3. Genomic variation

- 3.1 Genomic variations
- 3.2 SNP calling
- 3.3 Detect structural variations
- 3.4 Application of genomic variations

3.1 Genomic variations

- Types of genomic variations
- Genotyping approaches

Genomic DNA Variation

= variations between the genome of different individuals

Single Nucleotide Polymorphisms (SNPs) = substitutions, short indels (one to a few nucleotides)

Micro- and mini-satellite expansion and contraction (typically less than 100 bp variation)

Transposable Elements insertion/excision (ranging from ~100 bp to less than 10 kb)

Segmental Duplications = Low copy repeats (LCRs) (>1 kb- 3 Mb with similarity >90%) -- include copy number variants (CNVs)

Large chromosomal rearrangements: Mb-range duplication, insertion, deletion, inversion, translocation (microscopic structural variation)

Changes in chromosome numbers = aneuploidy (typically deleterious) (microscopic structural variation)

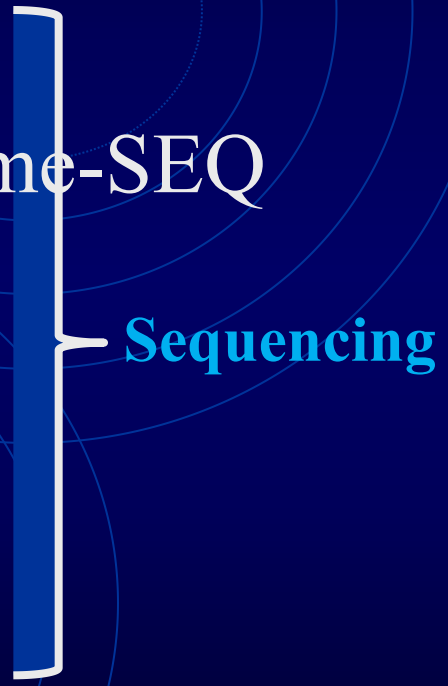
- **Types of genomic variation**
  - SNP
  - Indel: insertion/deletion
  - Structural variation (SV)
  - CNV: copy number variation; PAV
  - Inversion
  - Translocation: intra-/inter-chromosomal
  - Duplication
  - Rearrangement
  - Methylation
  - …

Two main types

- SNP (also small indel)
- Structural variation
  - It consists of many kinds of variation in the genome of one species, and usually includes microscopic and submicroscopic types, such as deletions, duplications, copy-number variants, insertions, inversions and translocations. Typically a structure variation affects a sequence length about 1Kb to 3Mb, which is larger than SNPs.

Genotyping approaches

- Genome re-sequencing
- Targeted sequence capture: Exome-SEQ
- RAD-SEQ
- RNA-SEQ
- Environmental (mixed) samples



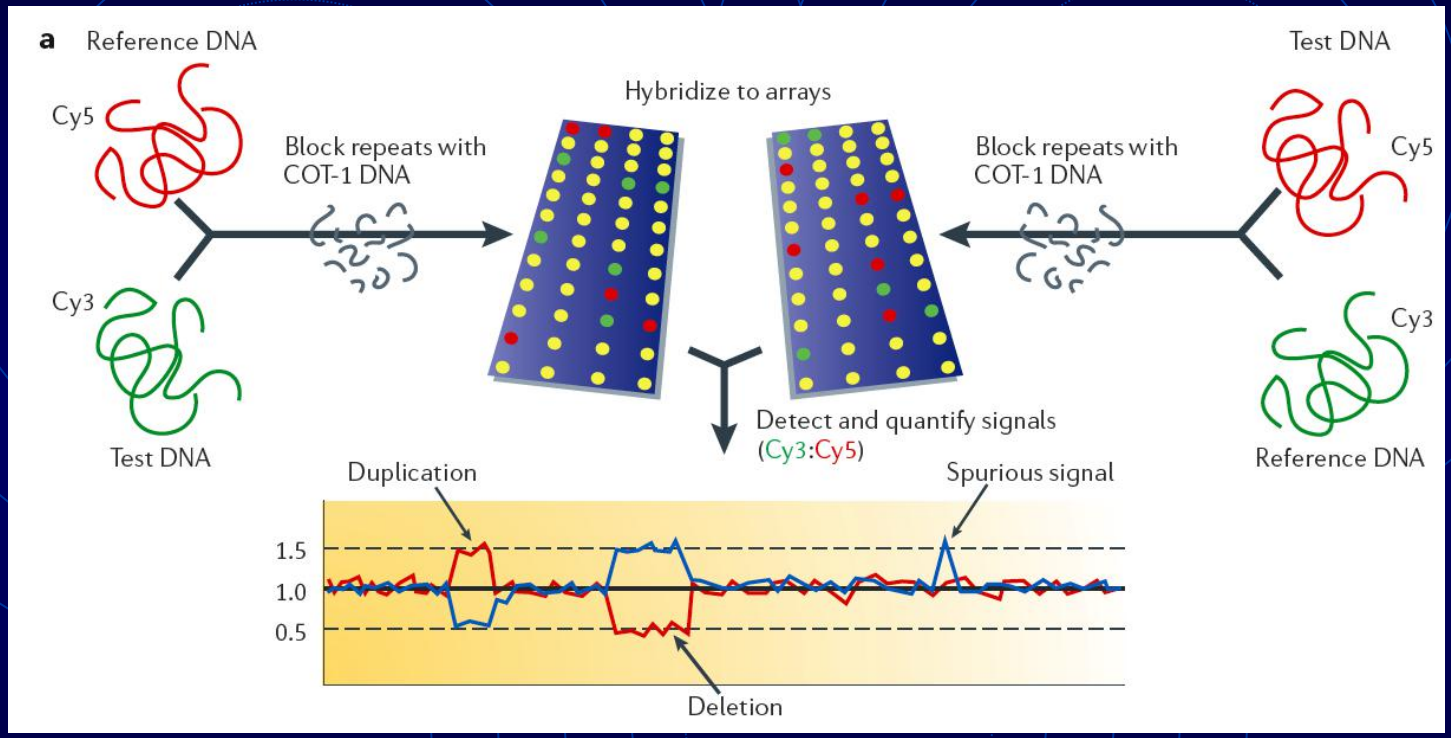
- Cytogenetic detection

Cytogenetic detection of structural genomic variation

- **FISH**
- **Array-based Comparative Genome Hybridization (CGH)**

CNVs detected by fiber FISH

Array-CGH technology



- Using this method, copy number changes at a level of 5-10 kilobases of DNA sequences can be detected. Today even high-resolution CGH (HR-CGH) arrays are accurate to detect structural variations (SV) at resolution of 200 bp (Urban et al. 2006).

3.2 SNP calling

- About SNP
- SNP calling

Single Nucleotide Polymorphism

- A Single Nucleotide Polymorphisms (SNP), is a genetic variation when a single nucleotide (i.e., A, T, C, or G) is altered and kept through heredity.

- SNP: Tag SNP
- Mutation
- Haplotype: haplotype block
- Genotype

SNPs are very common

- SNPs are very common in the human population.
- Between any two people, there is an average of one SNP every 1000 bases.
- Most of these have no phenotypic effect
  - only <1% of all human SNPs impact protein function (non-coding regions)
  - Selection against mis-sense mutations

SNP标记的优势

- 基因组上数量多密度大
  - **Maize：1 per 48 bp (non-coding)；1 per 131 bp (coding)**
  - **Soybean：1 per 294 bp**
  - **Arabidopsis：1 per 3,300 bp**
  - **Cotton：1 per 100 bp**
  - **Human: 1 per 1000bp**
- 可以通过简单但高通量方法检测，如测序和芯片检测等，效率高

SNP检测技术已被大量开发：

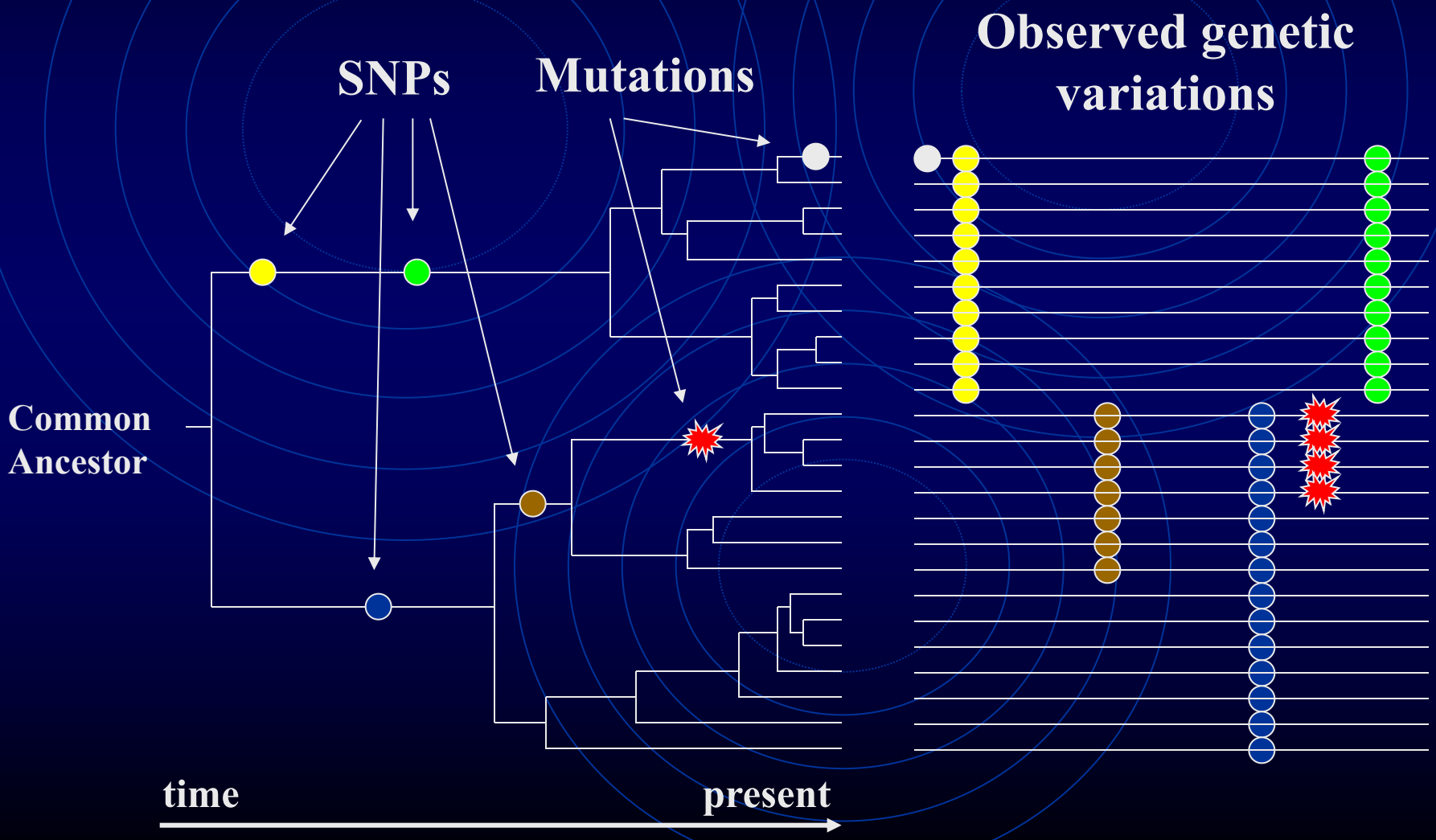
- MALTI-TOF MS: Sequenom, San Diego, CA,
- USATaqMan: Applied Biosystems, Foster City, CA,
- USAInvader: Third Wave Technologies, Madison, WI, USA
- SNPStream: Beckman-Coulter, Fullerton , CA, USA
- Pyroequencing: Uppsala, Sweden
- Illumina: La Jolla, CA, USA
- Biotrove OpenArray: Woburn, MA, USA
- Array Tape System: Alexandria, MN, USA

- Alleles of function related genes: Genetic markers that are linked to every gene
- Population diversity & history
- Genetic Association studies in populations
- Molecular assistant breeding …

SNP and mutation

- **SNP: Single DNA base variation found >1%**
- **Mutation: Single DNA base variation found <1%**

Mutations and SNPs



Major SNP

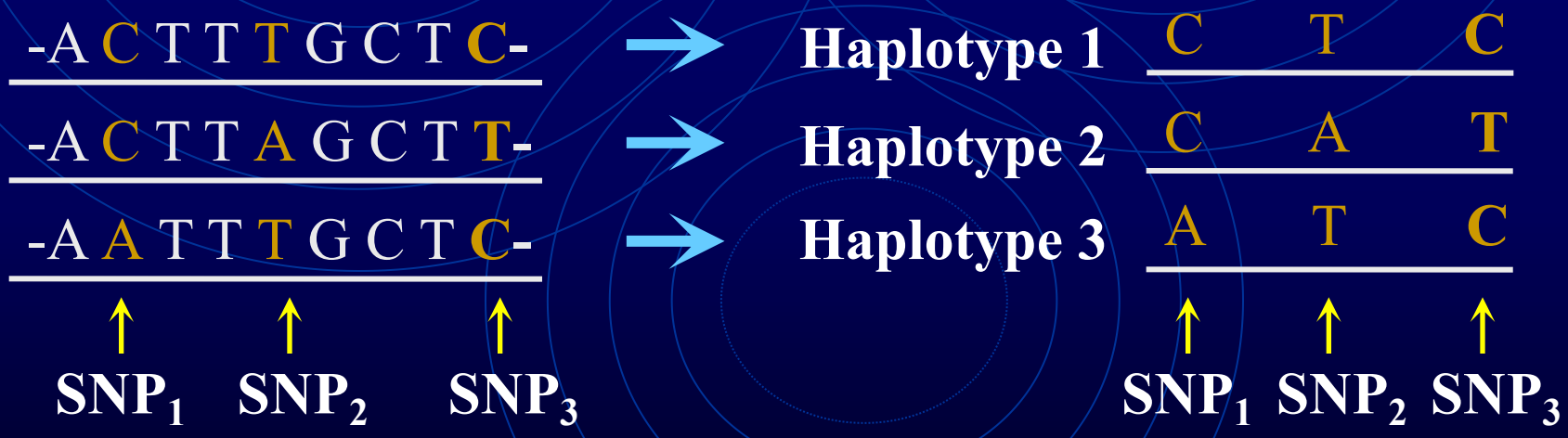
- A SNP is usually assumed to be a binary variable.
  - The probability of repeat mutation at the same SNP locus is quite small.
  - The tri-allele cases are usually considered to be the effect of genotyping errors.
- The nucleotide on a SNP locus is called
  - a major allele (if allele frequency > 50%), or
  - a minor allele (if allele frequency < 50%).

94% → A C T T A G C T **T** ← **T: Major allele**

6% → A C T T A G C T **C** ← **C: Minor allele**

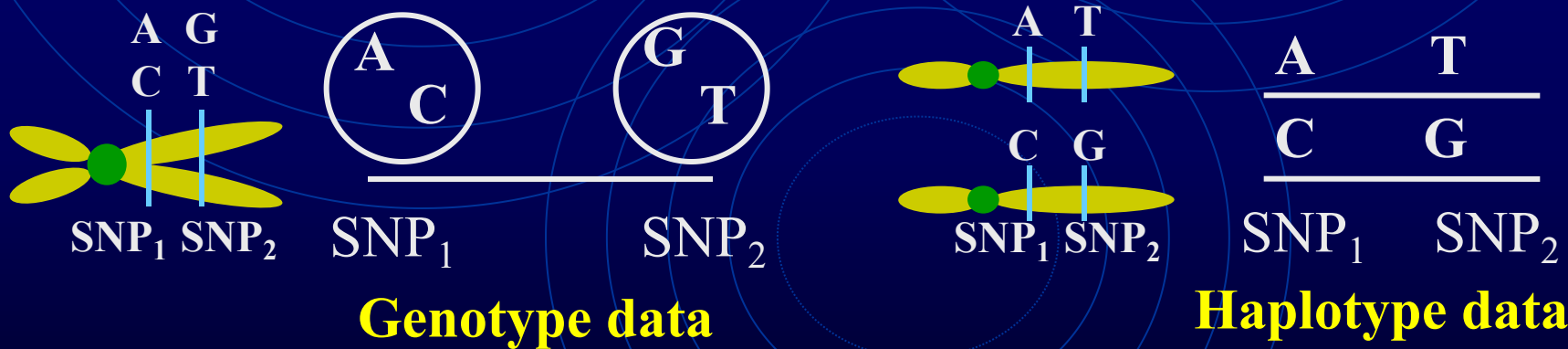
Haplotypes

- A haplotype stands for a set of linked SNPs on the same chromosome.



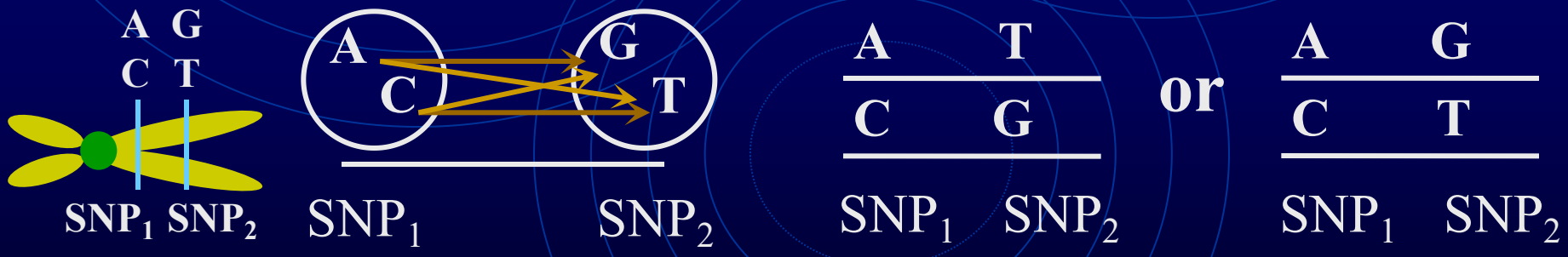
Genotypes

- In large sequencing projects, **genotypes** instead of haplotypes are collected due to cost consideration.
- Heterogenous genomic sites



Problems of Genotypes

- Genotypes only tell us the alleles at each SNP locus.
  - But we don't know the connection of alleles at different SNP loci.
  - There could be several possible haplotypes for the same genotype.



Genotype data

SNP calling

- With reference
  - Genome
  - transcripts
- Without reference

The reliability of short read alignment

- Repeats and sequencing errors
- Limited mismatch numbers within a region
- Always report a single alignment
- Fully utilize the mate-pair information of paired reads
- Produce a consensus genotype sequence from the alignment inferred from a statistical model

Finding higher quality SNPs

- Look at the number of reads covering the position with the SNP and discard those covered by three or fewer reads.
- Consensus quality is important, but SNP quality is more important. Discard a SNP with a quality score lower than 20.

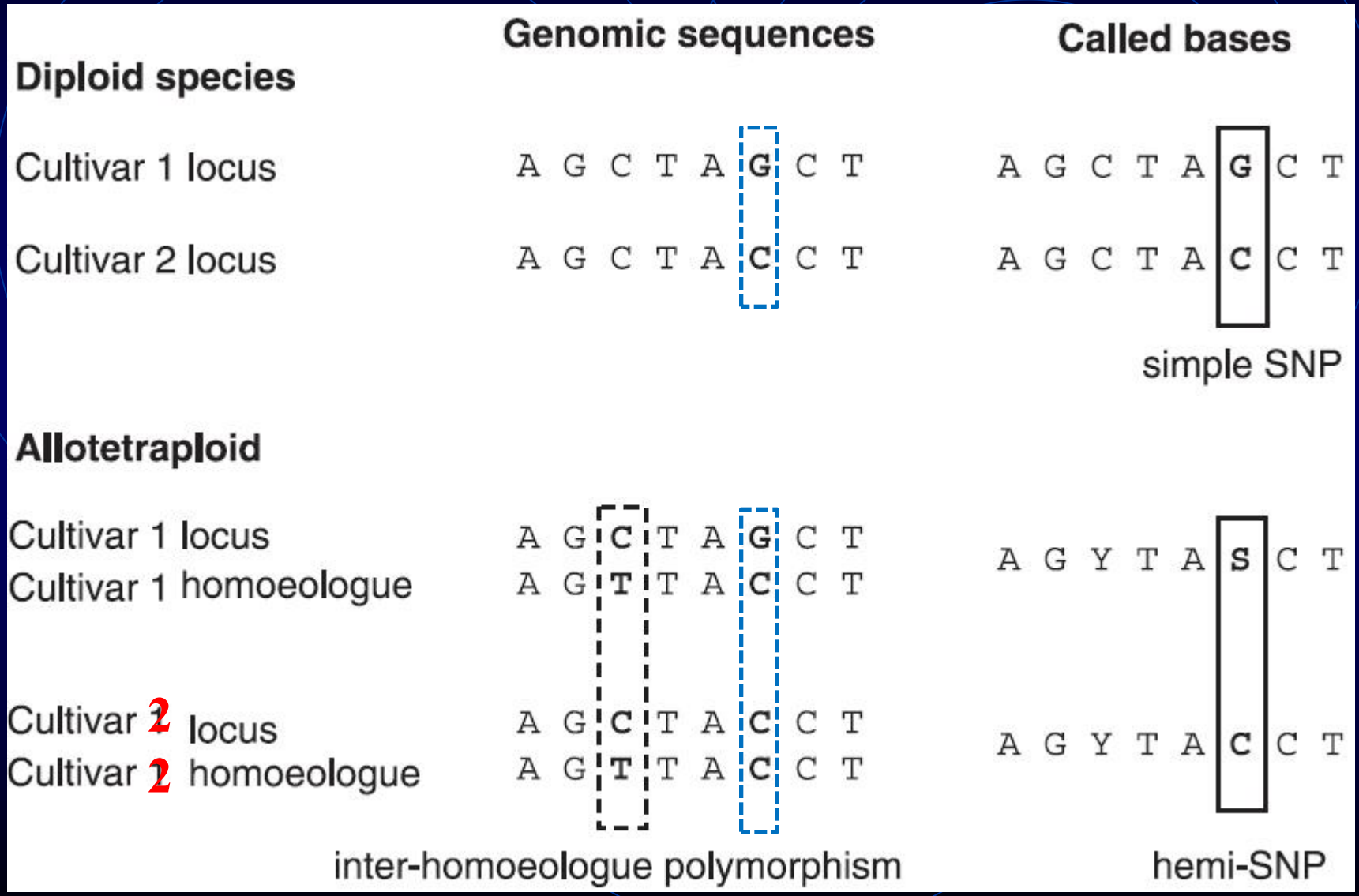
Challenges of mapping-based approaches

- Reference genome is not available
- Hemi-SNP
- Large size of genomic variation

Genotyping SNP when reference genome is not available

- SNP genotyping a genetic population/germplasm population…
- RAD-SEQ; RNA-SEQ; ect.

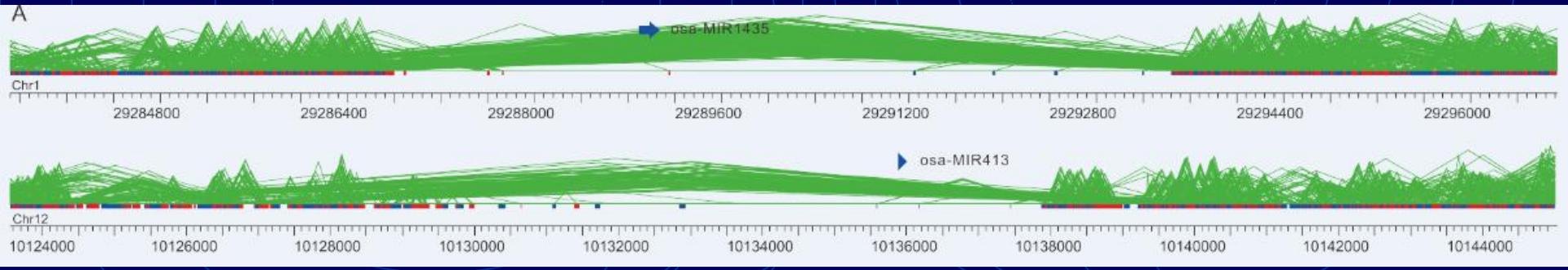
A complex case for SNP



3.3 Detect structural variations

- PE read-based
  - Deletion, insertion, inversion, translocation
- *de novo* assembly
- Read number-based
  - CNV

deletion



(Wang *et al*. 2012)

Insertion: T-DNA/ Tos17

No.	read ID	reference			T-DNA		results
		position		orientation	position	orientation	
25	FCC02UYACXX:1:1301:16055:36574#ACAGTGAT	Chr8	24761748	→	8109	→	An insertion
17	FCC02UYACXX:1:1106:15543:6427#ACAGTGAT	Chr8	24761768	→	8107	→	
23	FCC02UYACXX:1:1207:19712:37586#ACAGTGAT	Chr8	24761798	→	8096	→	
47	FCC02UYACXX:1:2106:5416:32126#ACAGTGAT	Chr8	24761809	→	8063	→	
27	FCC02UYACXX:1:1301:13037:138787#ACAGTGAT	Chr8	24761856	→	8029	→	
43	FCC02UYACXX:1:2104:3514:18972#ACAGTGAT	Chr8	24761860	→	8020	→	
49	FCC02UYACXX:1:2204:14422:140561#ACAGTGAT	Chr8	24761860	→	8031	→	
32	FCC02UYACXX:1:1304:13188:29929#ACAGTGAT	Chr8	24761866	→	7998	→	
28	FCC02UYACXX:1:1302:8239:66248#ACAGTGAT	Chr8	24761875	→	7998	→	
48	FCC02UYACXX:1:2107:11814:99931#ACAGTGAT	Chr8	24761897	→	7995	→	
29	FCC02UYACXX:1:1303:11475:145106#ACAGTGAT	Chr8	24761919	→	7950	→	
22	FCC02UYACXX:1:1204:8499:81818#ACAGTGAT	Chr8	24761931	→	7957	→	
10	FCC02UYACXX:1:1101:19509:197758#ACAGTGAT	Chr8	24761936	→	7960	→	
26	FCC02UYACXX:1:1301:8306:56369#ACAGTGAT	Chr8	24761943	→	7936	→	
14	FCC02UYACXX:1:1103:8427:29496#ACAGTGAT	Chr8	24761967	→	7922	→	
11	FCC02UYACXX:1:1102:5318:105782#ACAGTGAT	Chr8	24761995	→	7871	→	
50	FCC02UYACXX:1:2204:19802:142992#ACAGTGAT	Chr8	24762014	→	7847	→	
46	FCC02UYACXX:1:2105:2952:131262#ACAGTGAT	Chr8	24762016	→	7830	→	
40	FCC02UYACXX:1:2101:15782:16589#ACAGTGAT	Chr8	24762027	→	7856	→	
16	FCC02UYACXX:1:1105:18480:153332#ACAGTGAT	Chr8	24762032	→	7832	→	
42	FCC02UYACXX:1:2102:5977:29193#ACAGTGAT	Chr8	24762035	→	7836	→	
41	FCC02UYACXX:1:2101:3370:22111#ACAGTGAT	Chr8	24762059	→	7831	→	
19	FCC02UYACXX:1:1201:15479:158826#ACAGTGAT	Chr8	24762154	←	292	←	
33	FCC02UYACXX:1:1304:1811:36152#ACAGTGAT	Chr8	24762162	←	272	←	
38	FCC02UYACXX:1:1307:18790:58656#ACAGTGAT	Chr8	24762195	←	227	←	
9	FCC02UYACXX:1:1101:8482:25950#ACAGTGAT	Chr8	24762200	←	252	←	
35	FCC02UYACXX:1:1304:13736:176499#ACAGTGAT	Chr8	24762204	←	236	←	
18	FCC02UYACXX:1:1107:21276:118304#ACAGTGAT	Chr8	24762214	←	222	←	
45	FCC02UYACXX:1:2104:2002:113881#ACAGTGAT	Chr8	24762231	←	215	←	
13	FCC02UYACXX:1:1102:19788:183949#ACAGTGAT	Chr8	24762241	←	197	←	
39	FCC02UYACXX:1:1308:18679:38205#ACAGTGAT	Chr8	24762241	←	193	←	
21	FCC02UYACXX:1:1203:18103:28232#ACAGTGAT	Chr8	24762262	←	175	←	
36	FCC02UYACXX:1:1305:18309:8419#ACAGTGAT	Chr8	24762273	←	189	←	
34	FCC02UYACXX:1:1304:19011:56582#ACAGTGAT	Chr8	24762285	←	159	←	
12	FCC02UYACXX:1:1102:10398:163344#ACAGTGAT	Chr8	24762318	←	124	←	
37	FCC02UYACXX:1:1306:20471:51487#ACAGTGAT	Chr8	24762341	←	94	←	
44	FCC02UYACXX:1:2104:20778:95349#ACAGTGAT	Chr8	24762359	←	83	←	
30	FCC02UYACXX:1:1303:6195:154204#ACAGTGAT	Chr8	24762364	←	75	←	
15	FCC02UYACXX:1:1104:13088:72587#ACAGTGAT	Chr8	24762372	←	71	←	
31	FCC02UYACXX:1:1303:19490:191151#ACAGTGAT	Chr8	24762382	←	44	←	
24	FCC02UYACXX:1:1207:11152:93193#ACAGTGAT	Chr8	24762396	←	47	←	
20	FCC02UYACXX:1:1202:14593:104111#ACAGTGAT	Chr8	24762398	←	44	←	

Challenge: perfect mapping but failure of experimental validation. CNV?

(Li *et al*., 2015, unpublished)

3.4 Application of genome variations

- Reference genome-based:
  - Genetic diversity
  - Evolutionary issues
  - Artificial selection on crops
- Reference genome-free:
  - Genetic diversity
  - Molecular markers

Recent advances based on genomic variations

- The return of population genetics (Lecture 6)
- Intracultivar genomic heterogeneity was observed
- Intercultivar genomic variation is so big

The return of population genetics

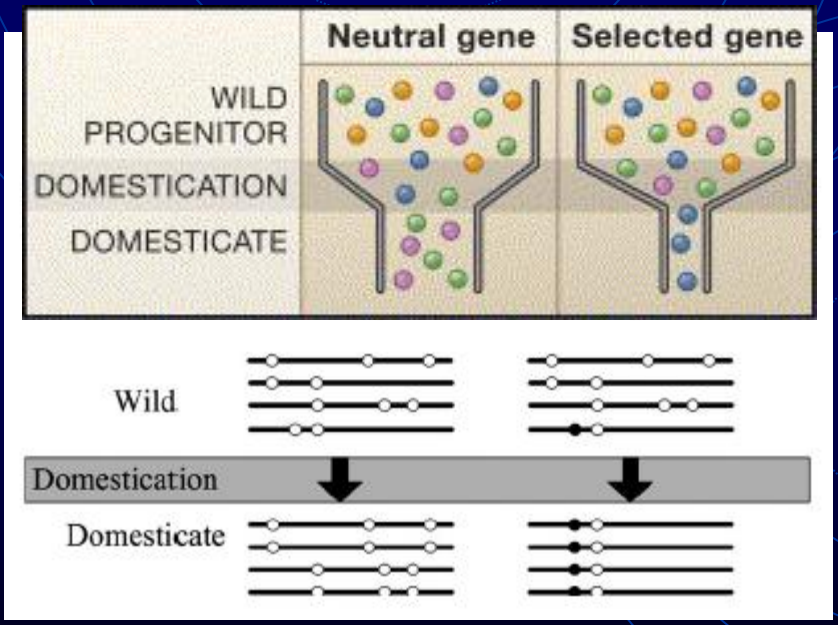
- Population genetics
  - Molecular population genetics: PCR-based
    - Molecular population genetics: high throughout-based
- A bridge between genomics and breeding: artificial selection
  - Breeding theory: genetic diversity, selection targets/strength, etc.
  - A bottom-up approach to find agronomic important genes

Two approaches to find positive/ adaptive genes

- The **top-down** approach:
  - QTL and LD mapping: from phenotype to candidate genes then molecular population genetics for signature of adaptation
- The **bottom-up** approach:
  - From molecular population genetics for signature of adaptation to candidate genes then find its function/phenotype

The effects of demography

- The effects of domestication bottleneck on genetic diversity



(Whitt and Gaut 2005)

Domestication bottleneck = domestication selection + demography effect

DNA diversity:
π (Tajima 1983)
θ (Watterson 1975)

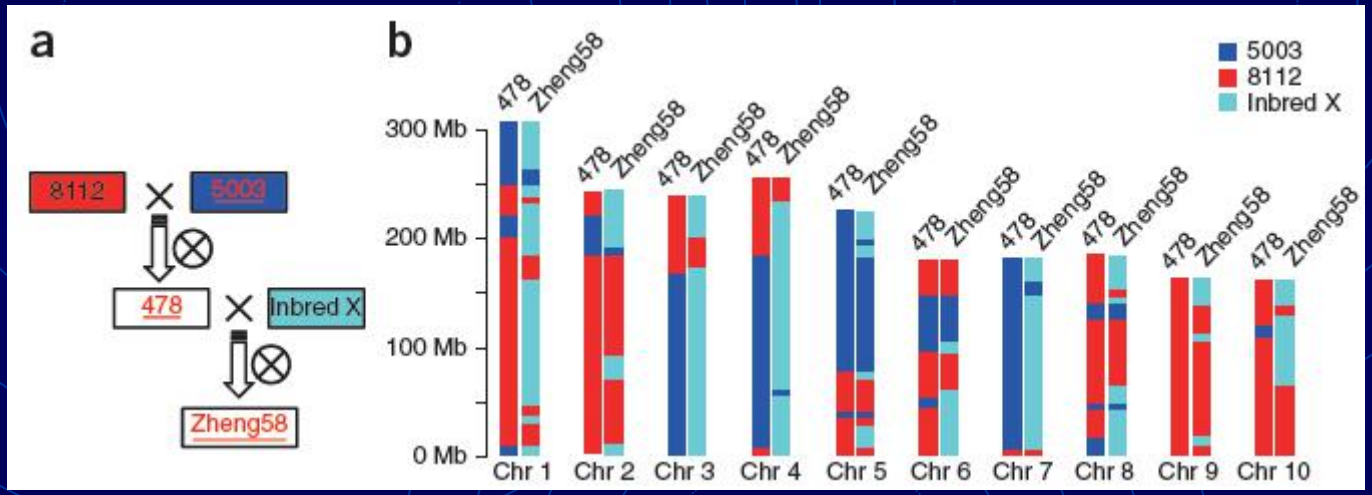
From the bottom up: Molecular population genetics

- Theoretical issues
  - Amount of diversity
    - Reduction of nucleotide diversity
  - Frequency distribution of polymorphisms
    - Selection skews the population frequency of genetic variants relative neutral equilibrium model (NE) expectations
    - An excess of rare variants relative to NE expectations
    - Or, with recombination, an excess of high-frequency derived (non-ancestral) mutations
  - Degree of association between polymorphisms/linkage disequibrium (LD)
    - Selective sweep increase LD

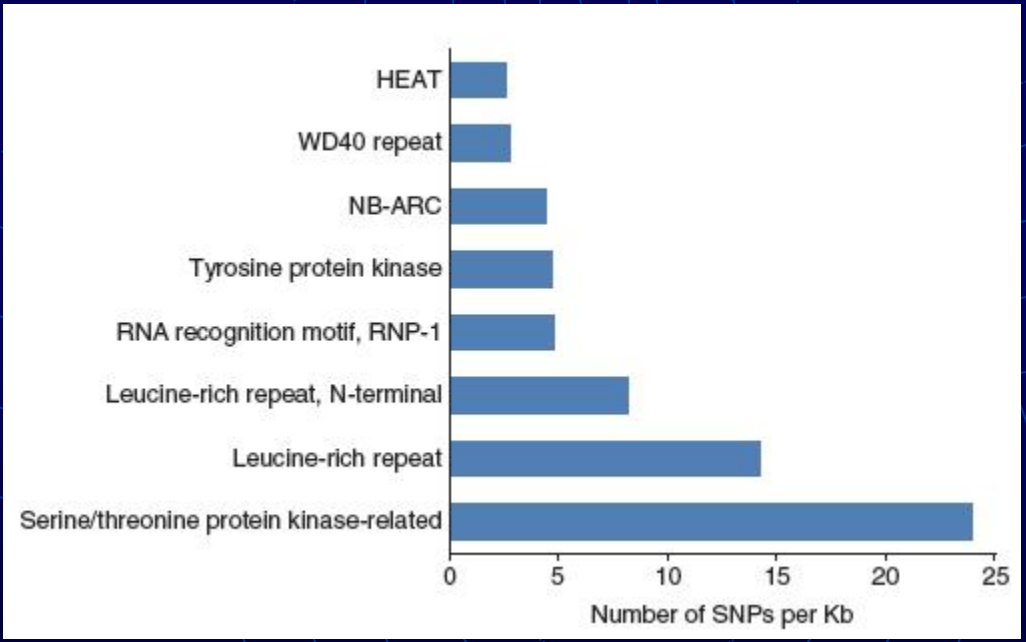
Intercultivar variation is so big

- A maize pedigree:
- Lai et al. 2010, Genome-wide patterns of genetic variation among elite maize inbred lines. Nature Genetics

Genetic background

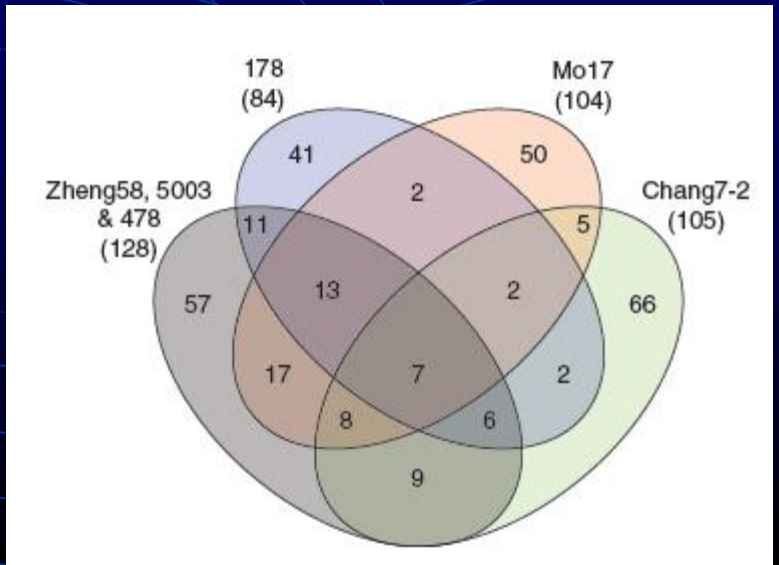


Annotation of large-effect SNPs



Numbers of PAVs relative to the B73 reference genome

- 296 high-confidence genes in B73 that were missing from at least one the six inbred lines.
- One large deletion between Mo17 and B73: ~2Mb with 24 genes



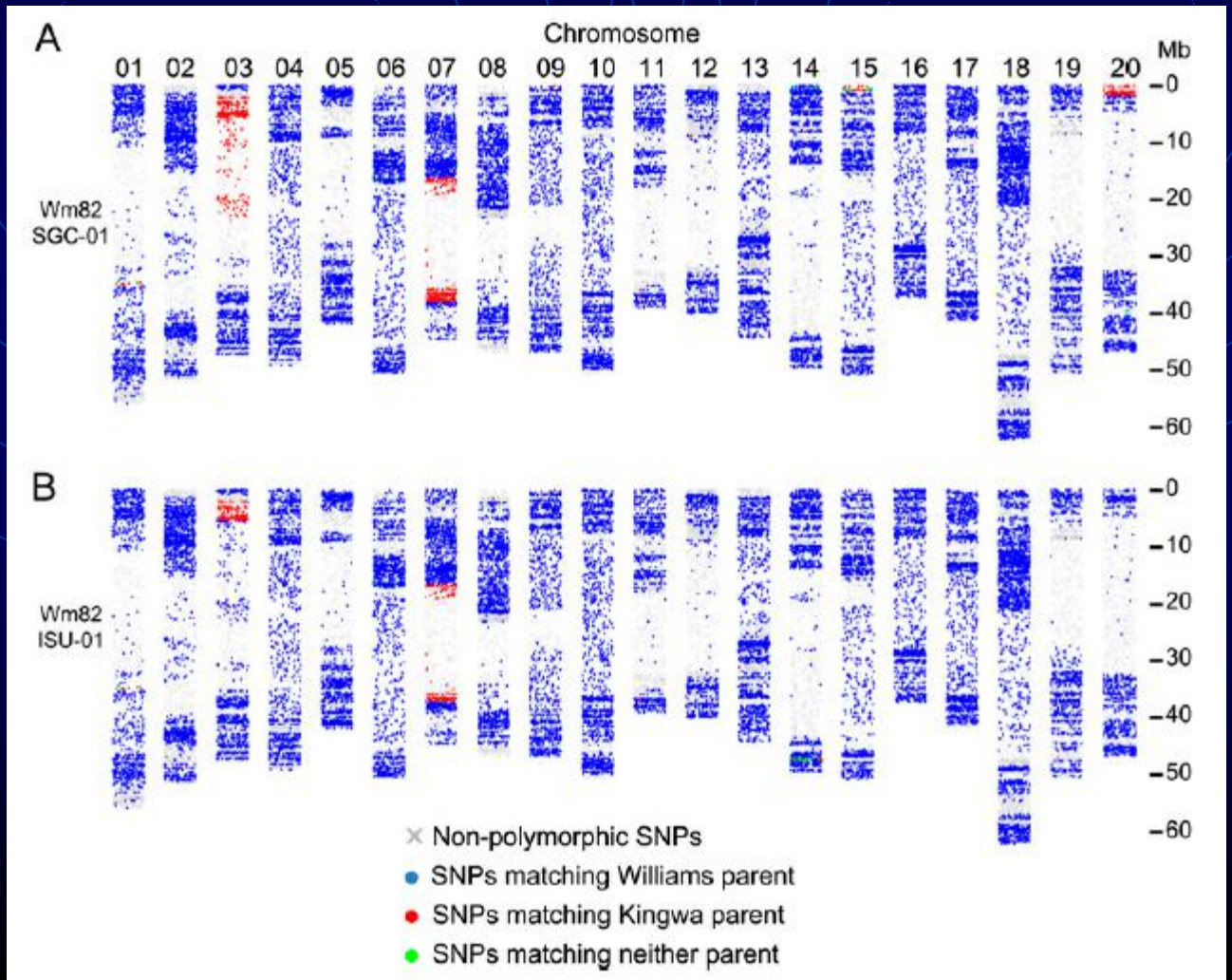
Intracultivar genomic heterogeneity was observed

- A same phenotype for individuals from a cultivar
- A reference genome of soybean (William 82): Haun et al. Plant Phiso., 2011

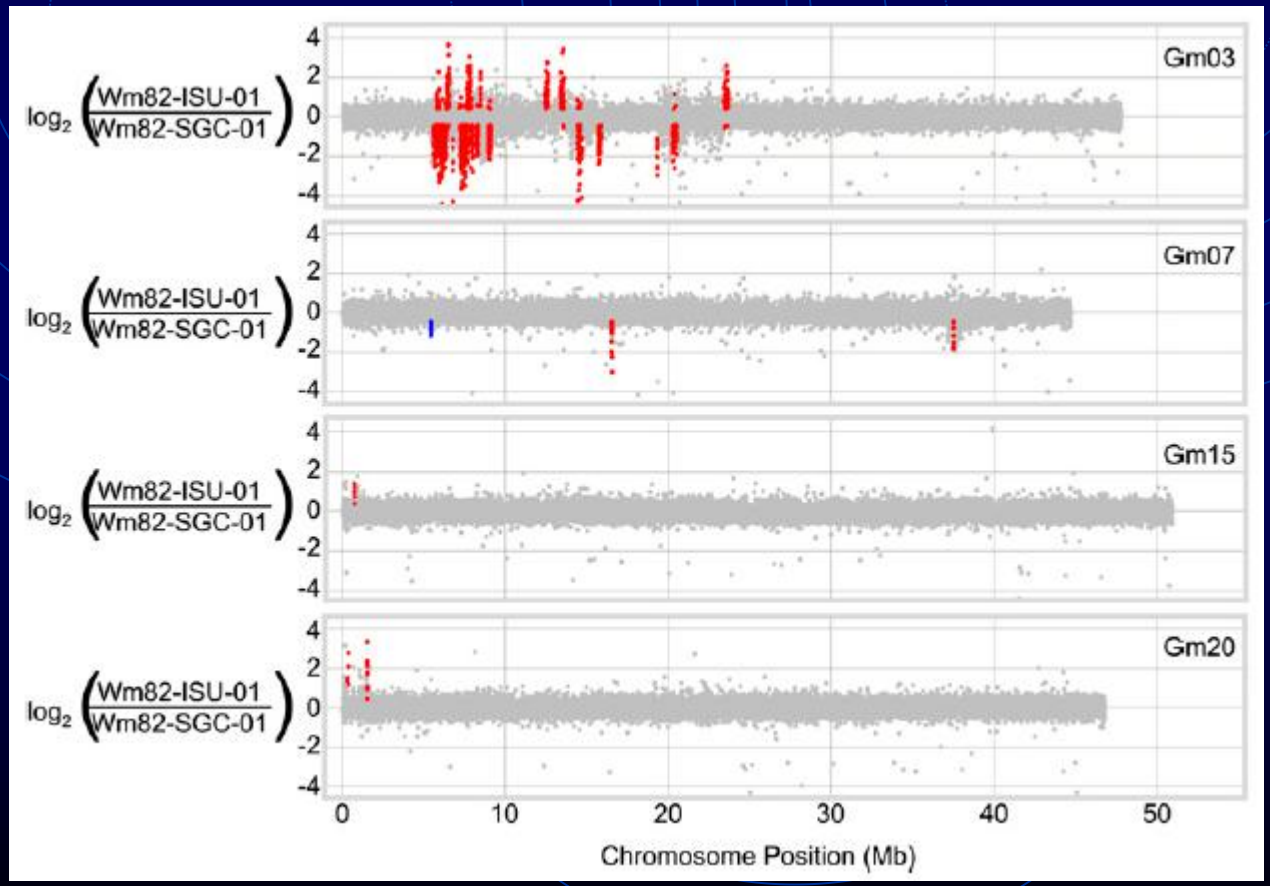
The composition and origin of genomic variation among individuals of the soybean reference cultivar Williams 82

- Haun et al. 2011, Plant Physio.
- Williams 82: a Williams × Kingwa BC_6F_3 generation

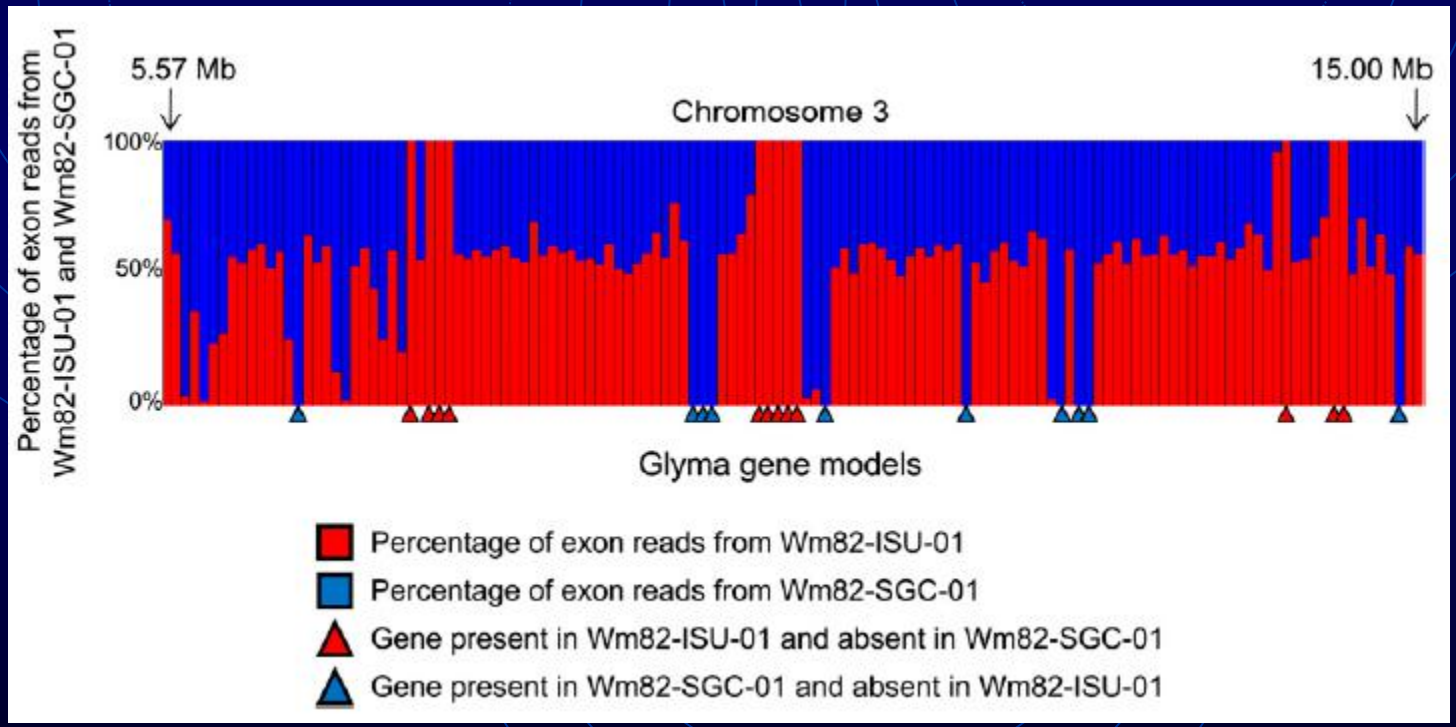
SNP genotyping (SNP chip) reveals the parental origins of Williams 82 genetic heterogeneity



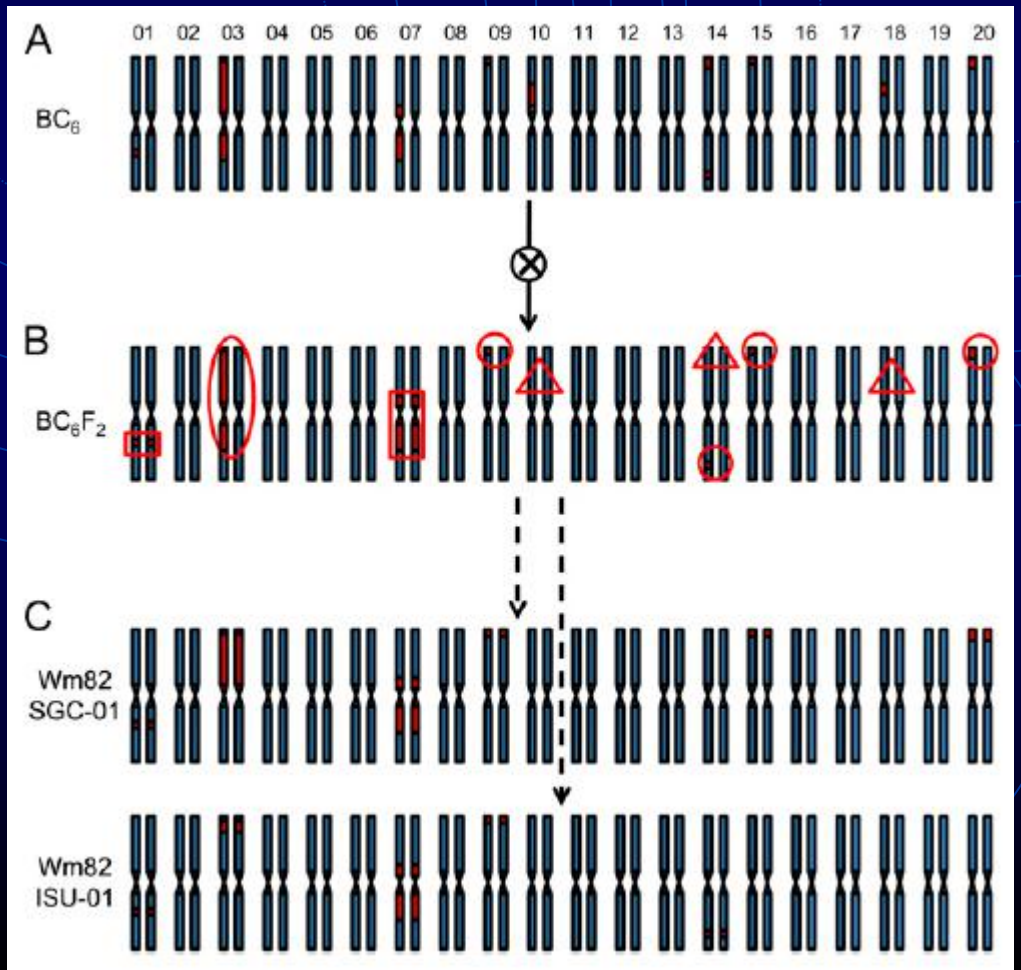
Structural variation (CGH) within regions of heterogeneity between two Williams 82 individuals



Exome resequencing reveals gene content variation between two Williams 82 lines



A model for the origin of genomic heterogeneity in two Williams 82 lines



Implications for the Williams 82 and other plant genome sequences

- Within regions of genetic heterogeneity, the reference sequences consist of a mosaic of the Williams and Kingwa haplotypes.
- Researchers investigating comparative studies of soybean that include Williams 82 as a reference genotype must factor in the inherent differences between each Williams 82 individual and the reference genome sequence.
- Similar considerations will need to be made for a variety of comparative methodologies, such as RNA-SEQ data.
- Similar circumstances may apply to the utility of other plant genome sequences