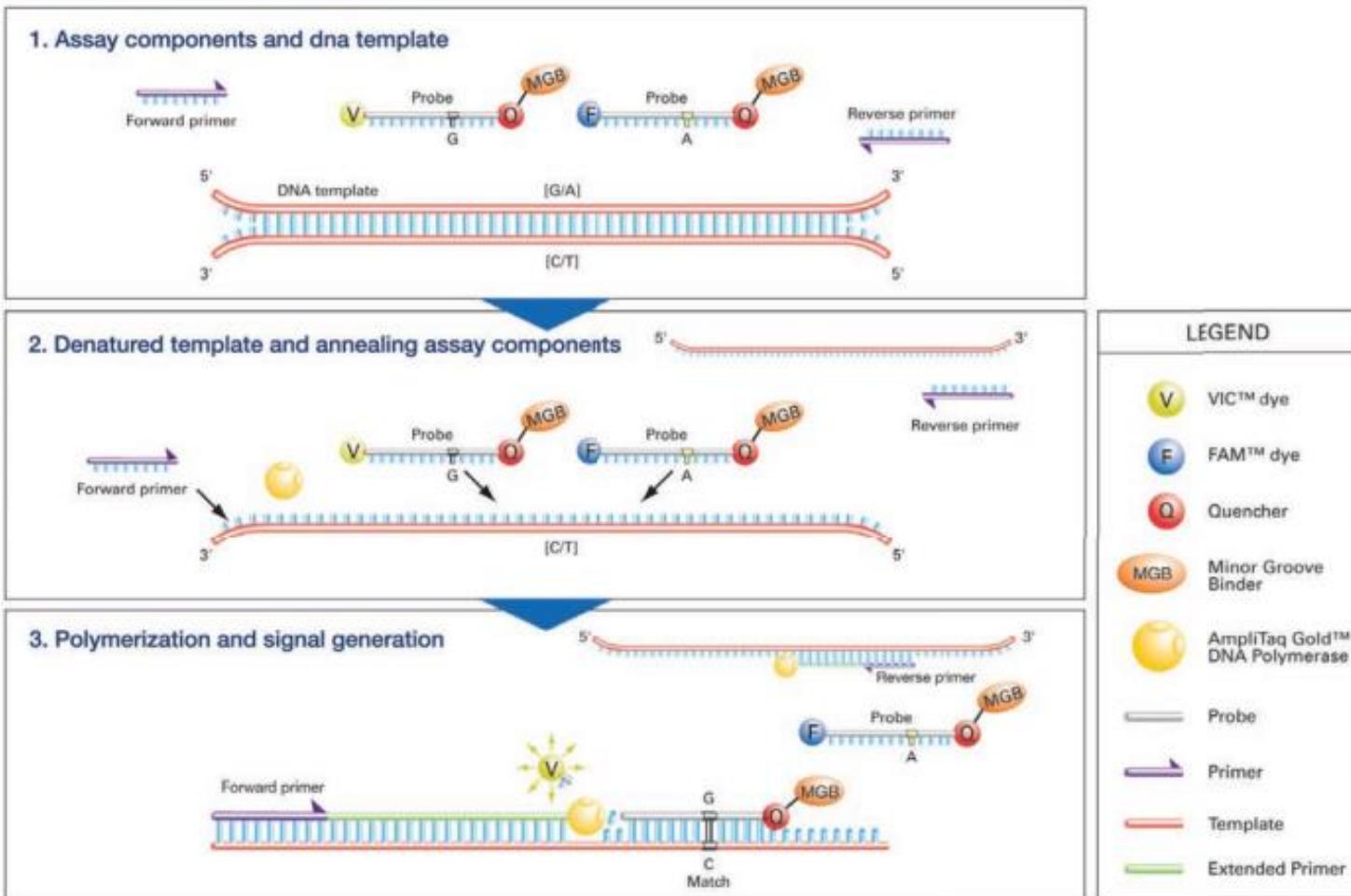
- co-dominant
- multiple alleles
- discovery is time consuming

<https://www.ncbi.nlm.nih.gov/books/NBK21116/>

Cleaved Amplified Polymorphic Sequence (CAPS) assay



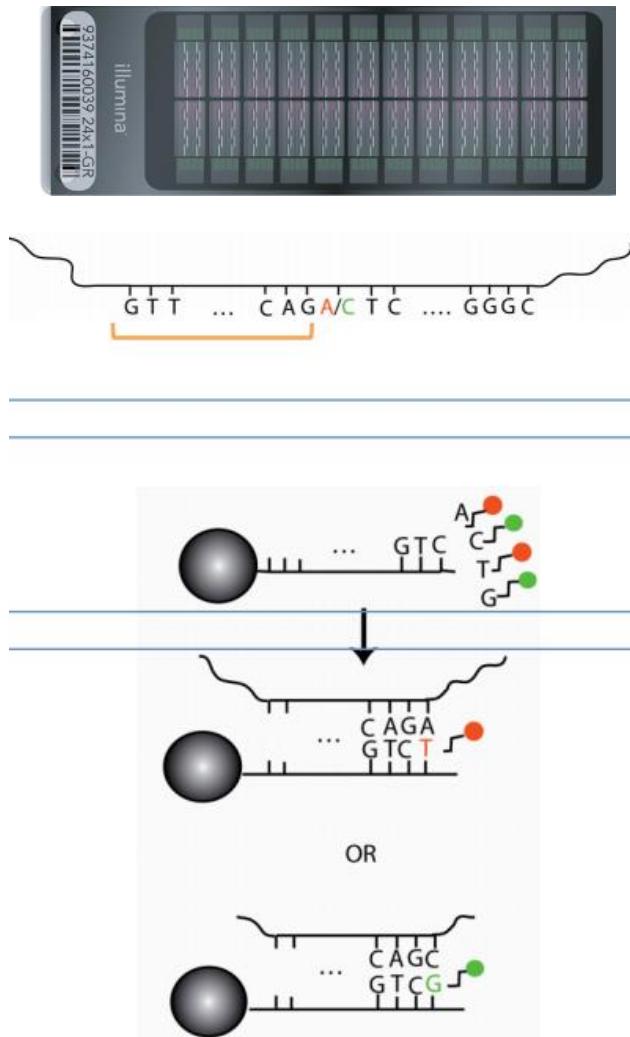
TaqMan SNP genotyping assay



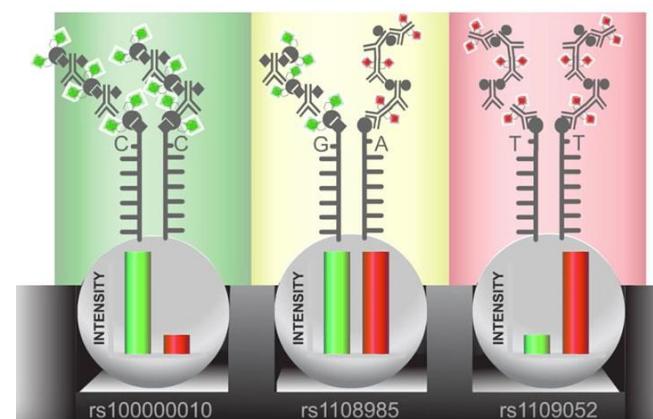
- co-dominant
- good for mapping specific genes
- probe is expensive

https://assets.thermofisher.com/TFS-Artists/LSG/manuals/cms_040597.pdf

Single nucleotide polymorphism (SNP) arrays



- An Illumina BeadChip contains up to millions of bead types.
- One bead type corresponds to each allele per SNP locus.
- Hundreds of thousands to millions of genotypes for a single individual can be assayed at once.
- Attached to each Illumina bead is a 50-mer sequence complementary to the sequence adjacent to the SNP site.
- The single-base extension (**T** or **G**) that is complementary to the allele carried by the DNA (**A** or **C**, respectively) then binds and results in the appropriately-colored signal (**red** or **green**, respectively).



<https://www.illumina.com/science/technology/microarray.html>

Genotyping by sequencing (GBS)

Step 1

construct reduced representation libraries by digesting each DNA sample with a restriction enzyme (ApeKI).

Step 2

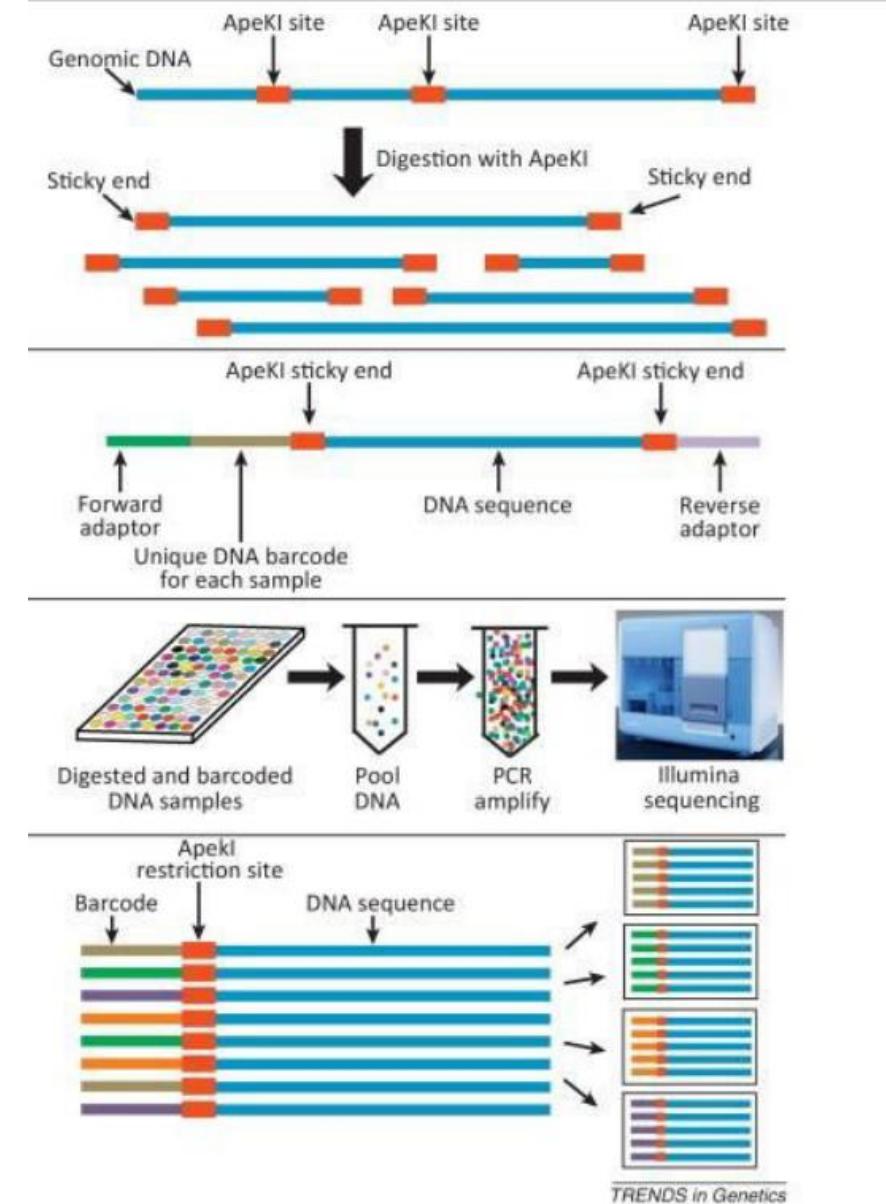
Ligate custom “barcoded adaptors to sticky ends of restriction site. Each sample has its own unique barcode sequence.

Step 3

Pool digested and barcoded DNA into a single tube. Perform PCR amplification, library preparation, and sequencing on illumina platform.

Step 4

Use barcodes to assign sequences to samples. Provide a file of DNA sequence data for each sample.



Genotyping by sequencing (GBS)

--- a multiplex targeted sequencing method

Step 1 Select a set of SNPs obtained through GBS and design a panel of multiplexed PCR primers based on flanking sequences of SNPs.

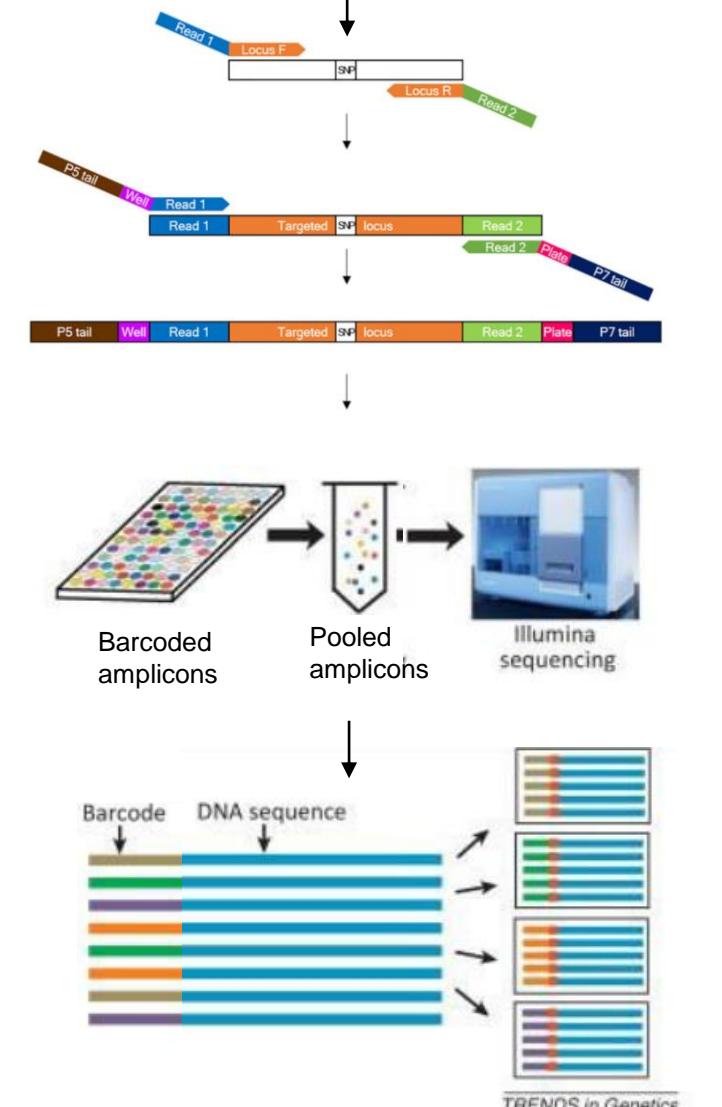
Step 2 Multiple target loci by amplified from a panel of multiplexed primers during first PCR step.

Step 3 Ligate index barcoded adaptors to amplicons through second PCR step. Each sample has its own unique index.

Step 4 Quantify and pool each library for sequencing.

Step 5 Use barcodes to assign sequences to samples. Provide a file of DNA sequence data for each sample.

A set of selected SNP markers obtained through GBS or based on other information resources.



Low-coverage sequencing

Step 1

Shear long genomic DNA into small fragments.

Step 2

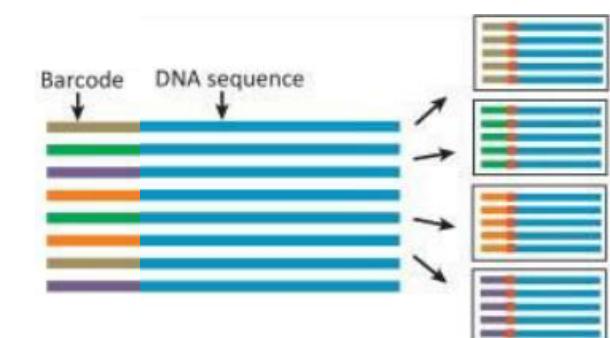
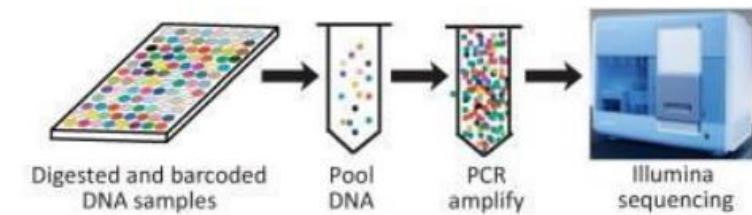
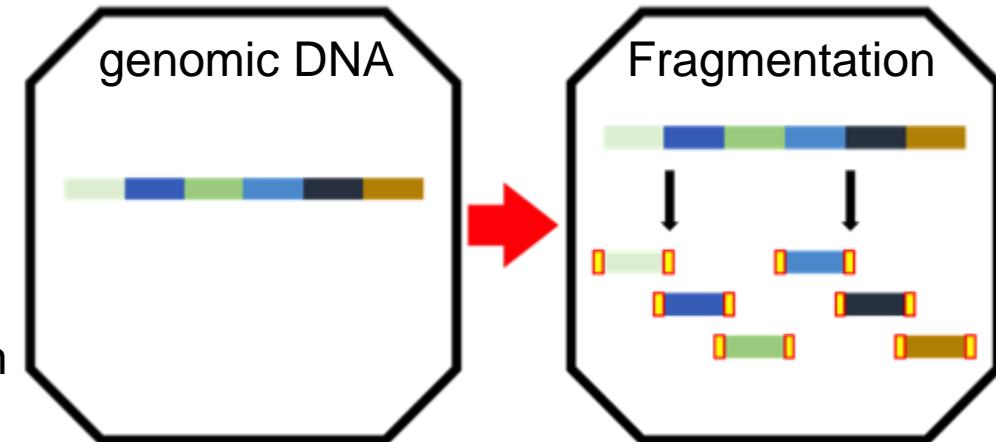
Ligate custom “barcoded adaptors to fragmented DNAs. Each sample has its own unique barcode sequence.

Step 3

Pool barcoded DNA into a single tube. Perform PCR amplification, library preparation, and sequencing on illumina platform.

Step 4

Use barcodes to assign sequences to samples. Provide a file of DNA sequence data for each sample.



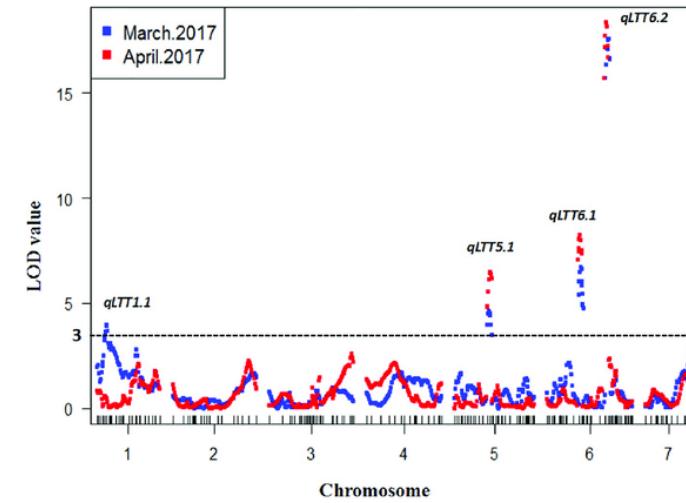
Uses of genetic markers---Genetic mapping

GWAS



<https://doi.org/10.3390/plants10050895>

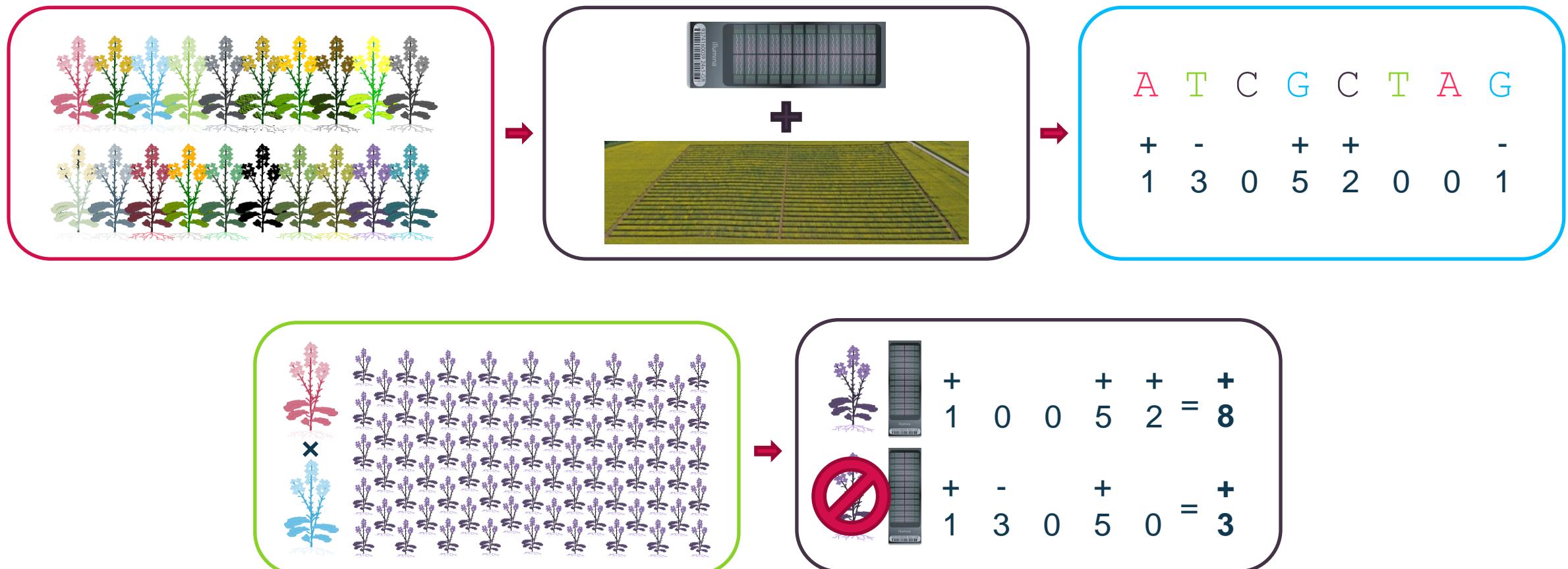
QTL analysis



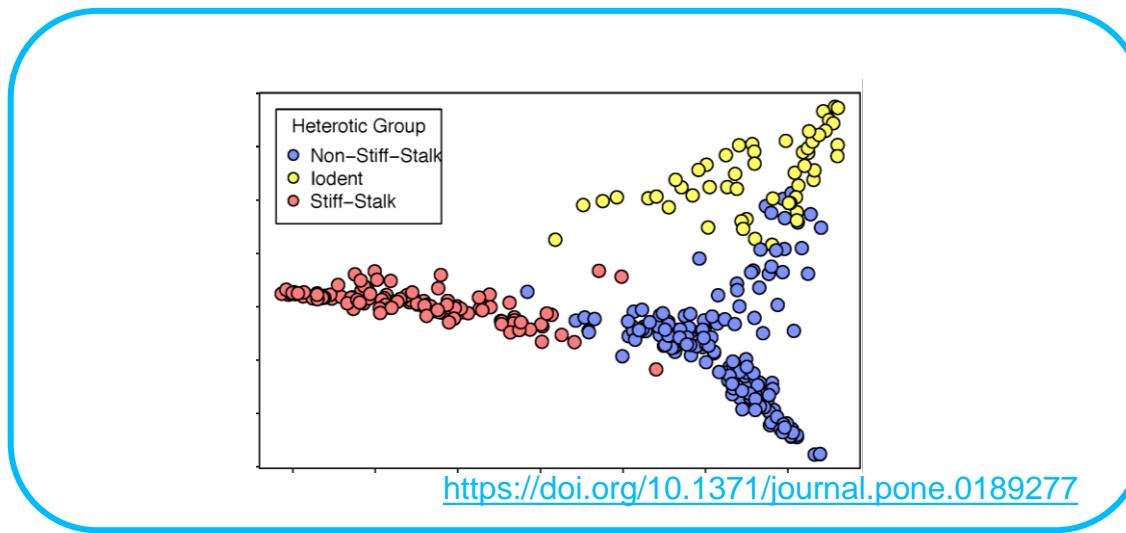
<https://www.frontiersin.org/articles/10.3389/fpls.2019.01620/full>

Uses of genetic markers

--- Marker-Assisted Selection using genome-wide markers



Uses of genetic markers—Genetic diversity analysis



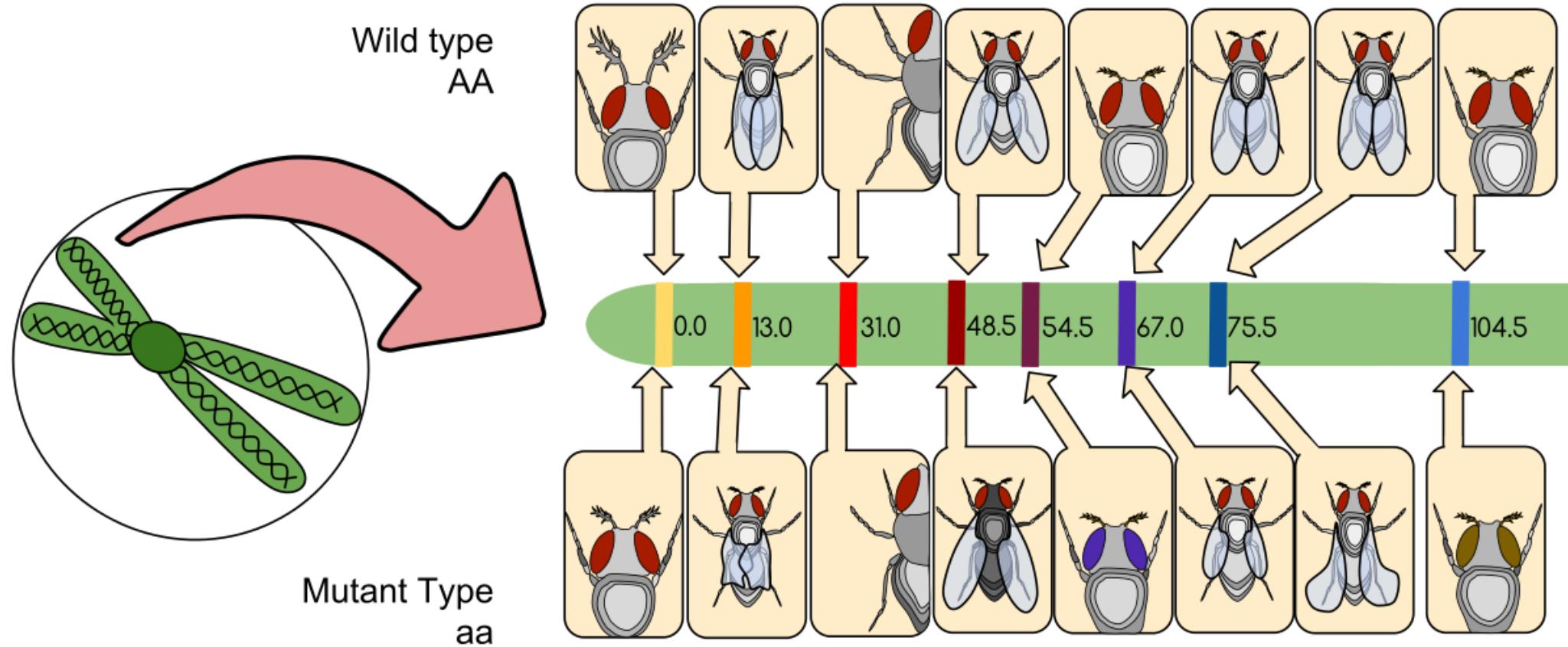
Bayer Russia Molecular Marker Training : Genetic Markers and Technologies



- 1** Genetic Markers for Discovery
- 2** Genetic Mapping
- 3** Trait-linked Marker Development
- 4** Markers for Genome Wide Selection
- 5** Bayer Genotyping Technologies

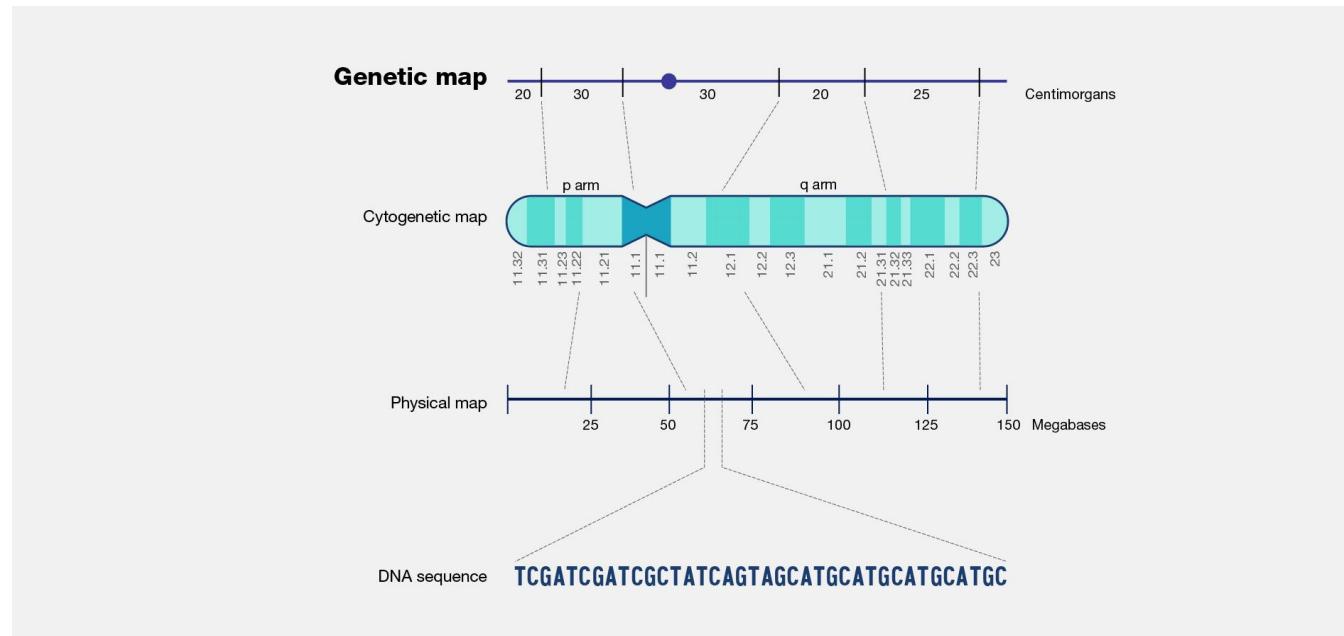
What is genetic mapping?

Gene mapping or genome mapping describes the methods used to identify the location of a gene on a chromosome and the distances between genes.



Mapping approaches

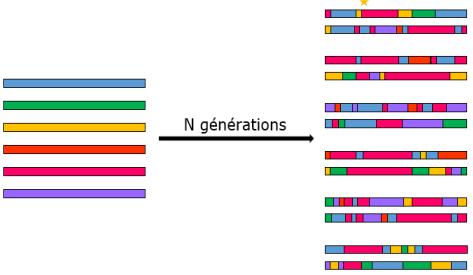
1. **genetic maps** (also known as linkage maps): distances are based on the genetic linkage information.
 - offering insights into the nature of different regions of the chromosome
 - recombination rates, and such rate is often indicative of gene-rich vs usually gene-poor regions of the genome
2. **physical maps**: use actual physical distances usually measured in number of base pairs.
 - a more accurate representation of the genome



<https://www.genome.gov/genetics-glossary/Genetic-Map>

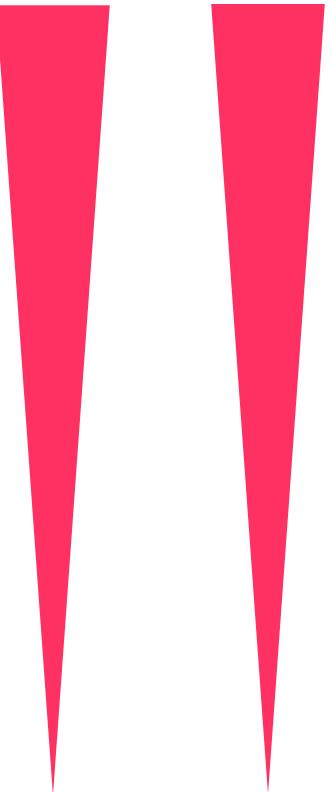
Native Trait discovery toolbox

GWAS – Genome Wide Association Study

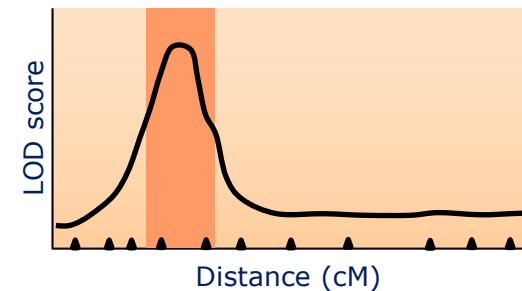
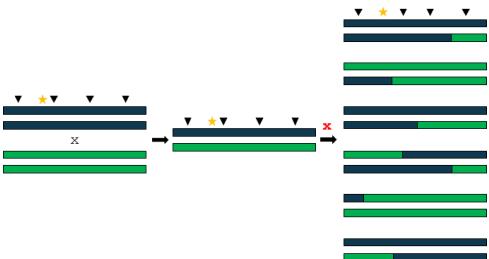


- Uses historical data
- Scans large germplasm pools
- Requires data QC

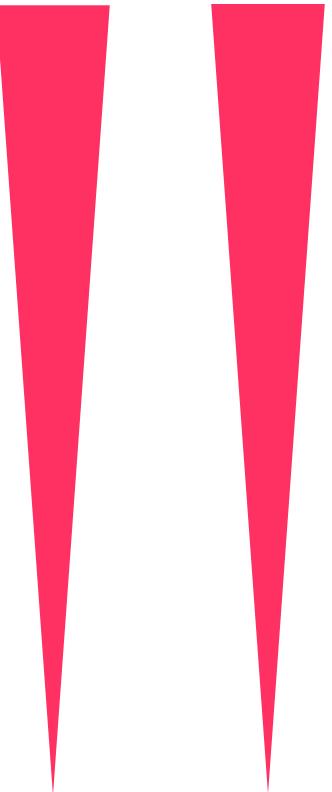
Amount of data Allele Frequency



Bi-parental mapping



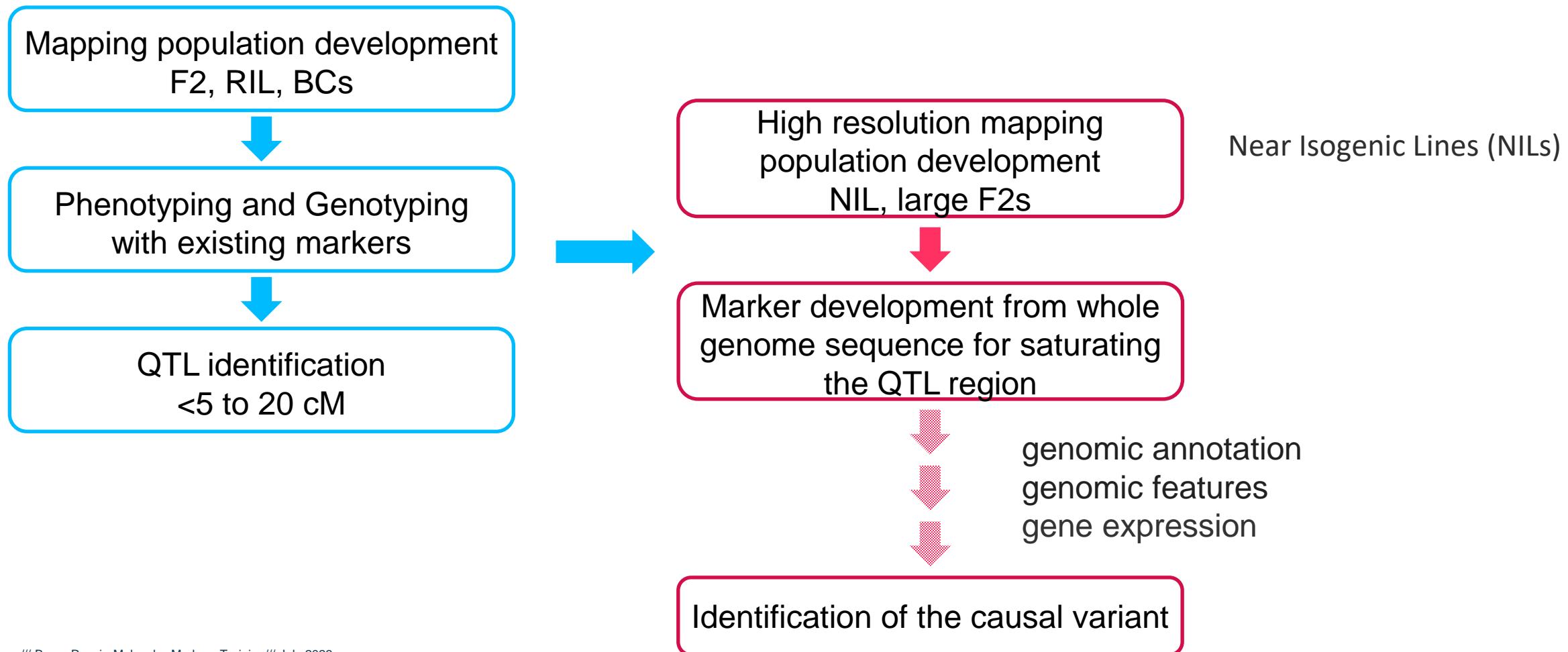
- Requires population development with contrasting parents for the trait of interest
- Targets specific germplasm pool

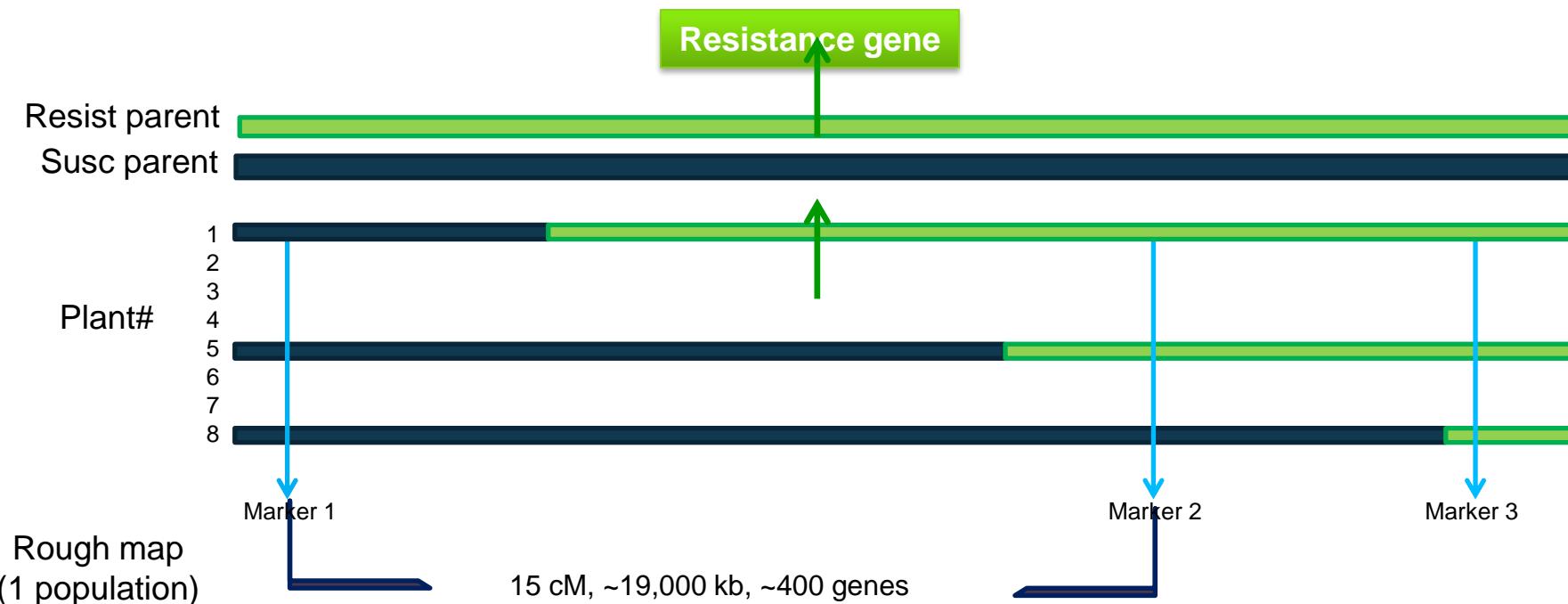


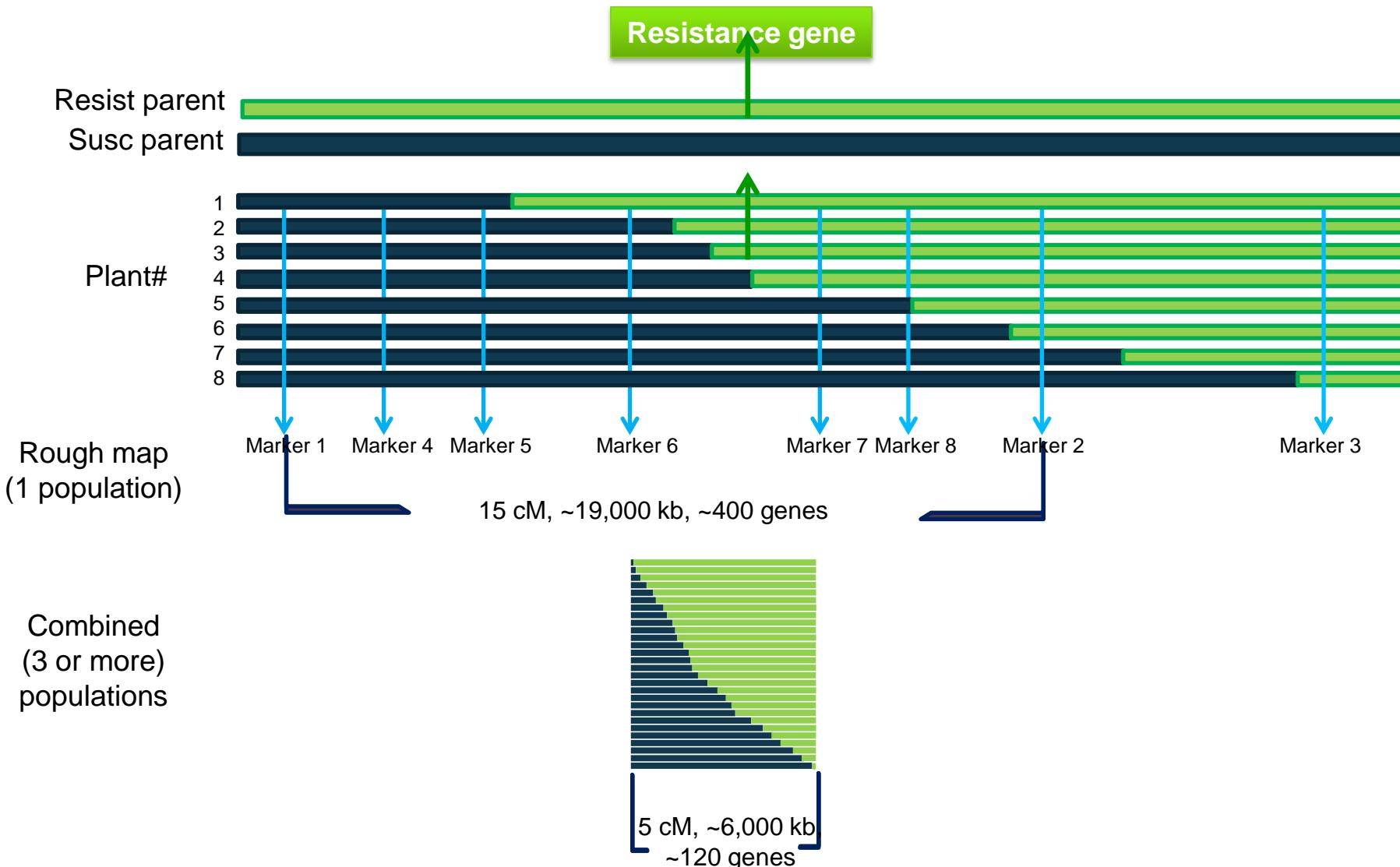
Adapted from Mackay *et al.*, 2009

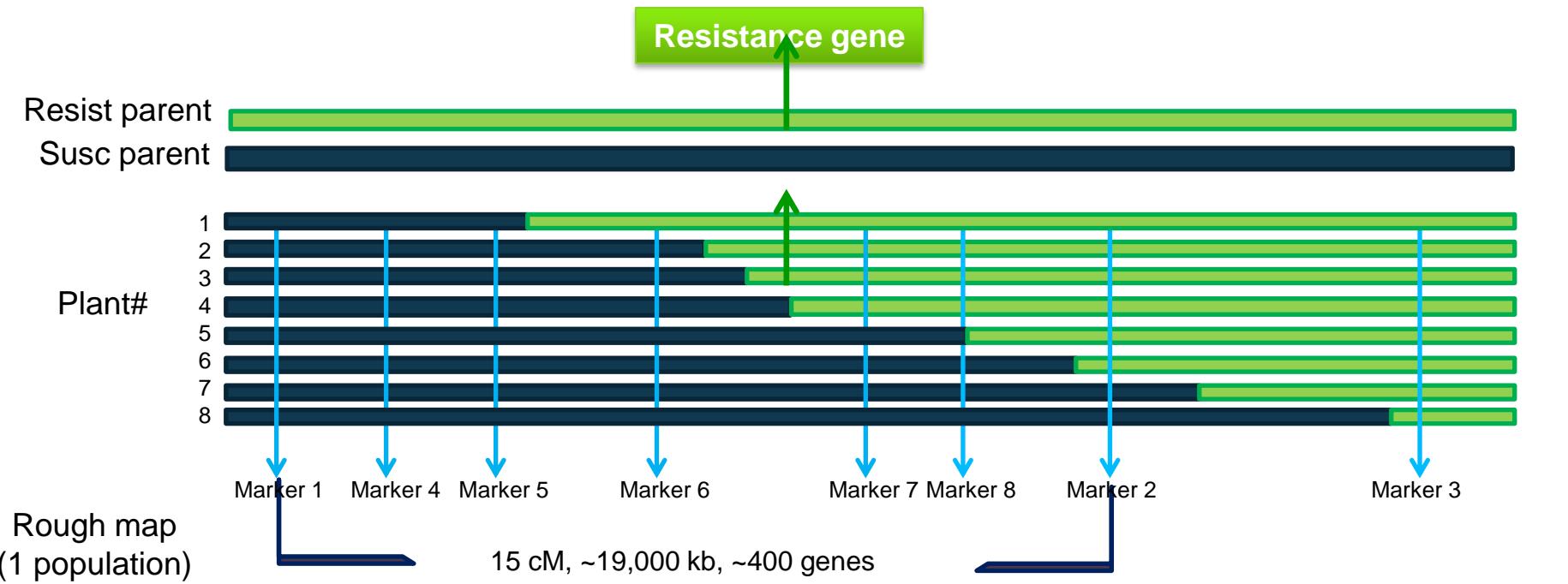
Fine-mapping: what are the causal SNPs?

Fine mapping is the process by which the location of a QTL is reduced from an initial interval of 20 cM or more to an interval of a few cM or less.







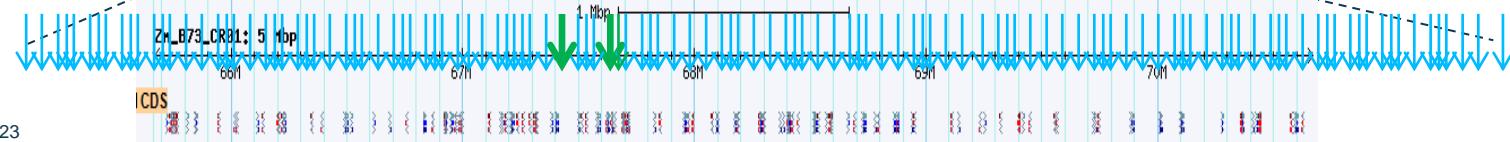


Combined
(3 or more)
populations

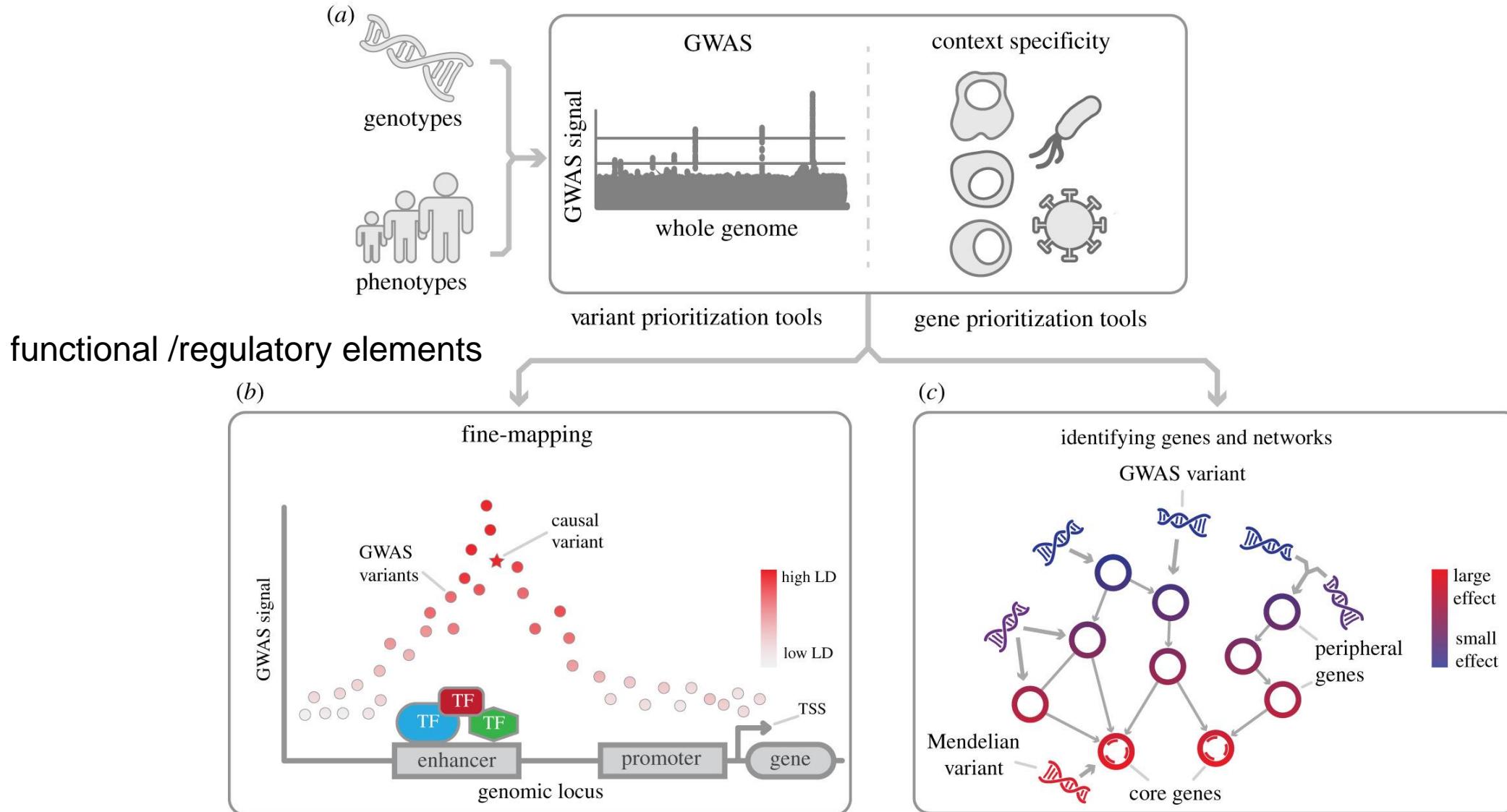


1000 kb

Fine mapping



Fine-mapping and gene prioritization approaches





Why fine-mapping?

---Enable rapid, routine high-resolution mapping of traits for Breeding and Disease program.

1. Reduction of a QTL linkage interval to a **smaller region**

- // fewer, higher predictive markers
- // more precise selection models
- // improve stacking
- // minimize the interpretation of pleiotropy and close linkage
- // Allele/region IP

2. Identification of causal **genes**

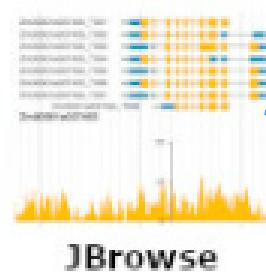
molecular characterization and functional studies of the relationship between gene and trait, and the wider genetic network

3. Pinpoint of the **causative genomic lesion**

Suggest experimental perturbation - editing



Bioinformatics tools—JBrowse



The MaizeGDB homepage features a header with the MaizeGDB logo, a search bar, and navigation links for Home, About, Community, Genome Browsers, Genomes, Tools, Data Centers, and Feedback. A blue arrow points from the JBrowse figure to the "JBrowse" link in the Quick Links section. The main content area includes sections for Reference Assembly (with links to B73 ASSEMBLY, B73 ANNOTATION, and ALL GENOMES), Common genome assembly/annotation tasks, Contribute data (with links to Contribute your data and Make your data FAIR), and @MaizeGDB Tweets. A "Nothing to see here - yet" message indicates no tweets are available. The page also lists Other USDA Crop Databases (GrainGenes, SorghumBase, SoyBase) and Funding Sources (USDA). A Database footer provides update information.

Chinese Version (中文版) Download

Search all data

Home About Community Genome Browsers Genomes Tools Data Centers Feedback

Welcome to MaizeGDB!

MaizeGDB is a community-oriented, long-term, federally funded informatics service to researchers focused on the crop plant and model organism Zea mays.

Quick Links

- JBrowse
- GBrowse
- NAM Genomes
- Downloads
- BLAST
- qTeller
- MaizeMine
- Protein Structure
- Metabolic Pathways
- Hot New Papers
- Newly Characterized Genes
- Diversity SNPs Traits
- Mutants and Phenotypes
- Maize Meeting
- Maize Genetics Cooperation
- AgBioData

Other USDA Crop Databases

Funding Sources

United States Department of Agriculture

Database

Last update: June 8, 2023
Next update: July 11, 2023

more news | archived news

View on Twitter

<https://www.maizegdb.org/>



Bioinformatics tools—JBrowse

The screenshot shows the JBrowse bioinformatics tool interface. On the left, there is a sidebar titled "Available Tracks" with sections for "Assembly" (31), "Annotations" (45), "Protein Alignments" (31), "Diversity" (35), "Epigenetics and DNA-Binding" (617), "RNA-Seq" (179), "Insertions" (4), and "Repetitive Elements" (2). Below these are expanded sections for "Assembly" (31) and "Annotations" (45). Under "Annotations", there is a checkbox for "select all from category" followed by several options: "Official gene models for Zm-B73-REFERENCE-NAM-5.0" (unchecked), "Non-coding gene models" (unchecked), "PANNZER predicted GO terms and functional descriptions (MaizeGDB - 2022)" (unchecked), "NCBI Annotations" (unchecked), "NCBI Gnomon Gene Models" (unchecked), "Zm-B73-REFERENCE-GRAMENE-4.0" (checked), and "Filtered Gene Set [from GMAP]" (unchecked). Under "Filtered Gene Set [from GMAP]", there are options for "B73 RefGen_v3 Filtered Gene Set [from GMAP]" (unchecked), "Transcription Start Site Root" (unchecked), "Transcription Start Site Shoot" (unchecked), and "B73 Mikado transcriptome" (unchecked). At the bottom of the sidebar, there is a section for "Pangenome" (35).

The main area of the interface shows a genomic track for "Zm-B73-REFERENCE-NAM-5.0". The top navigation bar includes "Select Genome" (set to "Zm-B73-REFERENCE-NAM-5.0"), "Track", "View", and "Help". The genome browser displays a genomic region on chromosome 2 (chr2) from approximately 228,762,500 to 228,812,500. The track shows several genomic features, including a large feature labeled "Zm00001d007345_T001" and several smaller features labeled "Zm00001d050424_T001", "Zm00001d007346_T001", "Zm00001d013628_T001", and "Zm00001d007347_T001". The interface includes various controls for zooming and navigating the genome.



Bioinformatics tools—JBrowse

Available Tracks

- Filter tracks
- Assembly 31
- Annotations 45
- Protein Alignments 31
- Diversity 35
- Select all from category
 - SNPs from dbSNP remapped from B73v4
 - SNPs from the EVA (Release 3 - 2022)
 - GWAS SNPs for 21 traits (Li. 2022 - Wang & Wang Labs)
 - Selective Sweeps for breeding improvement (Li. 2022 - Wang & Wang Labs)
- Pangenome 31
- Select all from category
 - Bins
 - Core Bin Markers
 - Pangenome markers
 - GWAS SNPs from GWAS Atlas database
 - Illumina MaizeSNP50 BeadChip remapped from B73v3
 - GWAS SNPs from Wallace et al. 2014 26
 - Select all from category
 - B73 SNPs
 - B97 SNPs
 - CML103 SNPs
 - CML228 SNPs
 - CML247 SNPs
 - CML277 SNPs
 - CML322 SNPs
 - CML323 SNPs
 - CML333 SNPs
 - CML52 SNPs
 - CML69 SNPs
 - HP301 SNPs
 - IL14H SNPs
 - KI11 SNPs
 - KI3 SNPs
 - Ky21 SNPs
 - M162W SNPs
 - M37W SNPs
 - Mo18W SNPs
 - M71 SNPs
 - NC350 SNPs
 - NC358 SNPs
 - Oh43 SNPs
 - Oh7B SNPs
 - P39 SNPs
 - Tx303 SNPs
 - Tzi8 SNPs

Select Genome Track View Help Zm-B73-REFERENCE-NAM-5.0 Share

0 20,000,000 40,000,000 60,000,000 80,000,000 100,000,000 120,000,000 140,000,000 160,000,000 180,000,000 200,000,000 220,000,000 240,000 chr2 chr2:228721001..228885900 (164.9 Kb) SNP ear infructescence position

228,725,000 228,750,000 228,775,000 228,800,000 228,825,000 Reference Sequence

Zoom in to see sequence Zoom in to see sequence Zoom in to see sequence Zoom in to see sequence

Zm-B73-REFERENCE-GRAMENE-4.0 Filtered Gene Set [from GMAP]

Zm00001d007344_T001 Zm00001d007345_T001
Zm00001d024585_T001

Zm00001d050424_T001
Zm00001d007346_T001
Zm00001d013628_T001
Zm00001d007347_T001

GWAS SNPs for 21 traits (Li. 2022 - Wang & Wang Labs)

EW_2_229287364
ear weight
KWPE_2_229287364
kernel weight per ear

Illumina MaizeSNP50 BeadChip remapped from B73v3

alignment:rs132004051

alignment:rs129237206

GWAS SNPs from GWAS Atlas database

cob length Common rust severity tassel spike length
leaf width cob diameter node count

ear infructescence position
ear infructescence position

Primary Data

Trait ear infructescence position
Type SNP
Position chr2:228790442..228790542 (+ strand)
Length 101 bp

Attributes

rs ID rs129237206
Ref allele G
Alt allele A
Pubmed 24514905
Title The genetic architecture of maize height
Journal Genetics
First Author Jason A. Peiffer
Chr chr2
Open_chromatin_intersect NA
Position 228790492
Pub_date 2/10/14
P-value 5.02E-05
R² NA
Seq_id chr2
Source GWAS_trait
Tf_intersect NA
Tissue plant
V4_pos Chr2:229288764
V5_gene_models NA
Region sequence Loading...

Bioinformatics tools—JBrowse

JBrowse interface:

The JBrowse interface displays genomic data for the *Zm-B73-REFERENCE-NAM-5.0* genome. The main panel shows a genomic track for chromosome 2 (chr2) from position 20,000,000 to 240,000,000. A specific region is zoomed in between 228,782,500 and 228,792,500. The track displays several tracks of data, including:

- Reference Sequence:** Shows the genomic sequence with various features highlighted.
- Zm-B73-REFERENCE-GRAMENE-4.0 Filtered Gene Set [from GMAP]:** Shows gene models with arrows indicating direction.
- GWAS SNPs from GWAS Atlas database:** Shows SNPs from the GWAS Atlas database.
- ear1-Total (ChIP-seq):** Shows ChIP-seq peaks for the ear1-TOTAL sample.
- Ear rep1 (ATAC-seq):** Shows ATAC-seq peaks for the Ear rep1 sample.
- 6 8 mm from tip of ear primordium (RNA-seq):** Shows RNA-seq peaks for the 6 8 mm sample.

Left sidebar (Legend):

Legend items shown in the sidebar:

- ear inflorescence position (green square)
- ear inflorescence position (dark green square)

Left sidebar (Category Tree):

- Diversity (35 items)
- Epigenetics and DNA-Binding (617 items)
 - Bolduc 2012 - Hake lab (4 items)
 - KN1 binding sites ChIP-Seq (4 items)
 - select all from category
 - ear1-ChIP region peaks (ChIP-seq)
 - ear1-ChIP summit peaks (ChIP-seq)
 - ear1-ChIP (ChIP-seq)
 - ear1-Total (ChIP-seq)** (selected)
 - Dong 2019 - Chuck lab (8 items)
 - Ricci 2019 - Schmitz and Zhang labs (166 items)
 - select all from category
 - Chromatin Accessibility, ATAC-seq (8 items)
 - select all from category
 - Leaf rep1 (ATAC-seq)
 - Ear rep1 (ATAC-seq)** (selected)
 - Input rep1 (ATAC-seq)
 - Input rep2 (ATAC-seq)
 - Leaf rep1 summit peaks (ATAC-seq)
 - Leaf rep1 narrow peaks (ATAC-seq)
 - Ear rep1 narrow peaks (ATAC-seq)
 - Ear rep1 summit peaks (ATAC-seq)
- Transcription Factor DNA Binding sites, DAP-seq (65 items)
- Histone Variant Binding sites, ChIP-seq (6 items)
- Histone Modification sites, ChIP-seq (84 items)
- Methylation (3 items)
- NAM Consortium (208 items)
- MOA-Seq (6 items)
- Oka 2017 - Stam lab (2 items)
- Tu & Mejia-Guerra 2020 - Buckler & Zhong Labs (220 items)
- Zeng 2022 - Dawe Lab (1 item)
- Hartwig 2023 (2 items)
- RNA-Seq (179 items)
 - select all from category
 - NAM Consortium (10 items)
 - Stelting 2016 - Kaeplke Lab (97 items)
 - Walley 2016 - Briggs Lab (23 items)
 - select all from category
 - Ear (2 items)
 - select all from category
 - 2 4 mm from tip of ear primordium (RNA-seq)
 - 6 8 mm from tip of ear primordium (RNA-seq)** (selected)
 - Root (5 items)
 - Kernel (6 items)
 - Leaf (4 items)
 - Internode/Meristem (3 items)
 - Reproductive (3 items)

Bioinformatics tools—TaqMan Marker design

The Custom TaqMan® Assays Design Pipeline

<https://www.thermofisher.com/order/custom-genomic-products/tools/genotyping/>

Please enter your sequence in the 5' to 3' direction. Sequences must be between 61 and 5000 nucleotides in length and composed solely of the nucleotides A, C, G, or T. Ensure successful design and performance by reading our [Design and Ordering Guide](#).

Open/Import File Search for Sequences Remove All

Keep all sequences confidential

Status	Name	Sequence	SNP	SNP #	SNP Name
		e.g., CAATTGTCATACGACTTACCGTAGTG[G/C]AGCGTCAGCCATG TTC[A/T]CGTCCAGGAAC			

After entering sequence, click 'Check Format' below. Remove

Do you wish to permanently hide your assay primer/probe sequences in all documents? More Info

No Yes

After entering sequence, click 'Check Format' below. Remove

Do you wish to permanently hide your assay primer/probe sequences in all documents? More Info

No Yes

After entering sequence, click 'Check Format' below. Remove

Do you wish to permanently hide your assay primer/probe sequences in all documents? More Info

No Yes

+ Enter More Sequences Select Species / Scale: -- Please Select --

- At least one target site
- Target sites that are more than 40 bases away from the 5' and 3' ends.
- Target sites that are more than two bases away from any Ns.

AGTGAACGCGATA[G/A]GCANCTCCTGCC

If this is your target site, verify that no Ns are within two bases.

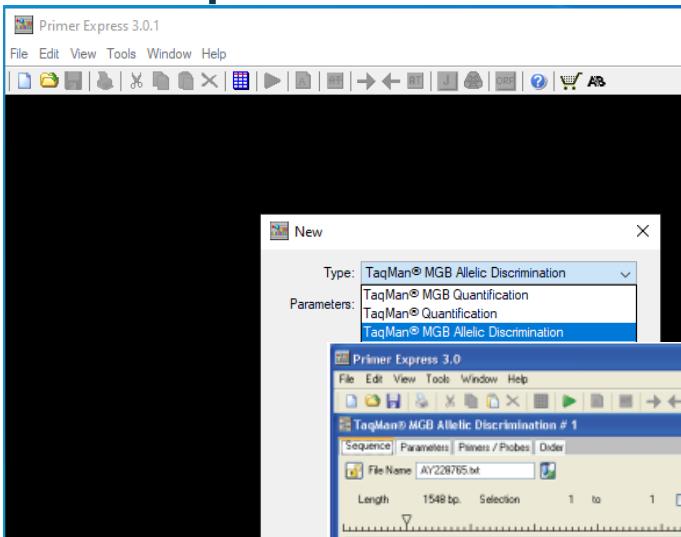
- Target sites that are more than two bases away from any other SNPs.
- At least one specific SNP for assay design (and mask all the remaining nontarget SNPs with Ns).



Bioinformatics tools—TaqMan Marker design



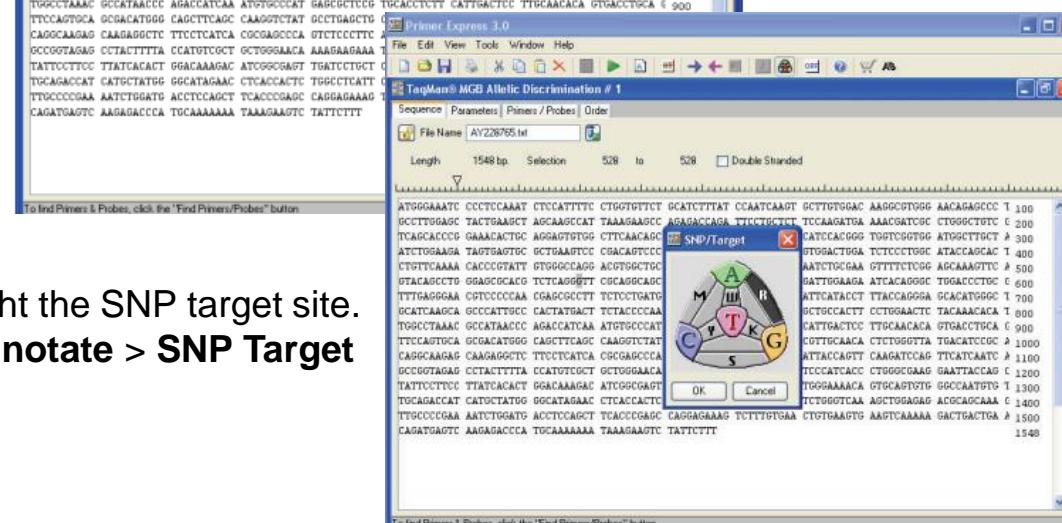
1. Select **File > New** to open the New dialog box



2. Select **Tools > Add DNA File**.
Note you can also copy and paste or type your sequence file in the Sequence tab



3. Highlight the SNP target site.
Edit > Annotate > SNP Target



Bioinformatics tools—TaqMan Marker design



4. Select Tools > Find Primers/Probes

TaqMan® MGB Allelic Discrimination # 1

Candidate Primers & Probes

#	Fwd Start	Fwd Length	Fwd Tm	Fwd %GC	Rev Start	Rev Length	Rev Tm	Rev %GC	Probe1 St...
1	71	27	58	33	171	26	58	31	118
2	71	27	58	33	171	26	58	31	118

Location

Secondary Structure

Length: 304 bp. Selection: 127 to 127 Double Stranded

50 results found.

Primer Express 3.0.1

TaqMan® MGB Allelic Discrimination # 1

File Edit View Tools Window Help

Length: 304 bp. Selection: 127 to 127 Double Stranded

50 results found.

Sequence:

```
TTTTGTAAAC ATCGGGTCGTTTCTTAATGCAATACA
ATTAAAGAAC AAAACATTAA TGACACTAC GTTCTTAATG CAATACAAC
TGCTCTTCTTCTT AATAAANNTA TTCTCAAAA CTTATCCTCG GATAAAAAAG
ATTGATTTAA GGGGTTTCACA TTCTCCAATA TTGTGTTCTAG CATACACAAA
GATGCTAGGA TCAAAACCTCT AACCACTAA TCAAGGGGTA TGTTTATTAA
CCAAACATGT TGTTTAAACCA TTITCTTTT AATACAACAA GTACTTTT
ACTT
```

5. In the **Primer/Probe tab**, select the primer and probe set you want to order. Click on the **Order tab** and Click **AB store** on the toolbar to go the online store

Primer Express 3.0.1

TaqMan® MGB Allelic Discrimination # 1

Order

Sequence: 05_26_23_com_root_biomass... User: Jiani Yang Date: 07/11/23

Type	Name	Sequence	Start bp	Stop bp	Tm
Forward Primer		TCACACTTACGTTCTTAATGCAATACA	71	97	58
Reverse Primer		ATGTGAACCCCTTAAATCAATCTTT	171	146	58
Probe1		TTAGTCTCAAAACTATTTC	118	136	66
Probe2		TTAGTCTCAGAAACTATTCCCTGA	118	141	67



Bioinformatics tools—R script

R is an open-source programming language for statistical analysis and graphing

Many resources are available ranging from fully developed code for certain applications to online communities that help troubleshoot code.

JOURNAL ARTICLE

AlphaSimR: an R package for breeding program simulations

R Chris Gaynor , Gregor Gorjanc, John M Hickey

G3 Genes|Genomes|Genetics, Volume 11, Issue 2, February 2021, jkaa017,

<https://doi.org/10.1093/g3journal/jkaa017>

Published: 07 December 2020 Article history ▾

Zhang et al. BMC Bioinformatics (2023) 24:199
<https://doi.org/10.1186/s12859-023-05318-9>

BMC Bioinformatics

SOFTWARE

Open Access

geneHapR: an R package for gene haplotypic statistics and visualization



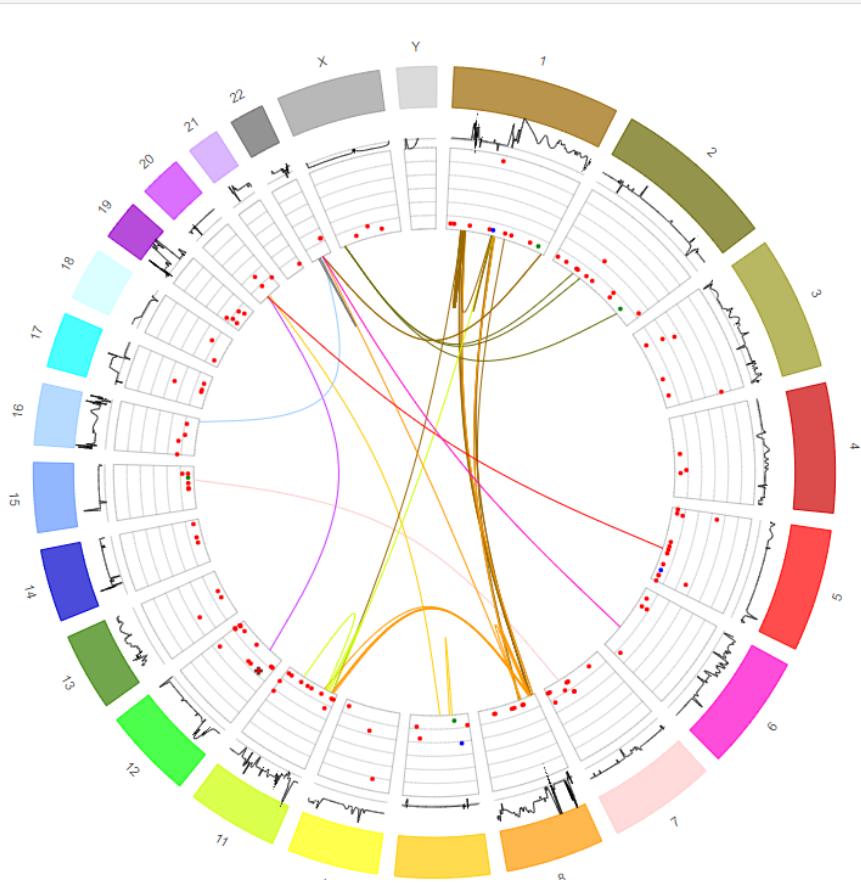
Renliang Zhang¹, Guanqing Jia^{1*} and Xianmin Diao^{1*}

Bioinformatics tools—Shiny for R: Genome Browser



More Information

ICGC PANCREATIC CANCER (DUCTAL ADENOCARCINOMA) - GENOME VIEWER



Cohort Top ClinVar Gene Summary:

HGNC	Chr	Start	From	To	Consequence	Count
SMARCA4	19	11144847	C	T	missense_variant	18
TP53	17	7578437	G	A	stop_gained	18
KRAS	12	25398284	C	T	missense_variant	12
TP53	17	7578437	G	A	exon_variant	10
SMARCA4	19	11144847	C	T	exon_variant	8
TP53	17	7577121	G	A	downstream_gene_variant	6
TP53	17	7577121	G	A	missense_variant	6
KRAS	12	25398285	C	G	missense_variant	4
SMARCA4	19	11144847	C	T	downstream_gene_variant	4
TP53	17	7578437	G	A	downstream_gene_variant	4

Please select a donor ID:

DO49184

SNP Consequences

Resize Factor:



Shiny is an open source R package for building web applications using R

For example shown here, visualization is based on Circos, a way of visualizing whole genomes.

<https://shiny.posit.co/r/gallery/life-sciences/genome-browser/>



Q&A Discussion



Bayer Russia Molecular Marker Training : Genetic Markers and Technologies

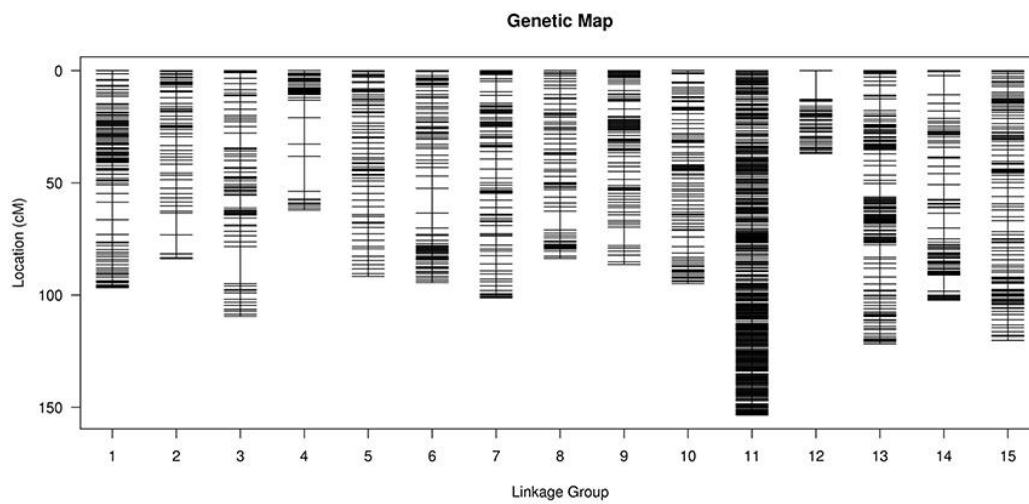


- 1 Genetic Markers for Discovery
- 2 Genetic Mapping
- 3 Trait-linked Marker Development
- 4 Markers for Genome Wide Selection
- 5 Bayer Genotyping Technologies

Two types of genetic markers

Regular markers

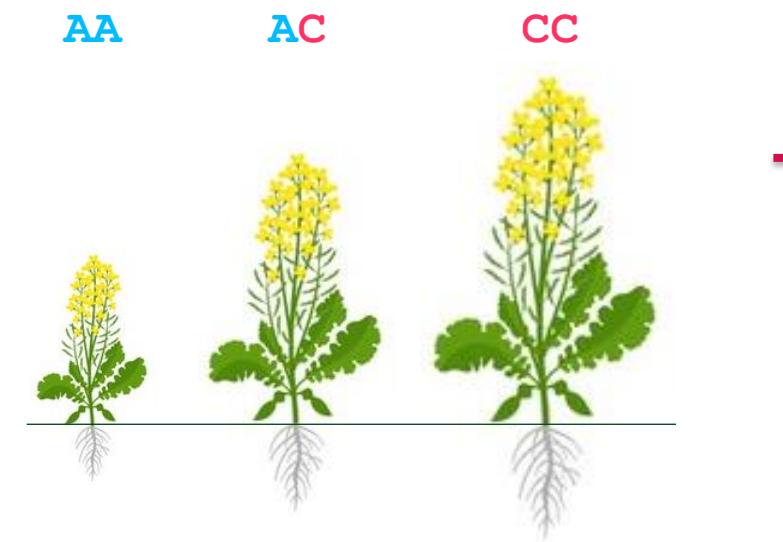
- // Spanned all over the genome
- // Polymorphism with no associated trait
- // Can be used to build genetic maps



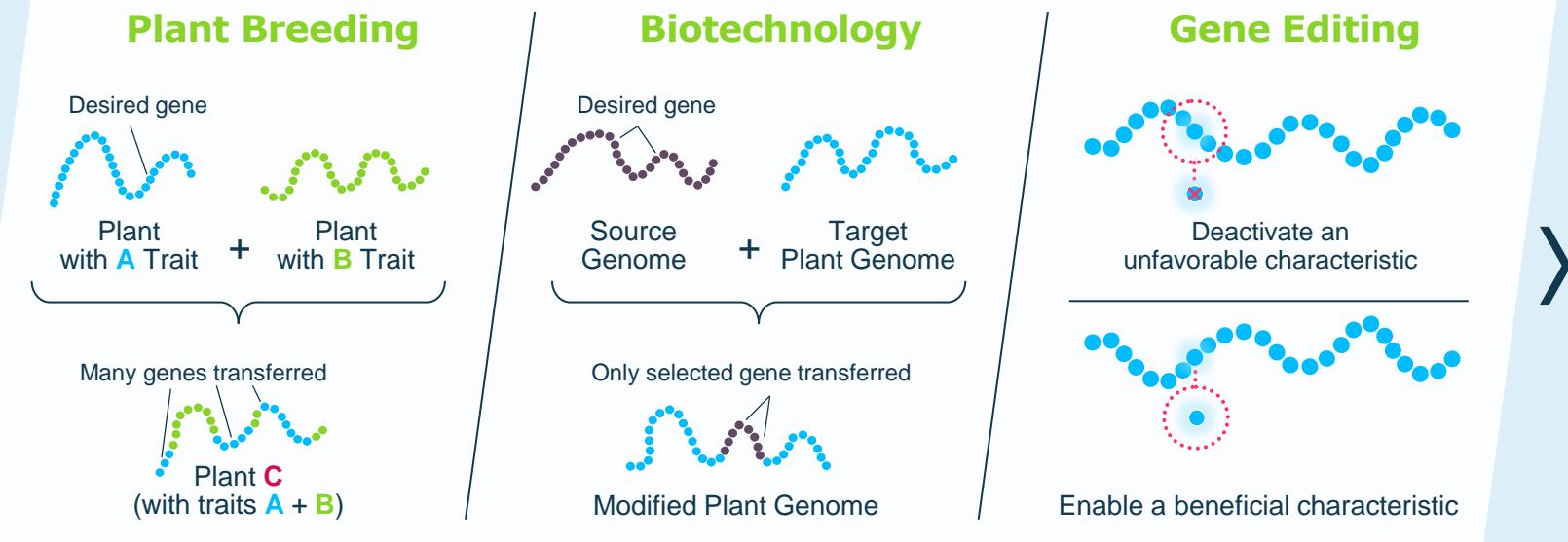
<https://doi.org/10.3389/fpls.2018.00885>

Trait markers

- // Located in specific genomic areas
- // Linked to the variation of a given trait
- // Can be used to follow said trait (MAS)



Technology we're leveraging to develop desired traits



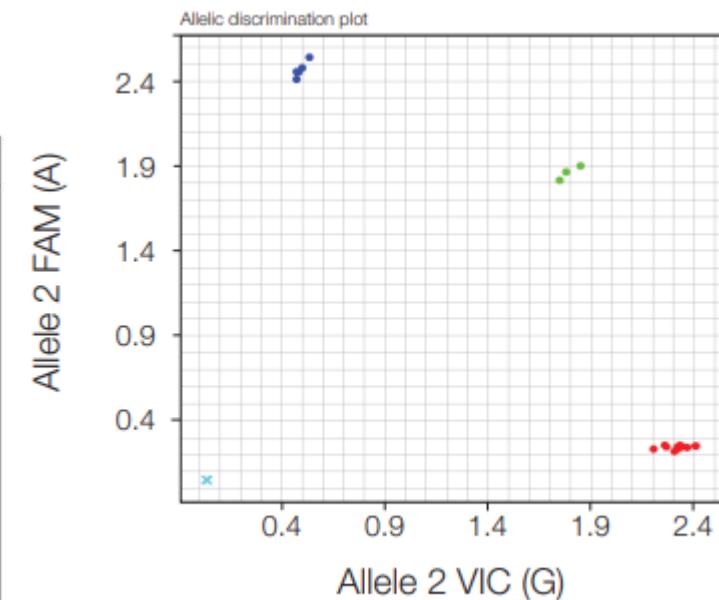
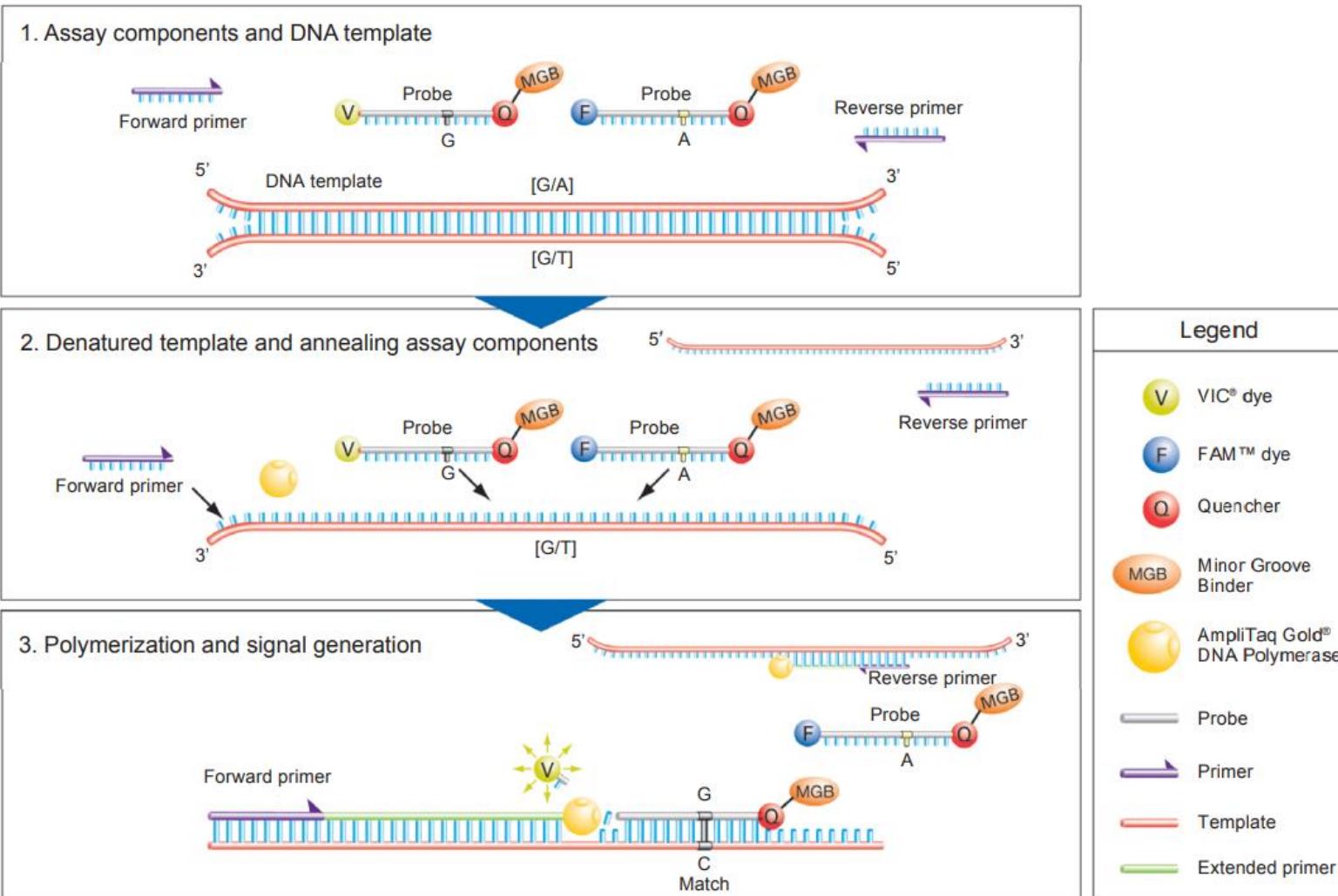


DNA assays are essential to crop science pipelines



Causal variant discovery for native traits

---TaqMan SNP genotyping assay



- co-dominant
- good for mapping specific genes
- Probe is expensive

https://assets.thermofisher.com/TFS-Artists/LSG/manuals/cms_040597.pdf

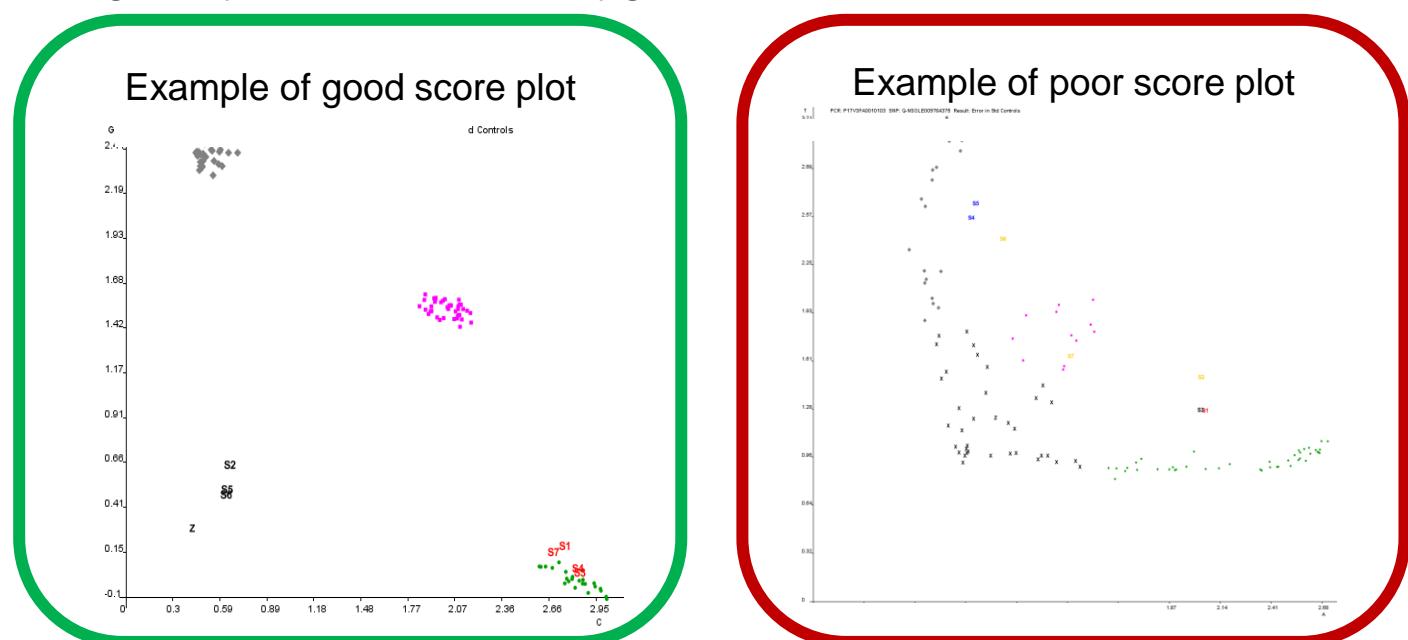
Causal variant discovery for native traits

TaqMan Assay Validation: Score Plot Analysis

1. Assay validation using historical data
2. Assay validation using new population:

F2 or any segregating population

Or, it can be a mixture of the parental genotypes (two homozygous classes) and an F1 (heterozygote) in a 1:2:1 ratio





Causal variant discovery for native traits

Genetic Accuracy

1. QTL effect requirements

- QTL identified via **linkage analysis**
 - Effect must be confirmed in two other populations (three pops total)
 - Minimum requirement for percent variation explained: $R^2 > 15\%$
 - QTL window $< 5\text{cM}$
- QTL identified via **association analysis**
 - Confirm QTL location and efficacy in at least one bi-parental population (F2 or BCX) or through haplotype mining
 - Minimum requirement for percent variation explained in bi-parental population: $R^2 > 15\%$
 - Haplotype mining: Pedigree and genotypic analysis via Trait Analysis Team of ancestral donor and recoveries combined with phenotypic data
- QTL or gene identified through **public literature**
 - Confirm in bi-parental or haplotype mining/Association analysis within breeding germplasm
 - Minimum requirement for percent variation explained
 - QTL window $< 5\text{cM}$



Causal variant discovery for native traits

Genetic Accuracy

2. Genetic accuracy panel requirements

Genetic Accuracy panel should constitute of the germplasm pool into which the trait will be deployed.

The germplasm or list of lines that constitute the deployment pool should be determined based on

- Pipeline materials (List of lines/inbreds that constitute the breeding pipeline for the crop)
- Heterotic pool (Male or Female)
- Geography
- Maturity
- Other

*Always remember to include the **trait donors** in the deployment pool.



Causal variant discovery for native traits

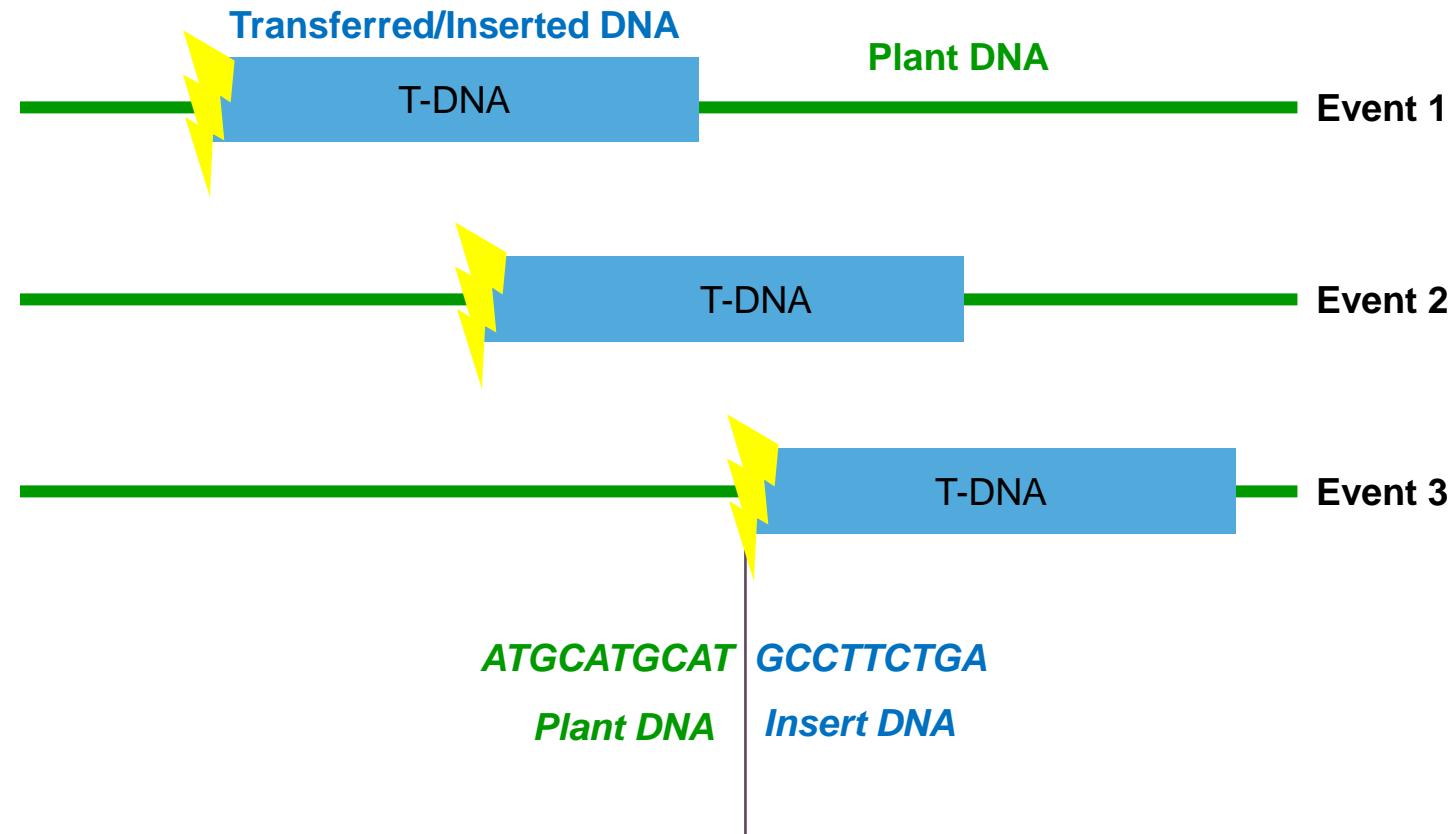
Genetic Accuracy

3. SNP accuracy requirements

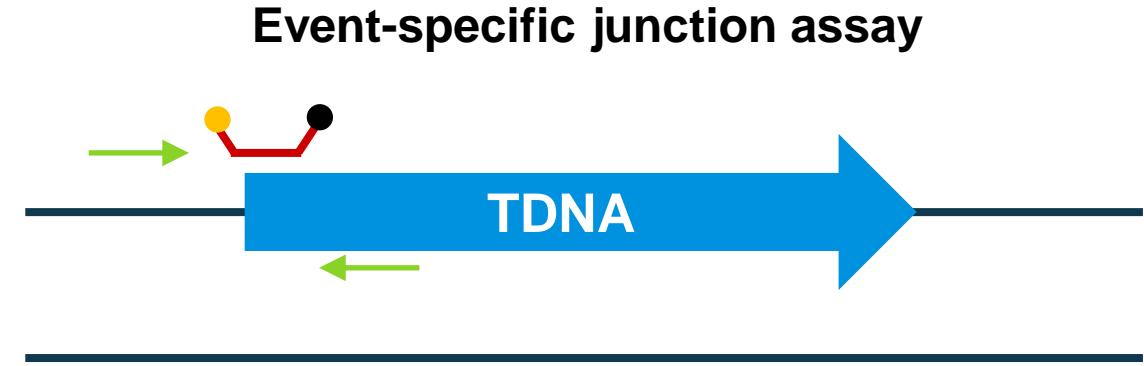
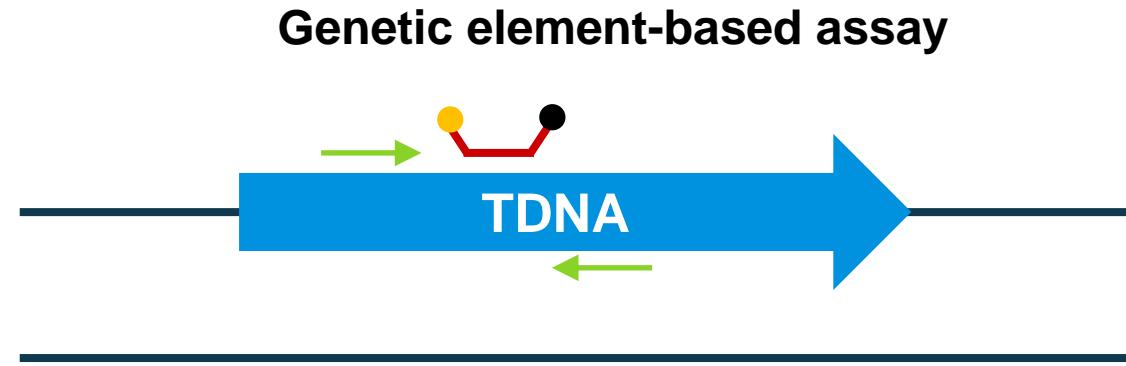
- SNP for trait association must be within 5cM of QTL in mapping analysis.
- False positive rate <5 %: Allele of SNP associated with favorable phenotype should be found in less than 5 % of unfavorable lines
- False negative rate <10%: Allele of SNP associated with unfavorable phenotype found in less than 10 % of favorable phenotype lines

Genomic characterization for biotechnology traits

A Transformation Event is Defined by T-DNA/Plant DNA Junction



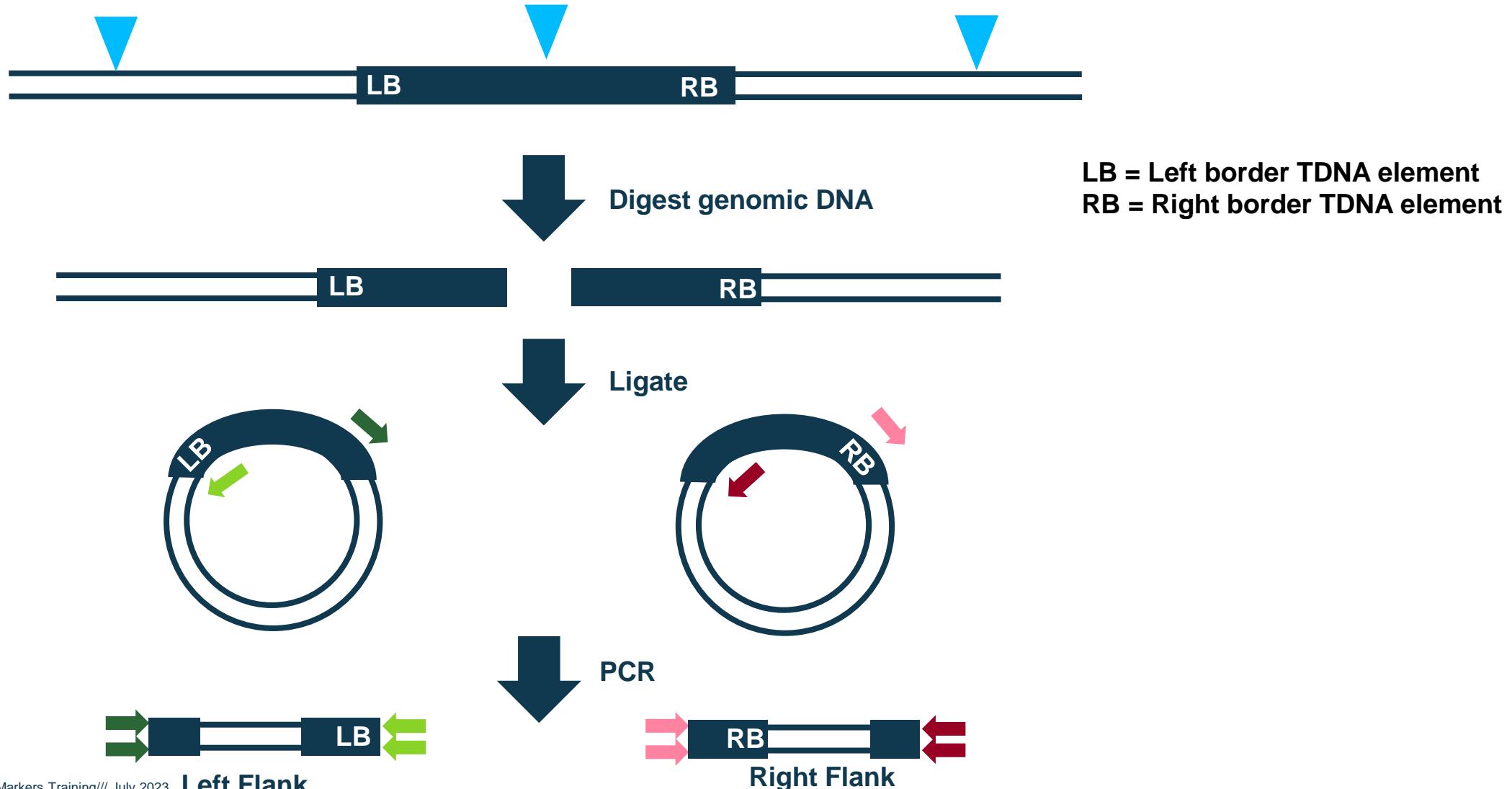
Genetic element-based assays and event junction assays can both be used to assay for transgenic events



- Detects all events from construct with the element sequence
- Can detect events from different constructs if constructs contains the same element sequence
- Used to determine copy number of TDNA
- Does not require genomic flank DNA sequence

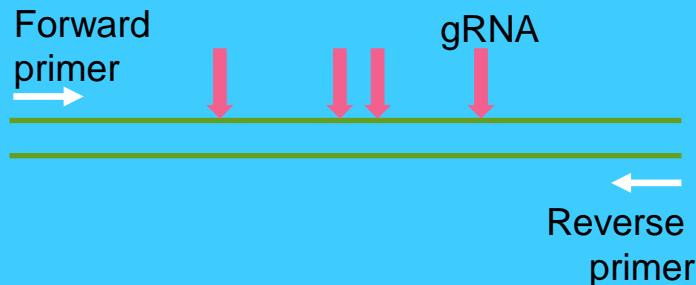
- Typically designed to detect one TDNA insertion
- Often used when the same event will be used or grown many times (as in commercial traits)
- Will not detect multiple events or events from a different construct
- Can be designed as RT-PCR assay or endpoint TaqMan® assay
- Requires isolation of genomic flank DNA for assay development

Inverse PCR can be used to obtain DNA sequence of genomic flank of transgene inserts to develop event-specific assays



Causal variant discovery for gene editing

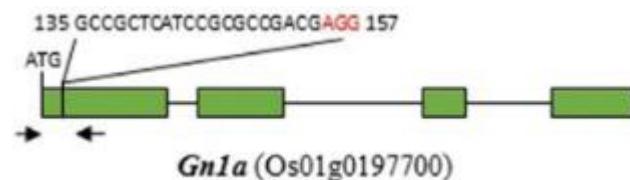
PCR based method



Presence of gene edits
Sequence confirmation of edits

Keep amplicon size within ideal PCR requirements (~300-600bp)

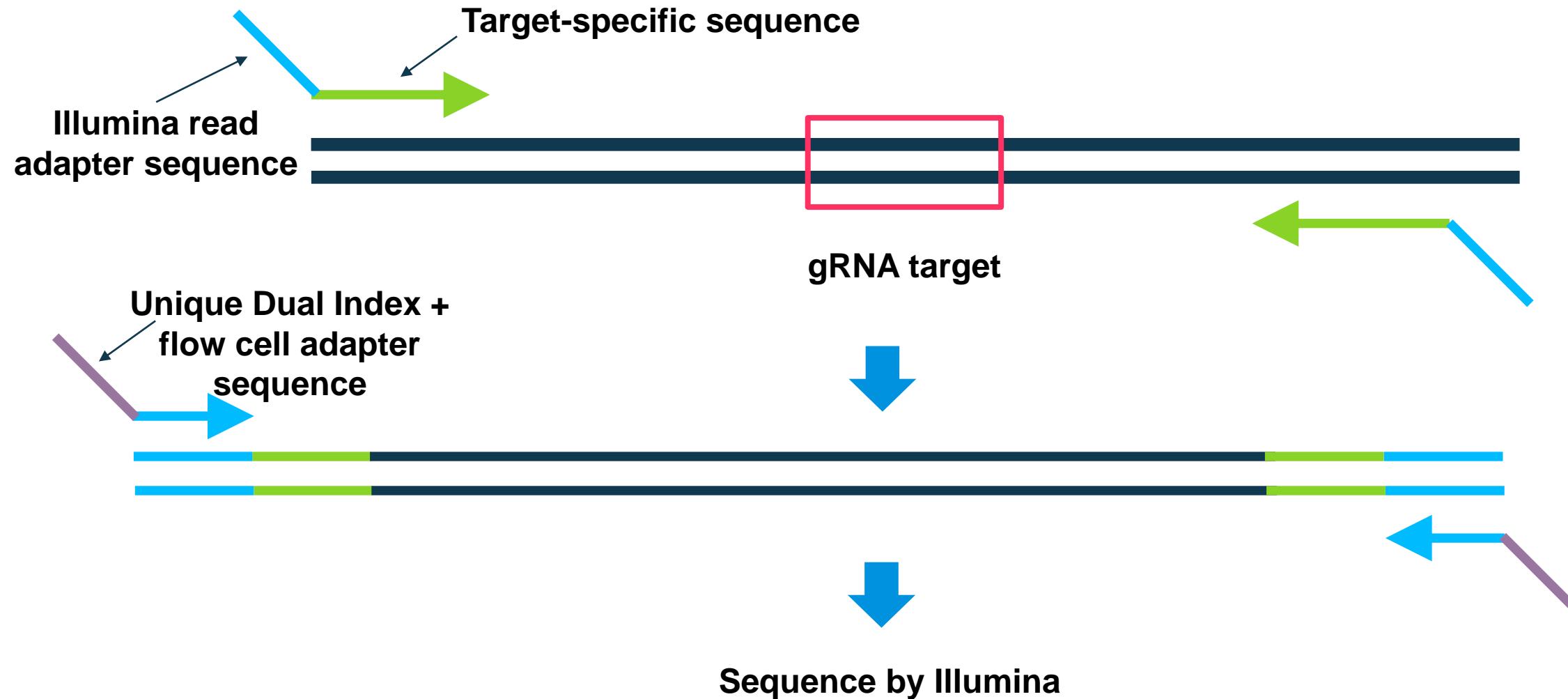
Have gRNA positioned <150bp from the gene specific primers, forward or reverse primers.



CGACCTCGGCATCGC **GCCC**GCTCATCCGGCGACGAGG CGGGCACCGCGCGCCTCCGCCGAC WT
CGACCTCGGCATCGC **GCCC**GCTCATCCGGCGC- ACGAGGCGGGCACCGCGCGCCTCCGCCGAC *gnla-1*
CGACCTCGGCATCGC **GCCC**GCTCATCCGGCGC- ACGAGGCGGGCACCGCGCGCCTCCGCCGAC *gnla -2*
CGACCTCGGCATCGC **GCCC**GCTCATCCGGCGC- - CGAGGCGGGCACCGCGCGCCTCCGCCGAC *gnla -3*
CGACCTCGGCATCGC **GCCC**GCTCA----- CGAGGCGGGCACCGCGCGCCTCCGCCGAC *gnla-10*
----- AC *gnla-112*
CGACCTCGGCATCGC **GCCC**GCTCATCCGGCGCG----- AC *gnla-115*

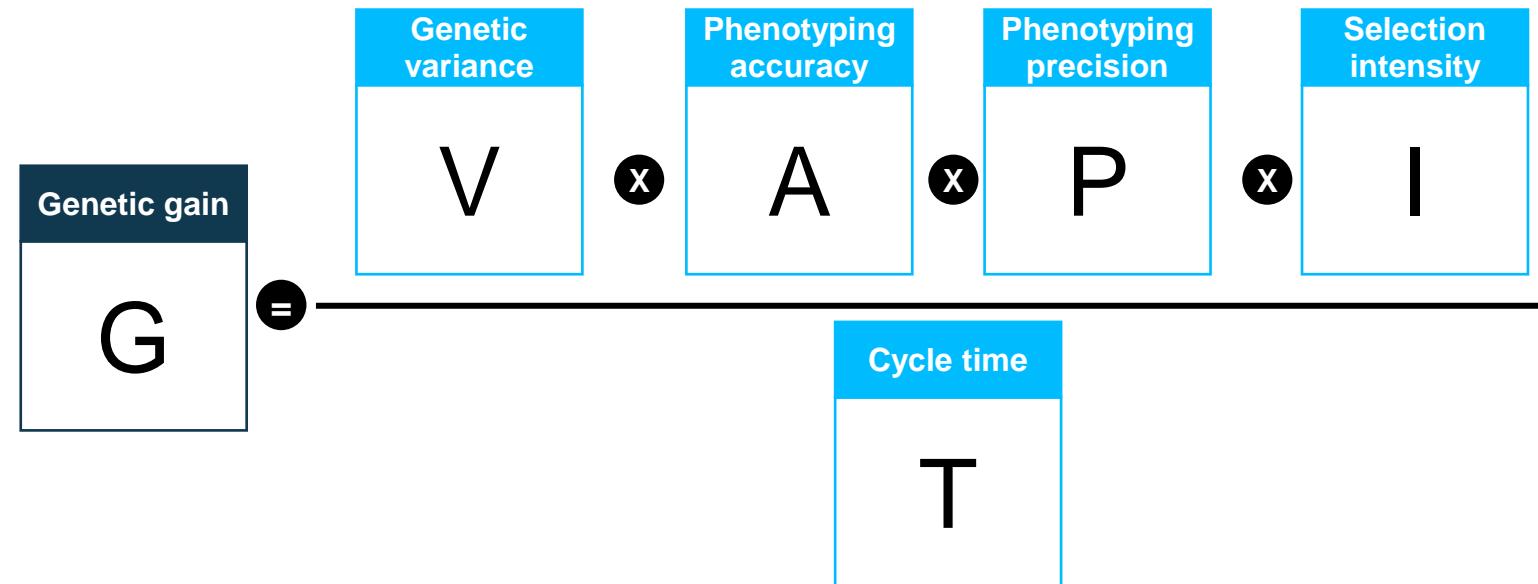
Li et al., 2016

Amplicon sequencing for gene edits involves a two-step PCR method



The breeder's equation:

Molecular breeding has positive impact to V and T

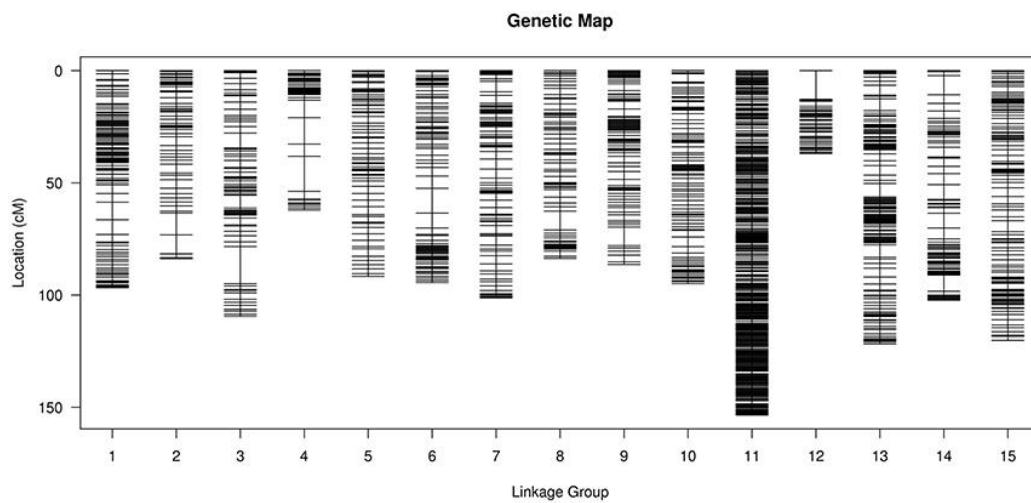


Breeders have been successful whenever they had access to useful **genetic variation** and selection has focused on the **right traits** measured with the **right protocol** in the **right environment**

Two types of molecular markers

Genome-wide markers

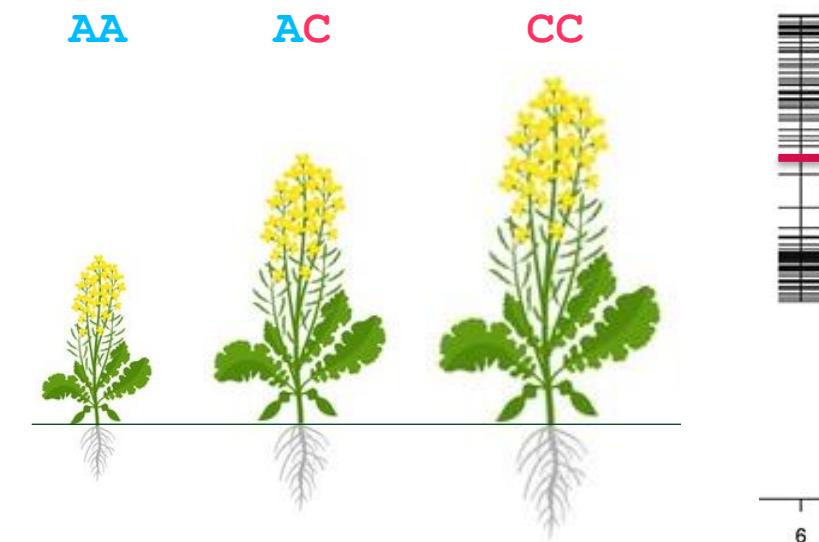
- // Span all over the genome
- // Polymorphic with no associated trait
- // Can be used to build genetic maps



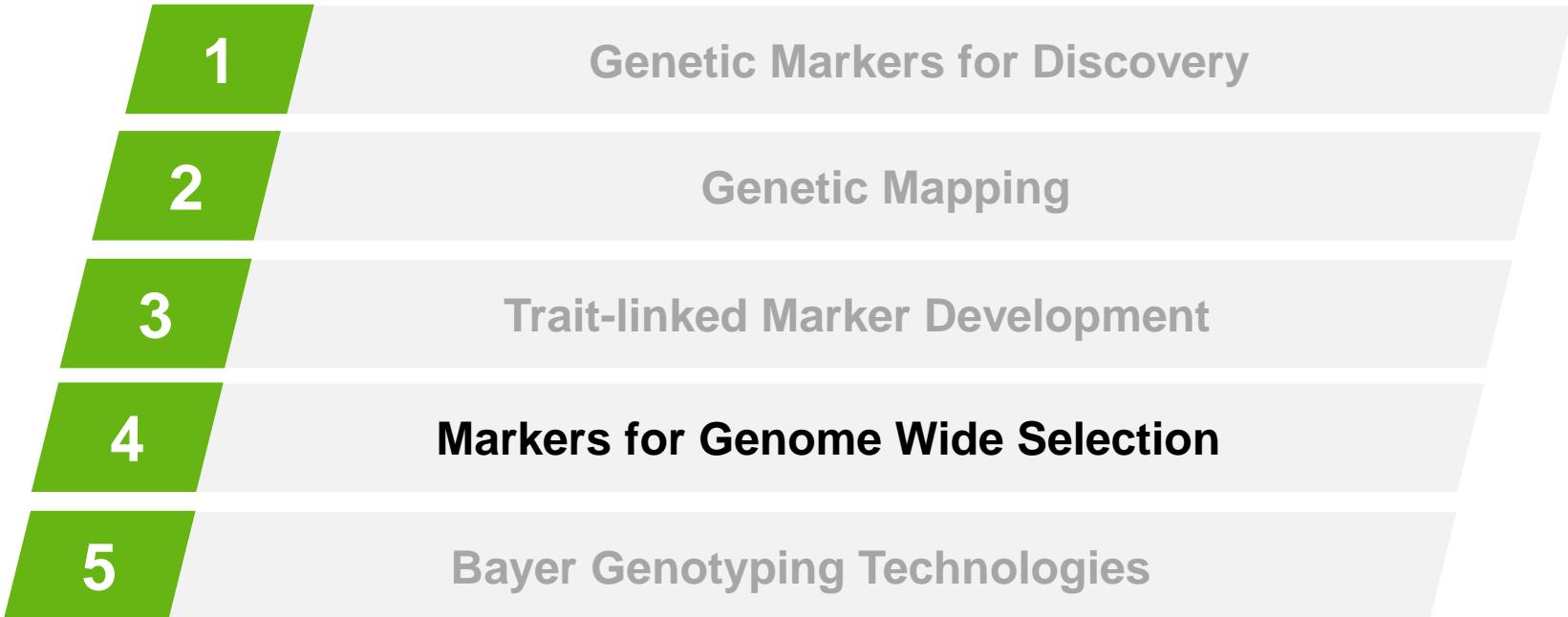
<https://doi.org/10.3389/fpls.2018.00885>

Trait markers

- // Located in specific genomic areas
- // Linked to the variation of a given trait
- // Can be used to follow said trait (MAS)



Bayer Russia Molecular Marker Training : Genetic Markers and Technologies





Agenda

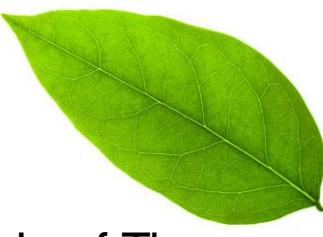
- // Overview
- // Selection of candidate markers
- // Designability screening
- // Marker Selection
- // Marker/Sample Validation
- // Deployment, Versioning and Maintenance

Marker data is generated by DNA sequencing and probe - based methods



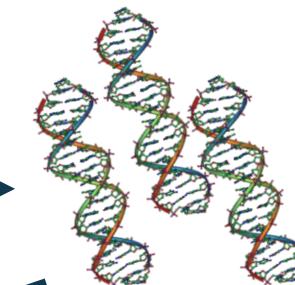
Seed
(Single or Bulk)

OR

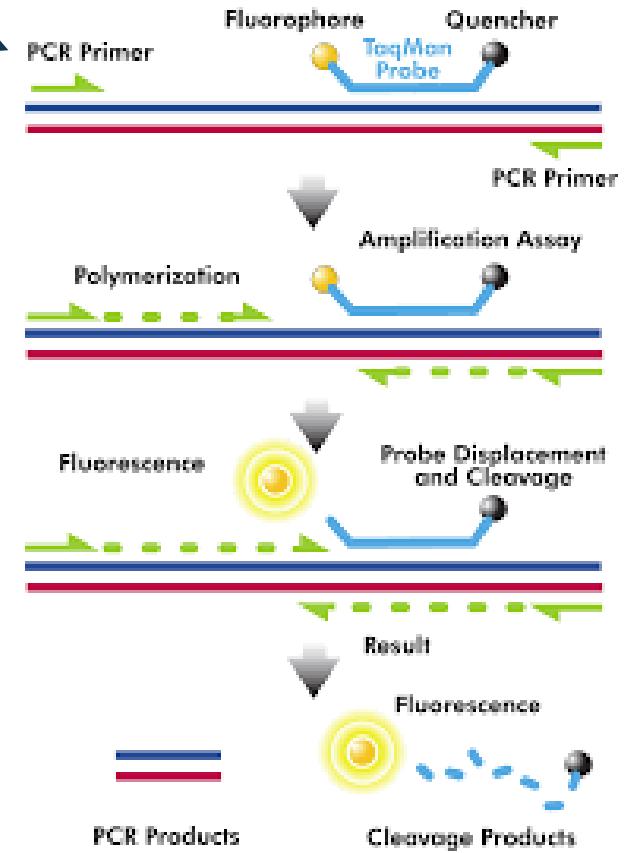


Leaf Tissue

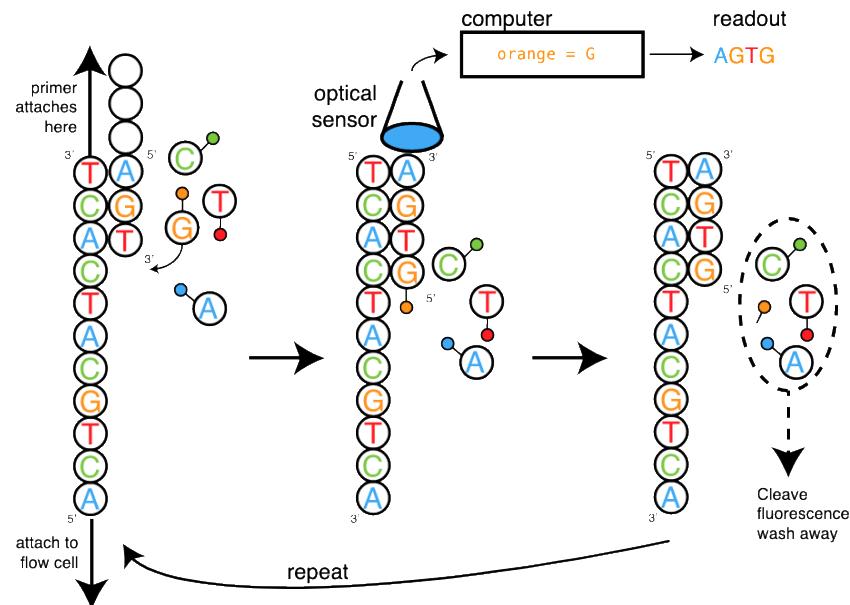
DNA
Extraction



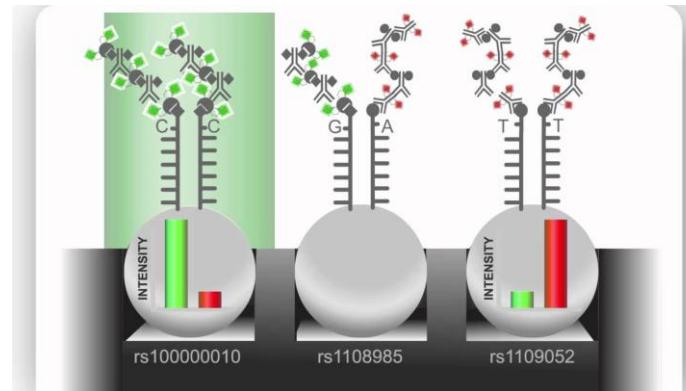
TaqMan Assay



DNA Sequencing (WGS, GBS, etc)

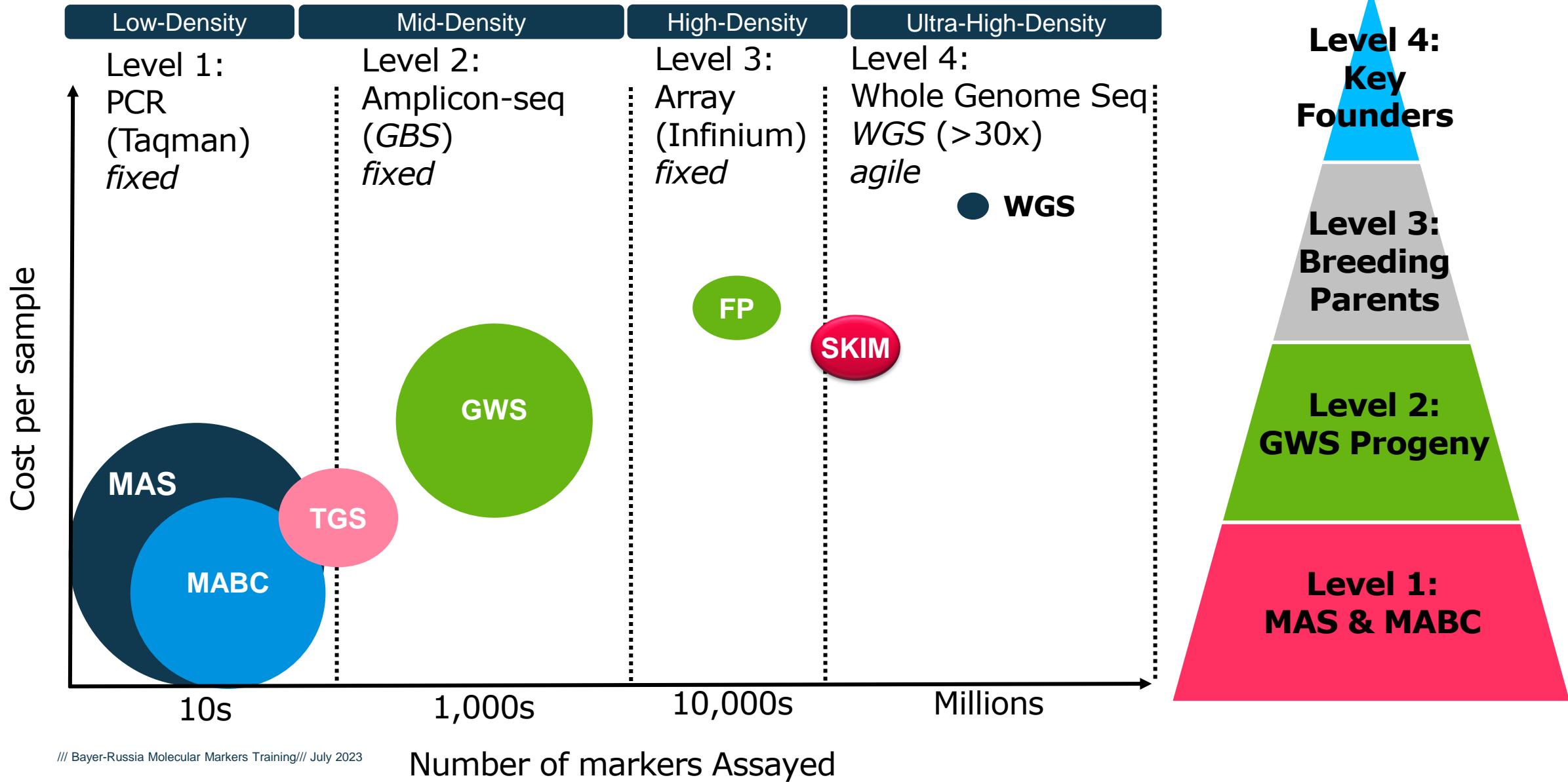


Infinium Assay (Fingerprinting)

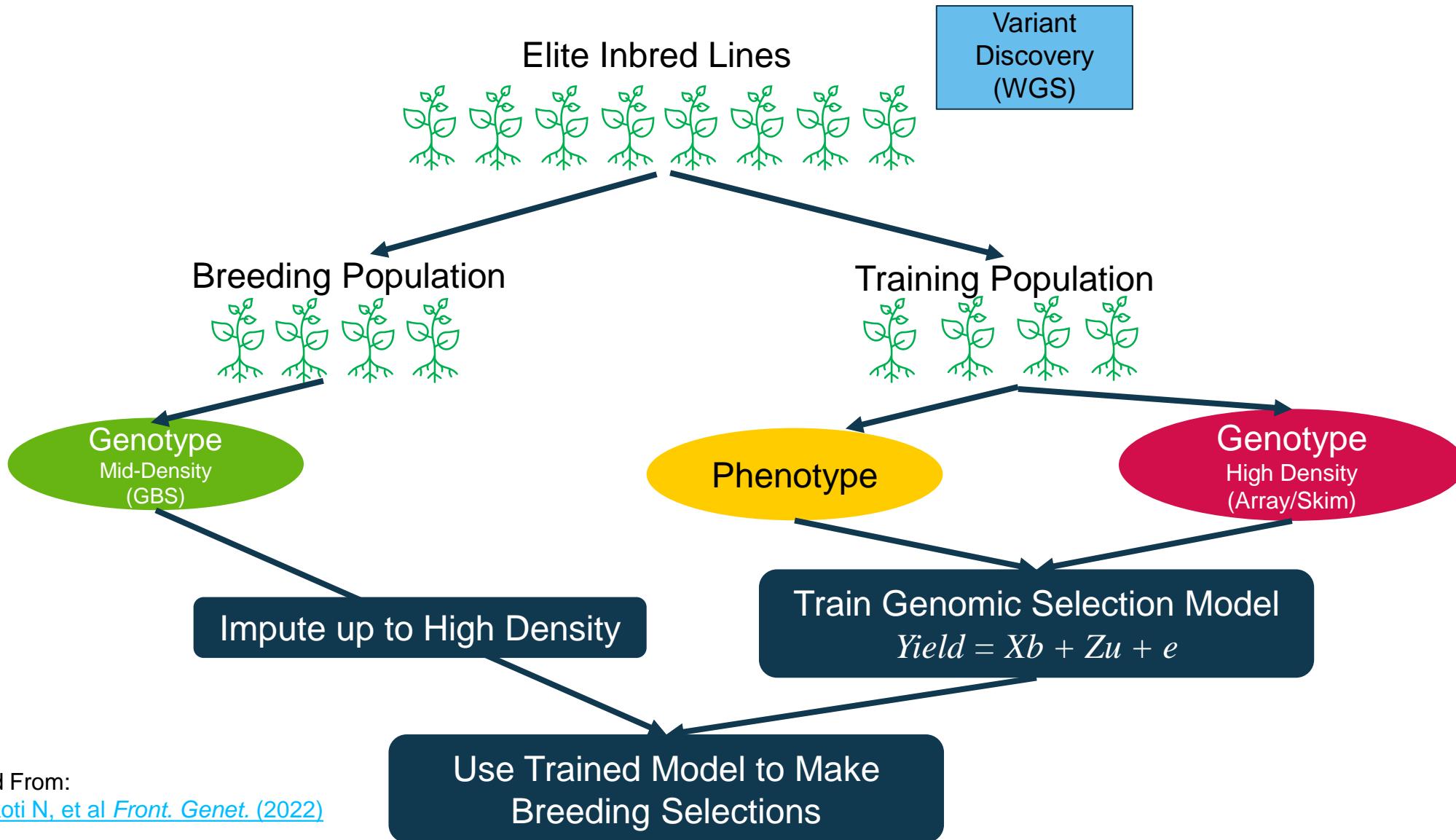


AA AT TT

Production genotyping technologies at Bayer Crop Science - Precision Genomics



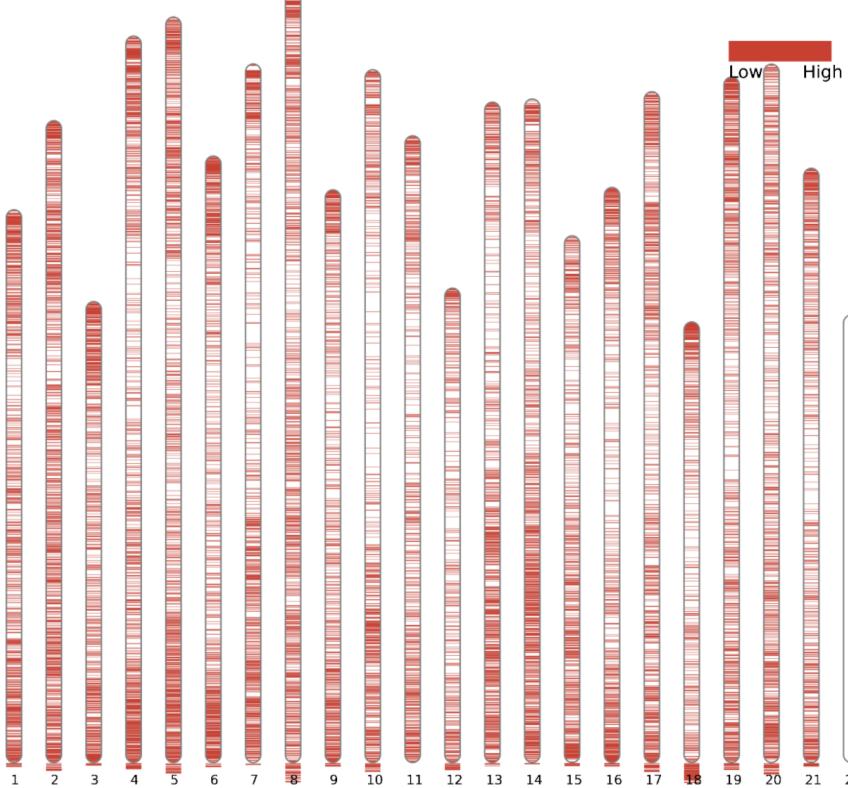
Genomic Selection



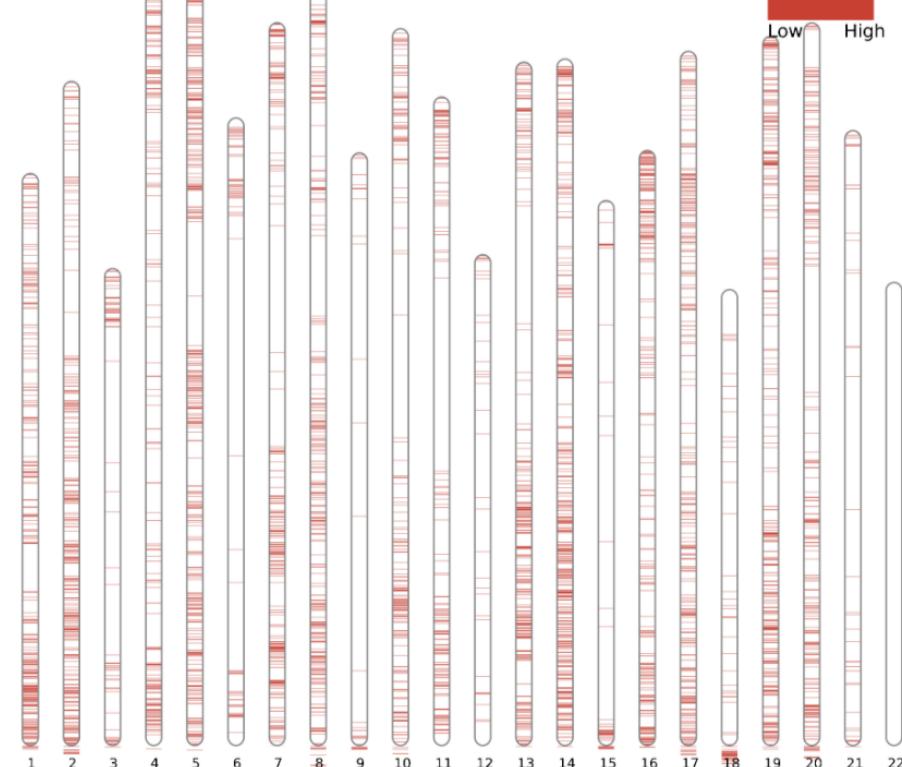
Adapted From:
[Buhdakoti N, et al *Front. Genet.* \(2022\)](#)

Objective: Develop reduced-representation marker set(s) that covers as many haplotypes as possible, constrained by per-sample assay cost

Full Set of Markers
(\$\$\$\$)

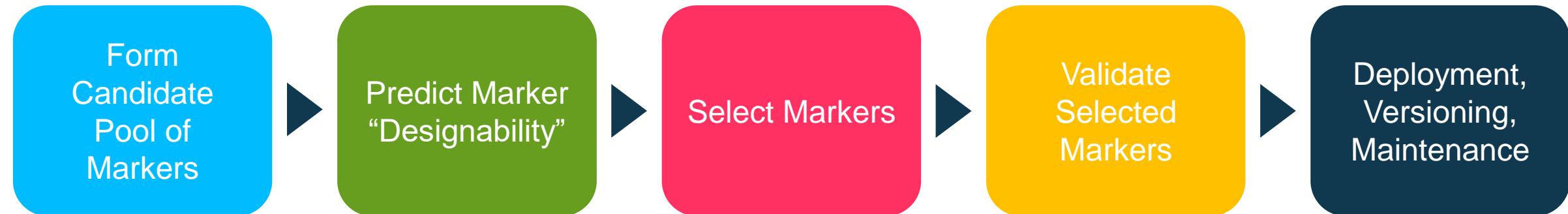


Reduced-
Representation Set
(\$)



- Select Markers by
- Parental Haplotypes
 - Assay performance
 - Required/Must-in Status

How to Build a Marker Set



- Collect all marker sequences and metadata
- Define germplasm to test
- Correct inconsistencies in alleles
- Correct genetic map

- Select markers to screen
- Determine genotyping platform
- Send markers to vendor

- Select markers to validate
- Manufacture assays/probes

- Measure success metrics compared to legacy data
- Drop low-performing markers
- Backfill must-in markers

- Put marker set into production
- Update production pipelines
 - Lab pipeline
 - Analysis pipeline

Density

TaqMan



Genotyping Platforms
Genotyping By Sequencing



Thermo Array

Fingerprinting



Low coverage Sequencing



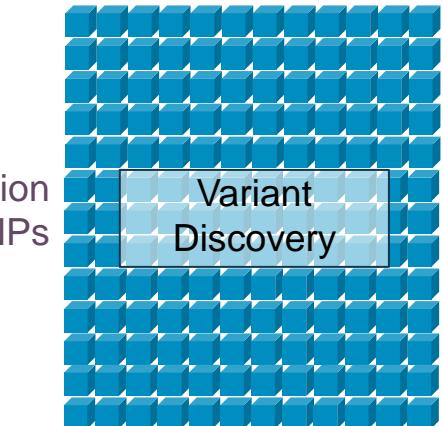
Illumina Array

Designing Markers for Genomic Selection

Discovery: Whole Genome Sequencing

- Discovered with >30x Whole Genome Sequencing
- Purpose: Deep genetic characterization of elite lines
- Cost: \$\$\$\$

100-1000 Individuals



~30 Million
SNPs

20k-60k Individuals



10,000-90,000
Markers

60k- >100k Individuals



1,000-2,000
Markers

Critical Trait/QC Markers
(TaqMan)

A/G **A/T**
G/T **C/T**
C/G

Filter on

- GATK Variant Quality Score
- Informativeness
- Simulated Crosses
- Genome Coverage

Filter on

- Fingerprint Performance
- Mappability to Reference Genome
- Informativeness
- Simulated Crosses
- Genome Coverage

Production: Genotyping Microarray

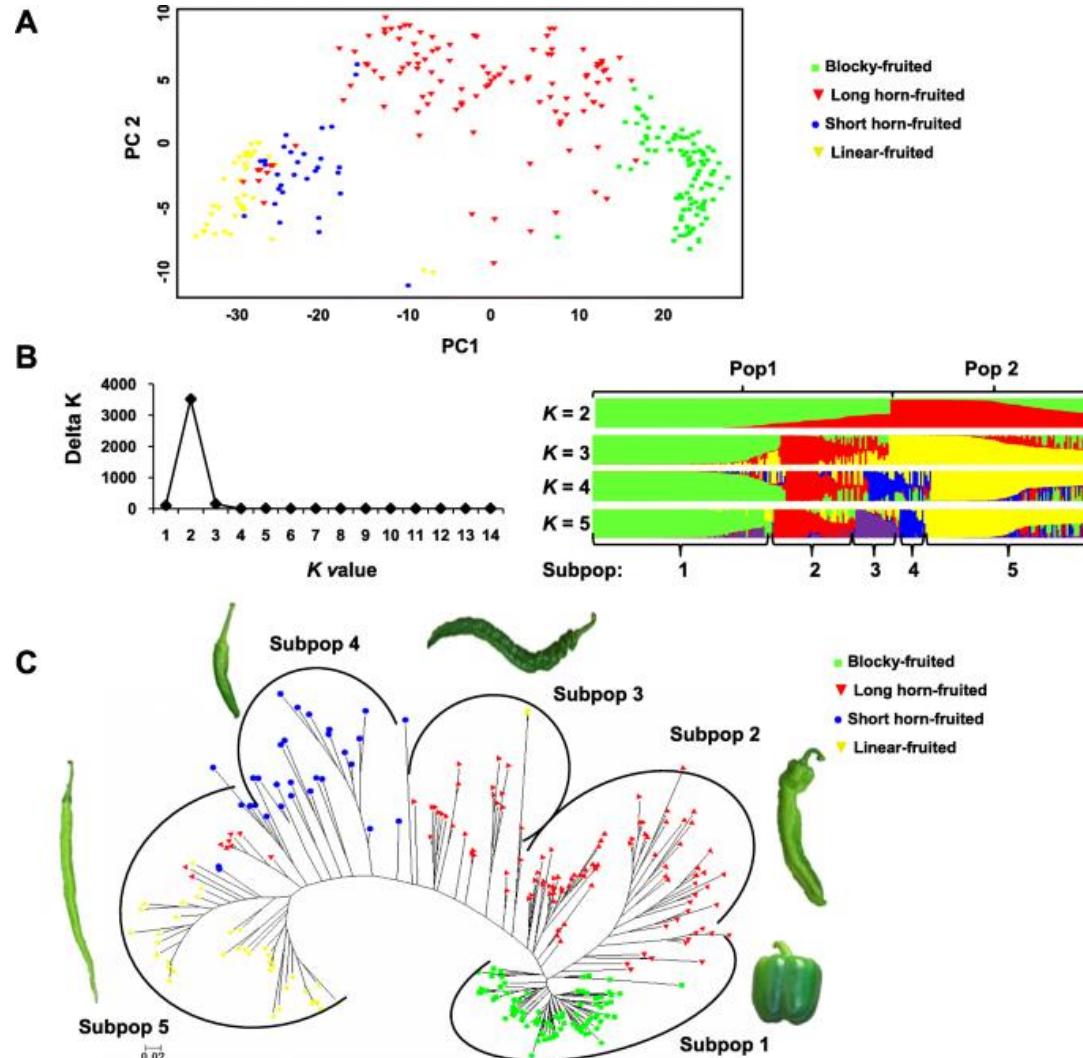
- Genotyped on a microarray or low coverage-seq
- Purpose: Deep genetic characterization *with high-quality markers*
- Cost: \$\$

Production: Genotyping by Sequencing

- Multiplexed amplicon sequencing
- Purpose: Genomic Selection
- Cost: \$



Developing a Diversity Panel – Pepper Example

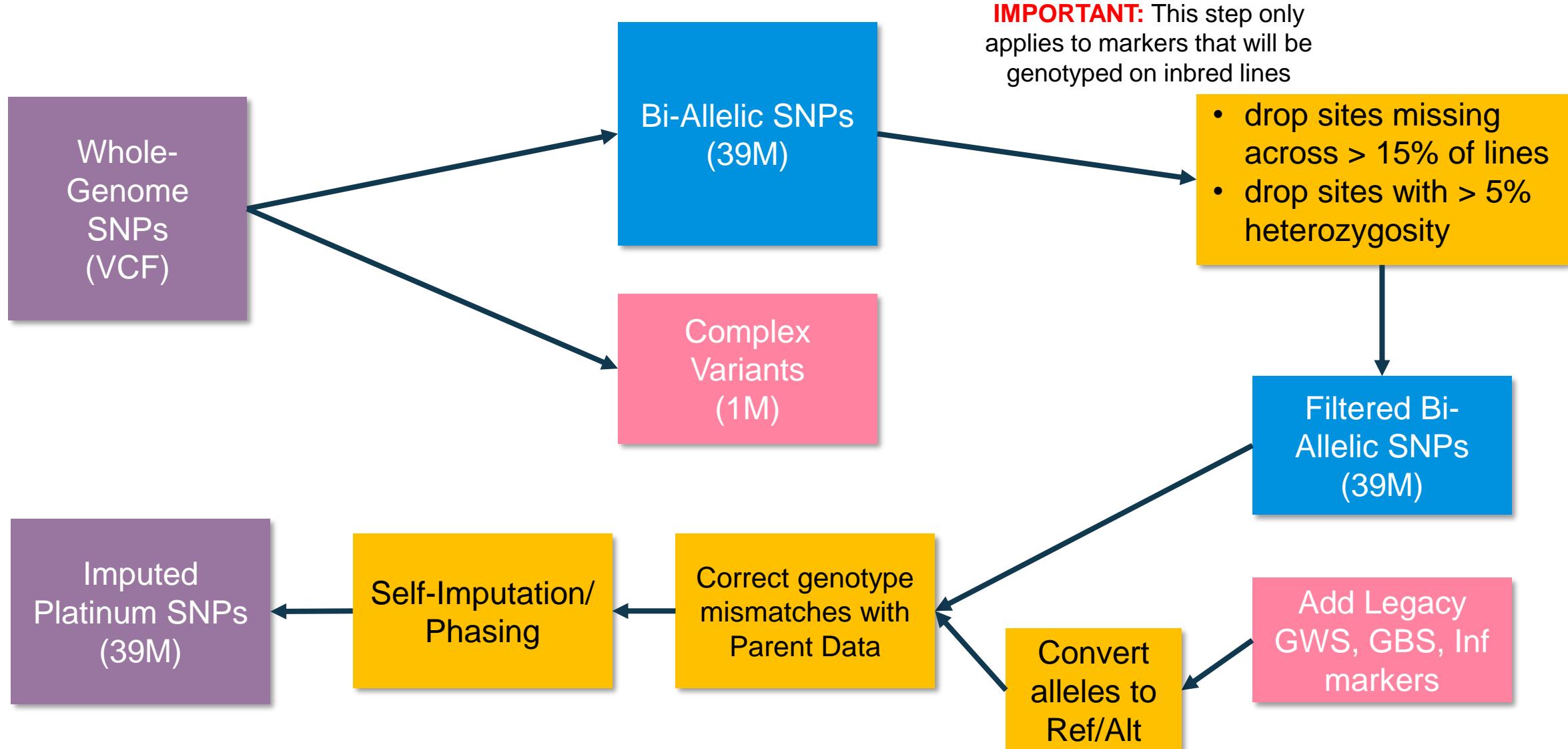


Du, H., Yang, J., Chen, B. et BMC Plant Biol 19, 578 (2019).

- When selecting a marker validation diversity panel select markers that can distinguish between:
 - Major sub-populations
 - Major phylogenetic groups
 - Major phenotypic groups



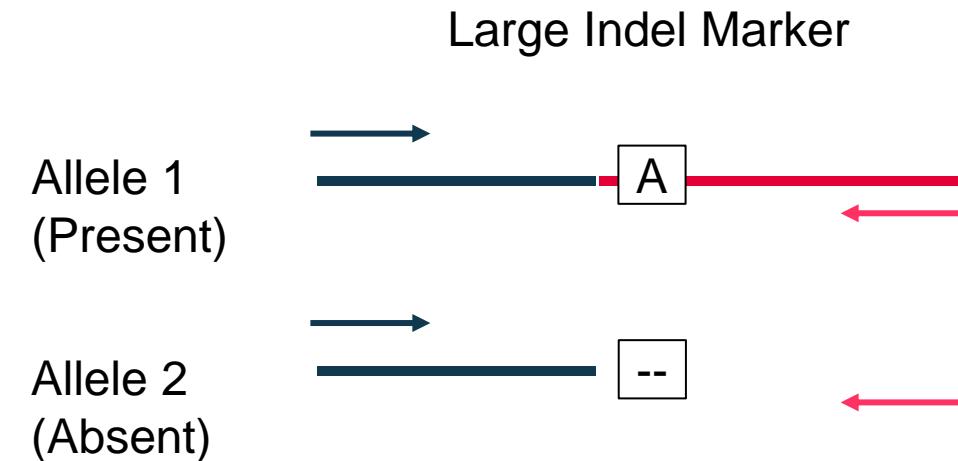
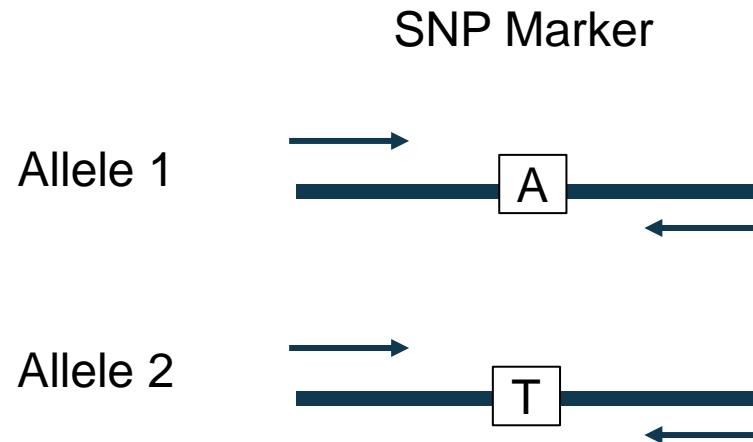
Selecting Reliable Markers from Whole-Genome SNPs





Required/Must-in Markers

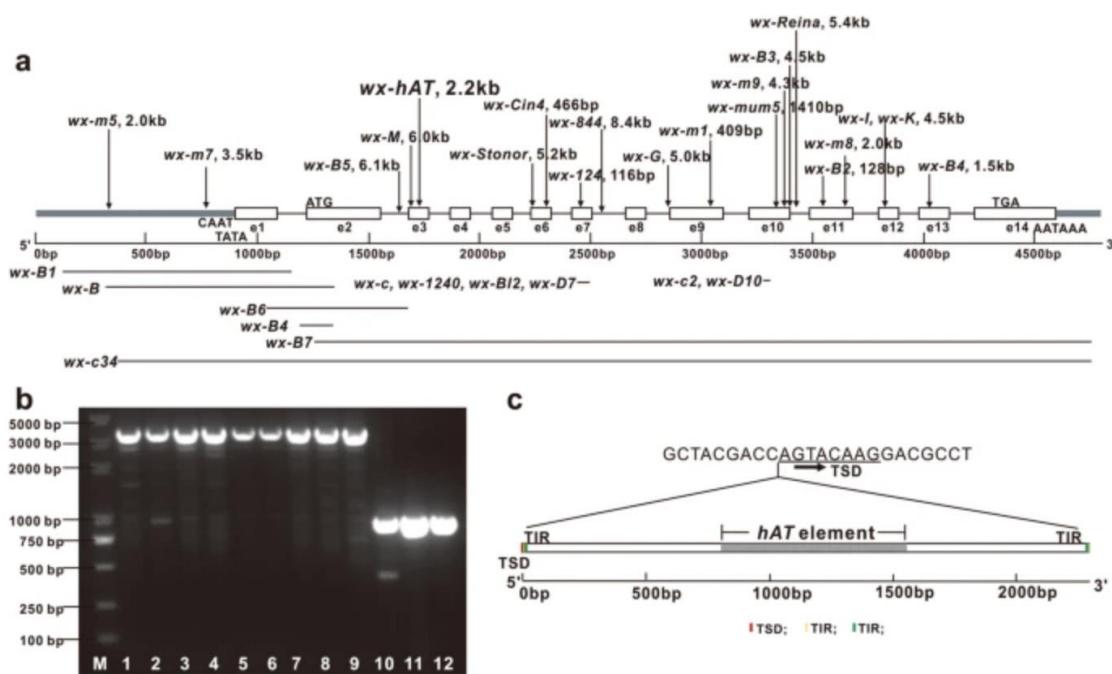
- These are markers that are specifically requested by the breeders
 - Also contains supply chain QA/QC markers (10-30 markers)
 - Most of these markers are designed to track markers associated with a trait, and some of the designs can be somewhat complex



Trait markers can have complex/indel designs

Examples

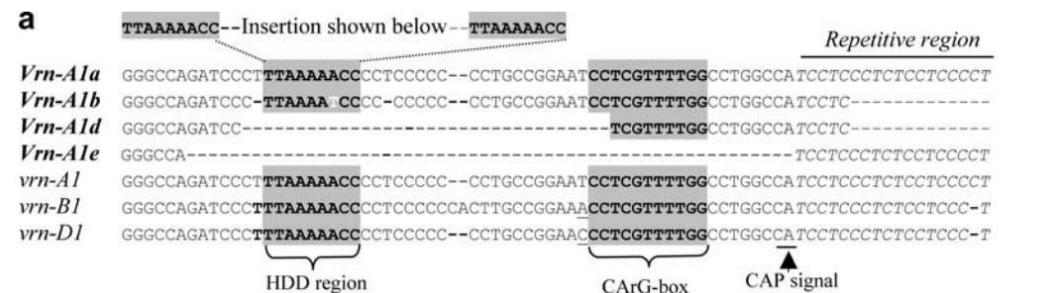
Waxy Locus in Maize: 2kb *hAT* Element insertion



M. Luo, *Nature Sci Reports* (2020)

7 Vernalization Genes in Wheat

(the fact that this is in a known repetitive region adds to the pain fun)

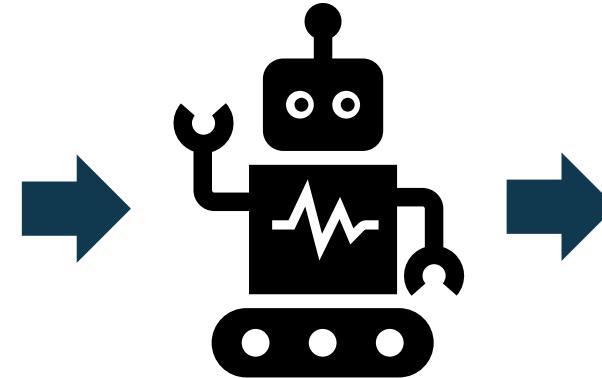


Predicting Designability based on Sequence

Marker Flanking
Sequence

Marker	Sequence
M ₁	ATACAGA [C/G] CTGAACA
M ₂	GTCCCGT [A/T] CAGGGGT
...	...
M _N	GTGGGAT [A/C] CAAGGCC

Vendor Software



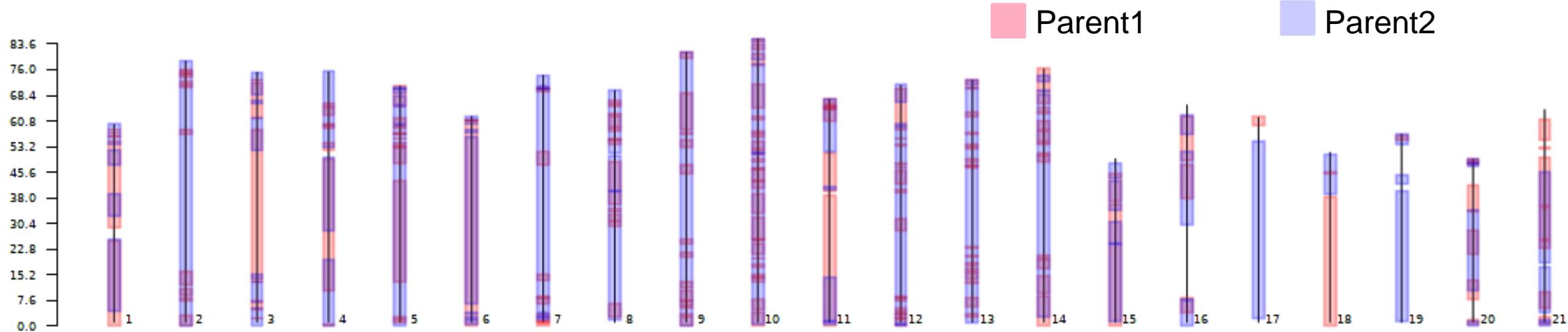
Scores and Designs

Marker	Score	Primers/Probes
M ₁	100%	ATACAGA
M ₂	50%	CAGGGGT
...
M _N	90%	CAAGGCC

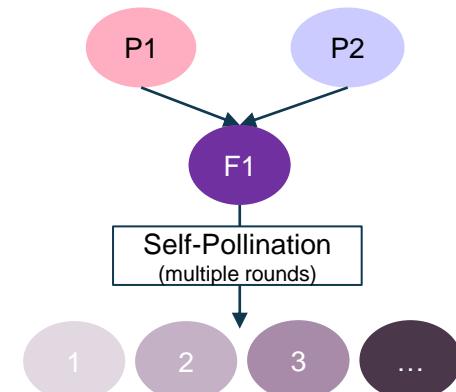
- Marker Flanking Sequence can come from:
 - A reference genome
 - Targeted sequencing
 - Previous design from a different platform

- Marker Flanking Sequences are scored by:
 - Repetitiveness
 - Probability of cross-hybridization
 - Additional sequence features
- Include markers with favorable designability scores
- Remove markers with unfavorable scores

Selecting markers using parental information



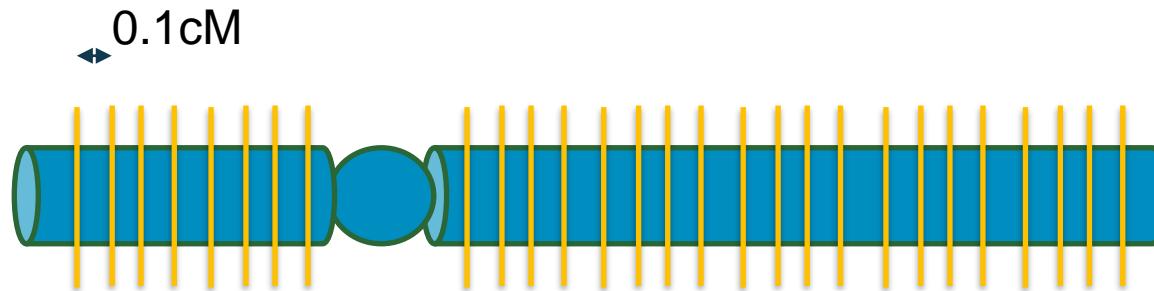
1. Using genotype data on the parents, assign marker alleles to each parent
2. Determine the minimum number of markers that will cover the maximum number of haplotypes across all pedigrees



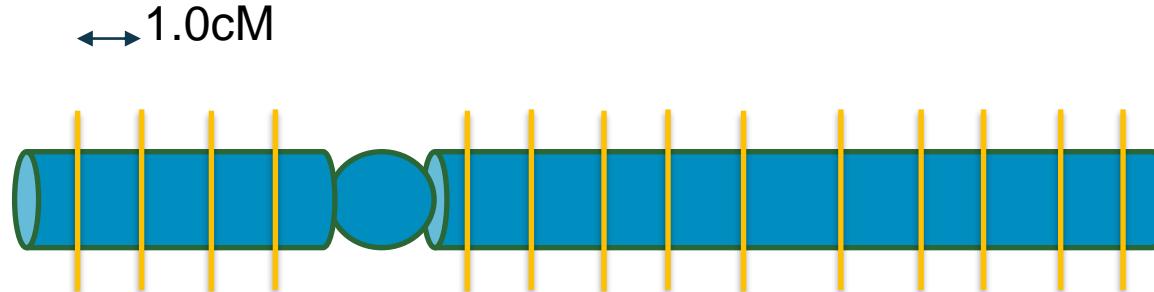
Selecting the minimum number of markers to maximize genetic map coverage

Select
Markers

All Possible Markers



High-Density



Medium-Density



- The exact amount of spacing is decided by breeding stakeholders and optimized by genome size and recombination rate
- If gaps appear, they can be filled by gap-filling algorithms

Selecting markers using assay performance metrics

	Plant1	Plant2	Plant3	Plant4	Plant5	Plant6
Marker1	AA	AA	AA	BB	BB	BB
Marker2	AA	BB	BB	BB	BB	BB
Marker3	BB	Missing	BB	Missing	Missing	AA
Marker4	BB	BB	BB	BB	BB	BB
Marker5	AB	AB	AB	AB	AB	AB

- Low het
 - High MAF
 - No missing
- 
- Low MAF
- 
- High Missing
- 
- All homozygous BB
- 
- High heterozygosity
- 

Determining Genotyping Accuracy with Controls

Genotypes From
New Marker Set

	Plant1	Plant2
Marker1	AA	AA
Marker2	AA	AA
Marker3	BB	AA
Marker4	BB	BB
Marker5	AB	AB

Control Genotypes
(WGS, Array, or GBS)

	Plant1	Plant2
Marker1	AA	AA
Marker2	AA	BB
Marker3	BB	BB
Marker4	BB	BB
Marker5	AB	AB

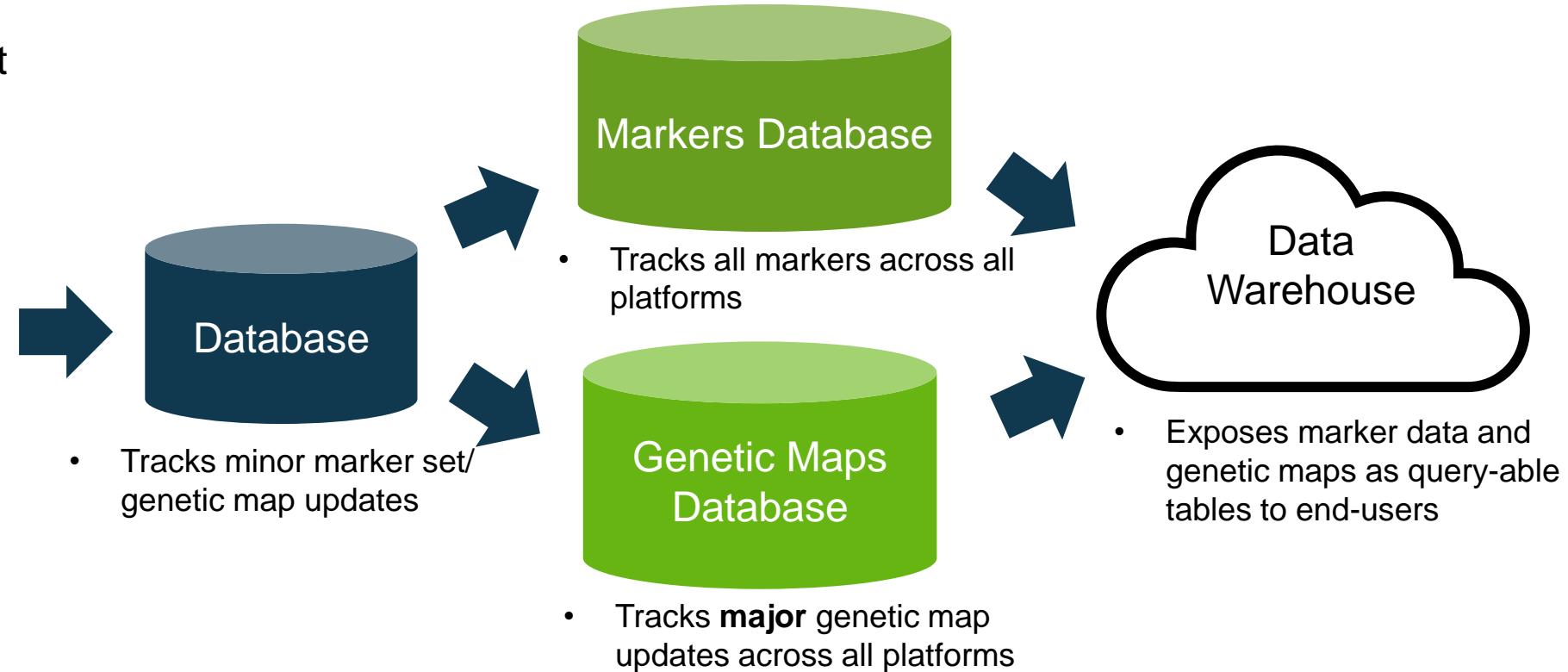
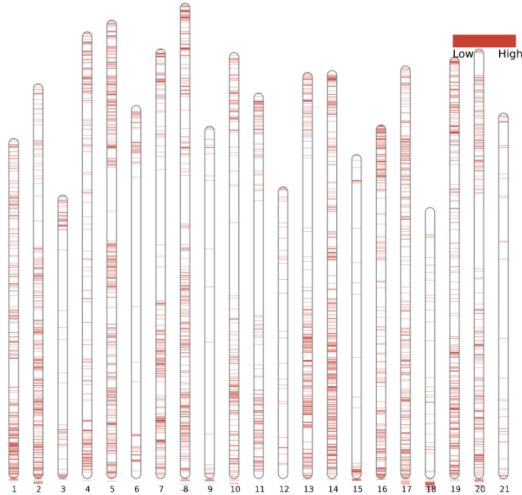
Concordance Matrix

	Plant1 Match?	Plant2 Match?	Marker Concordance
Marker1	1	1	100%
Marker2	1	0	50%
Marker3	1	0	50%
Marker4	1	1	100%
Marker5	1	1	100%
Sample Concordance	100%	60%	



Marker set versioning and deployment

New, Validated Marker Set



- // Both marker sets and genetic maps are given unique Identifiers, which are updated with every new version
- // New markers require an additional registration step in database, where they are assigned new IDs



Q&A Discussion



Bayer Russia Molecular Marker Training : Genetic Markers and Technologies

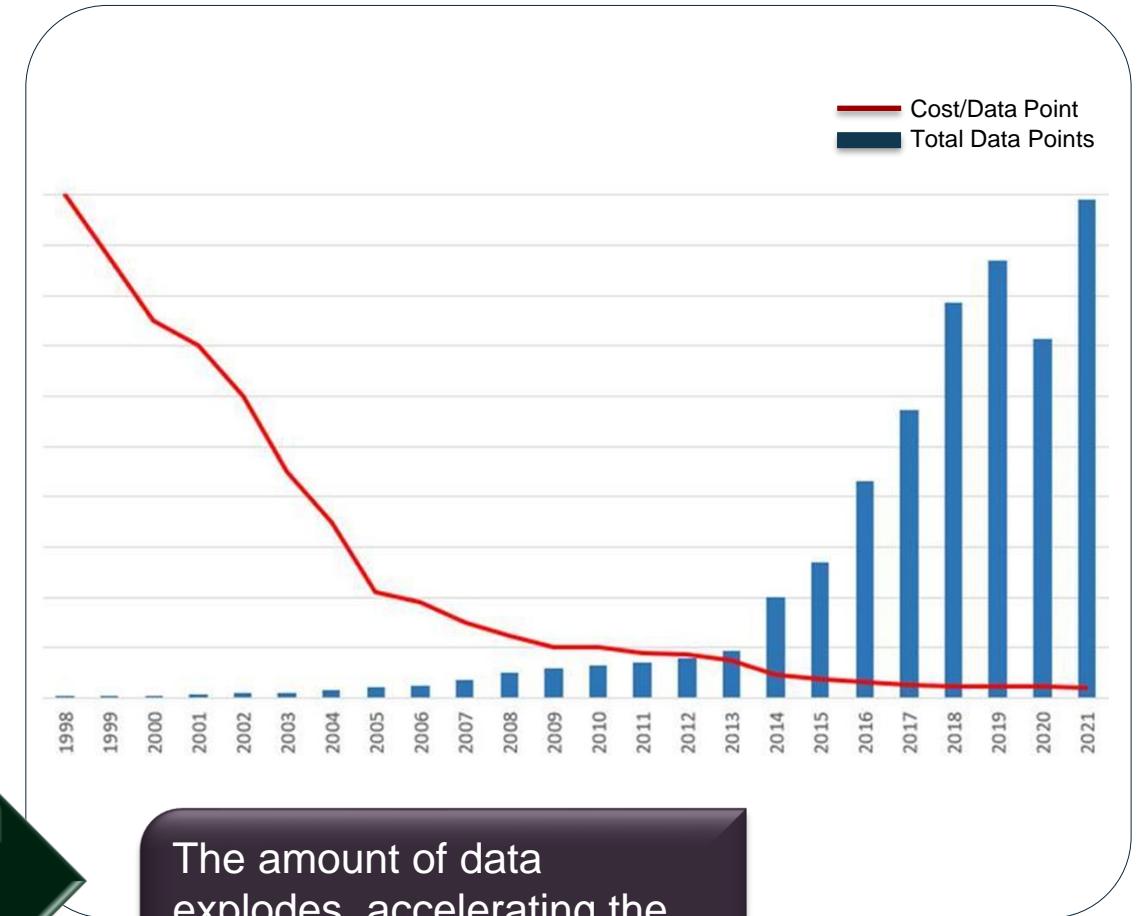


Pipelines need to be scalable to find the rare, high performers.

Scaling is enabled by increased capacity and decreased cost



Robotics Automation



The amount of data explodes, accelerating the ability to realize step-change improvements

Genotyping Technology Comparison

~10k to ~800k
markers

High density markers

HD-Array by Affymetrics or Infinum. Selected markers which are best fit for the breeding pool as [Fingerprinting \(FP\)](#) platform. Used on parental / elite germplasm in later breeding pipeline.

Medium density markers

[Genotyping by sequencing \(GBS\)](#), pre-designed marker set based on FP. Used on progeny of the breeding populations and imputed to FP level for genomic prediction. Can also be used in TI / Discovery.

Low density markers

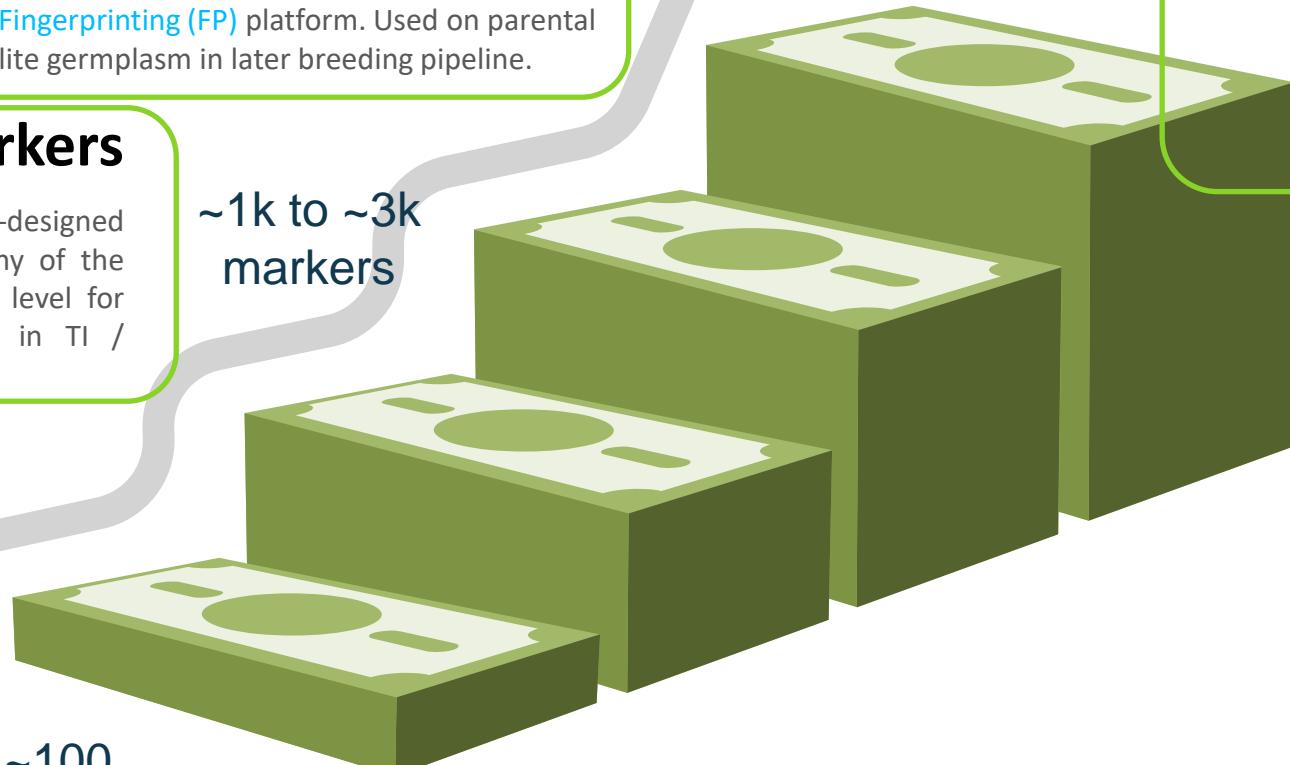
Taqman / KASP markers. Used to track traits controlled by limited number of genes.

1 to ~100
markers

~1k to ~3k
markers

Genomic sequence

Elite / core / founder germplasm to identify necessary polymorphisms to strengthen the genomics platform and additional coverage to support traits discovery.





Low density genotyping

Equipment

Thermocycler or
Hydrocycler

Plate Reader

OR

qPCR instruments

Talent

Assay Design

Experimental
Design

Bench Scientists

Assay Scoring

Scaling

Liquid handling
instrumentation
(Tecan, Biomek, Firefly)

Robotics for plate
movement

Consider shift
work for quicker
turn-around times



Medium density genotyping

Equipment

Thermocycler or
Hydrocycler

Short-read
sequencer

Talent

Marker Selection

Experimental
Design

Bench Scientists

Computational
Biology

Scaling

Liquid handling
instrumentation
(Tecan, Biomek, Firefly,
etc)

Robotics for plate
movement

Consider shift
work for quicker
turn-around times



High density genotyping and genomic/transcriptomic sequencing

Equipment

Thermocycler or
Hydrocycler

Appropriate array
preparation and
reader
instrumentation

Short-read
sequencer
(for low coverage seq)

Talent

Marker Selection

Experimental
Design

Bench Scientists

Computational
Biology

Scaling

Scalable high
quality DNA
preparation

Liquid handling
instrumentation
(Tecan, Biomek, Firefly)

Genotyping does not have to be developed in-house as several global companies offer genotyping services



Examples for genotyping/sequencing companies

- ❖ Eurofins
- ❖ LGC
- ❖ SGS Institut Fresenius
- ❖ Gencove
- ❖ Paragon Genomics

Considerations when leveraging a 3rd party

- ❖ When scoping the project, important to have key information available
 - ❖ Expected number of samples/ batch
 - ❖ Single project or recurring submissions
 - ❖ Expected turn around time
 - ❖ What development or customization will be needed? Will you work together or do you need 3rd party to develop?
- ❖ Data flow from 3rd party back to breeding program
 - ❖ Is analysis required by 3rd party or will raw data be supplied?
 - ❖ Will standard data report be sufficient?
 - ❖ How will information be exchanged?

*All company names are used for illustrative purposes only. Inclusion does not imply Bayer Crop Science endorsement.



Q&A Discussion





DISCLAIMER

THE INFORMATION CONTAINED HEREIN IS EXPERIMENTAL IN NATURE AND IS PROVIDED "AS IS". BAYER MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO THE MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE INFORMATION WILL NOT INFRINGE ANY THIRD-PARTY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS.