Development of SNP genotyping assays and applications in rice authentication

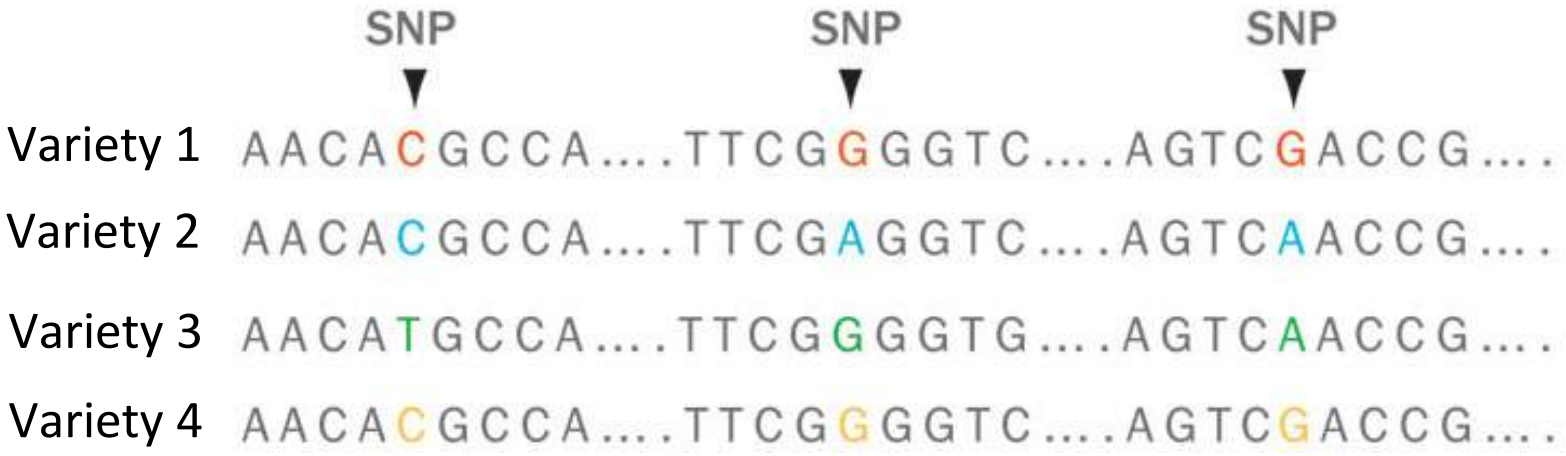
Chih-Wei Tung
Cornell University

Plant Authentication Workshop
IRMM, JRC, EC, Geel, Belgium
September 21, 2009

Outline

- SNP genotyping assays for rice
  - SNP genotyping platform choices
  - SNP discovery and SNP selection
  - QC and validation of SNP assays
- Rice authentication
  - Population structure
  - Phenotype-genotype relationships
  - Genetic similarity
  - “Reference” varietal profiles
- Rice diversity databases
  - Collection of varietal phenotypic profiles
  - Collection of critical alleles/haplotypes

Single Nucleotide Polymorphism (SNP)



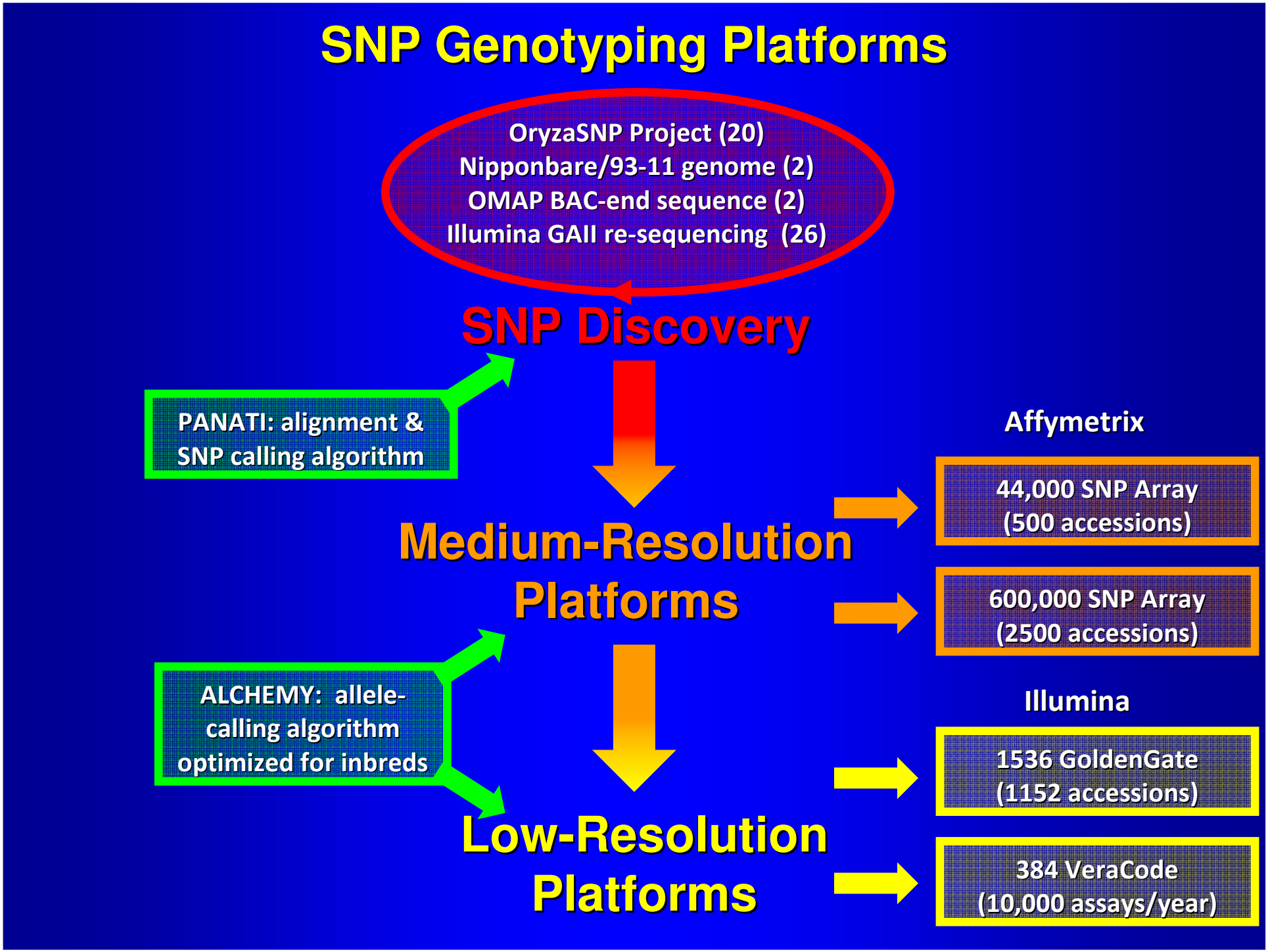
A **single-nucleotide polymorphism** (**SNP**) is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species

SNP discovery and genotyping platforms

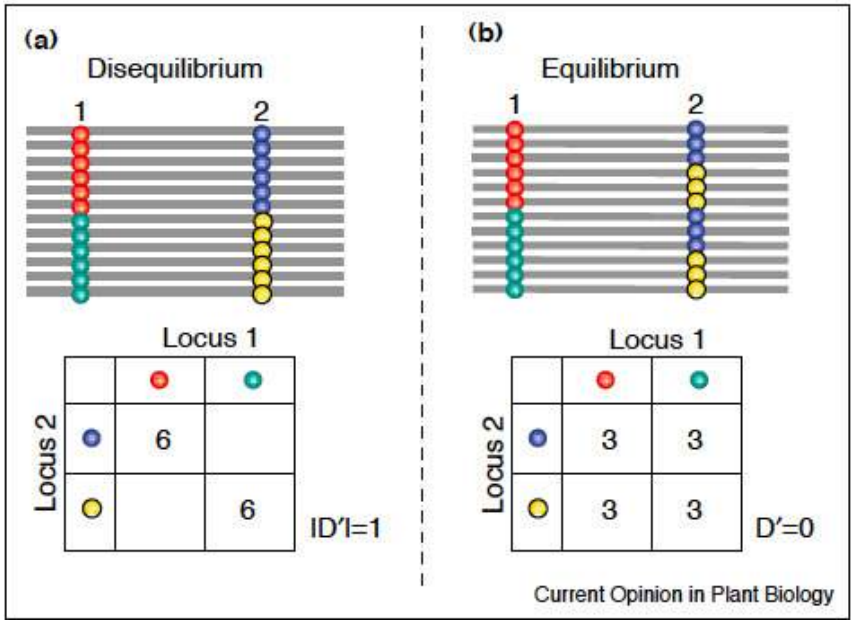
Technique	Scale	Equipment and platform cost	Reagents cost
Sequencing			
Resequencing (Sanger)	Low- to medium-throughput	High	High
Pyrosequencing	Medium-throughput	High	Medium
454 sequencing	High-throughput	High	Low
DNA conformation			
SSCP	Low- to medium-throughput	Low	Low
DGGE	Low- to medium-throughput	Low	Low
dHPLC	Low- to medium-throughput	Medium	Low
Allele-specific amplification	Low-throughput	Low	Low
Enzymatic cleavage methods			
CAPS and dCAPS	Low-throughput	Low	Low
TILLING	Medium-throughput	Medium	Low
Invader assay	Medium- to high-throughput	Medium	Medium–High
Allele-specific oligonucleotides			
Microarray-based	High-throughput	High	High
Taqman	Medium-throughput	Medium	High
Oligonucleotide ligation assay			
ELISA colorimetric assay	Medium-throughput	Low	Low
Rolling-circle amplification	High-throughput	Medium	Medium
Minisequencing			
Allele-specific primer extension	Medium- to high-throughput	High	Medium

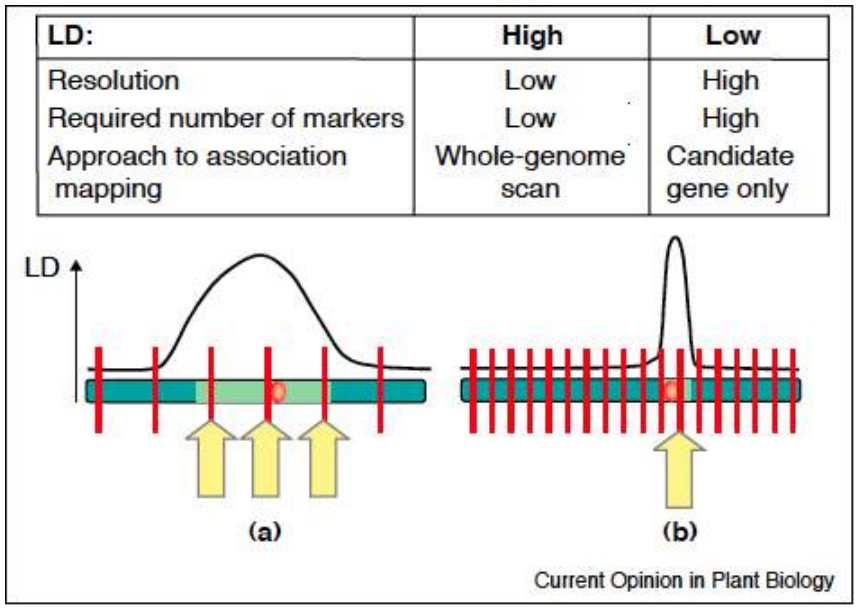
Chagne et al (2007) SNP genotyping in plants

SNP Genotyping Platforms



Linkage disequilibrium and association mapping

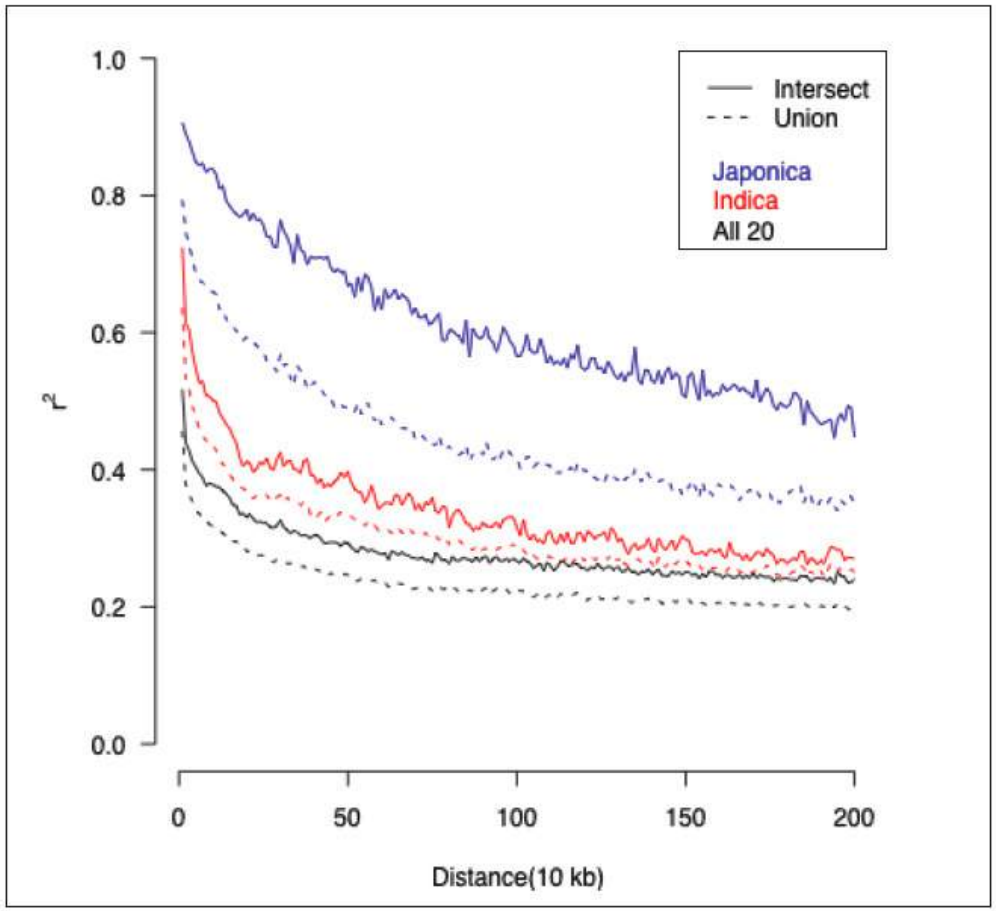




The high-LD chromosomal region around a marker locus defines the predictive range of a certain genetic marker. If LD within this genomic range is complete, any polymorphism within this range will have the same predictive value for the association with the phenotype. Hence, as a result of a significant marker-phenotype association, it can be concluded that the causative polymorphism locates within this high LD region around the marker locus.

Linkage disequilibrium in rice

LD in rice is expected to decay at around 100-200kb in *indica*, while it extends to over 500kb in *japonica*



LD was measured as r^2 between SNPs

K. Zhao (2009)

Criteria for SNP selection
Illumina 1536 SNP assay

1) Bi-allelic, single copy

   TACAACTTACAAGGATTATTGAGCTTCTTTCTTTCTCTGCCGTGGCTTCTATCTTTACCT
   [T/G]CTATGCATCTTCAACCTGACTTGATGATTGTTTGAAACCAATTTTAATGGAAGTTAAATG

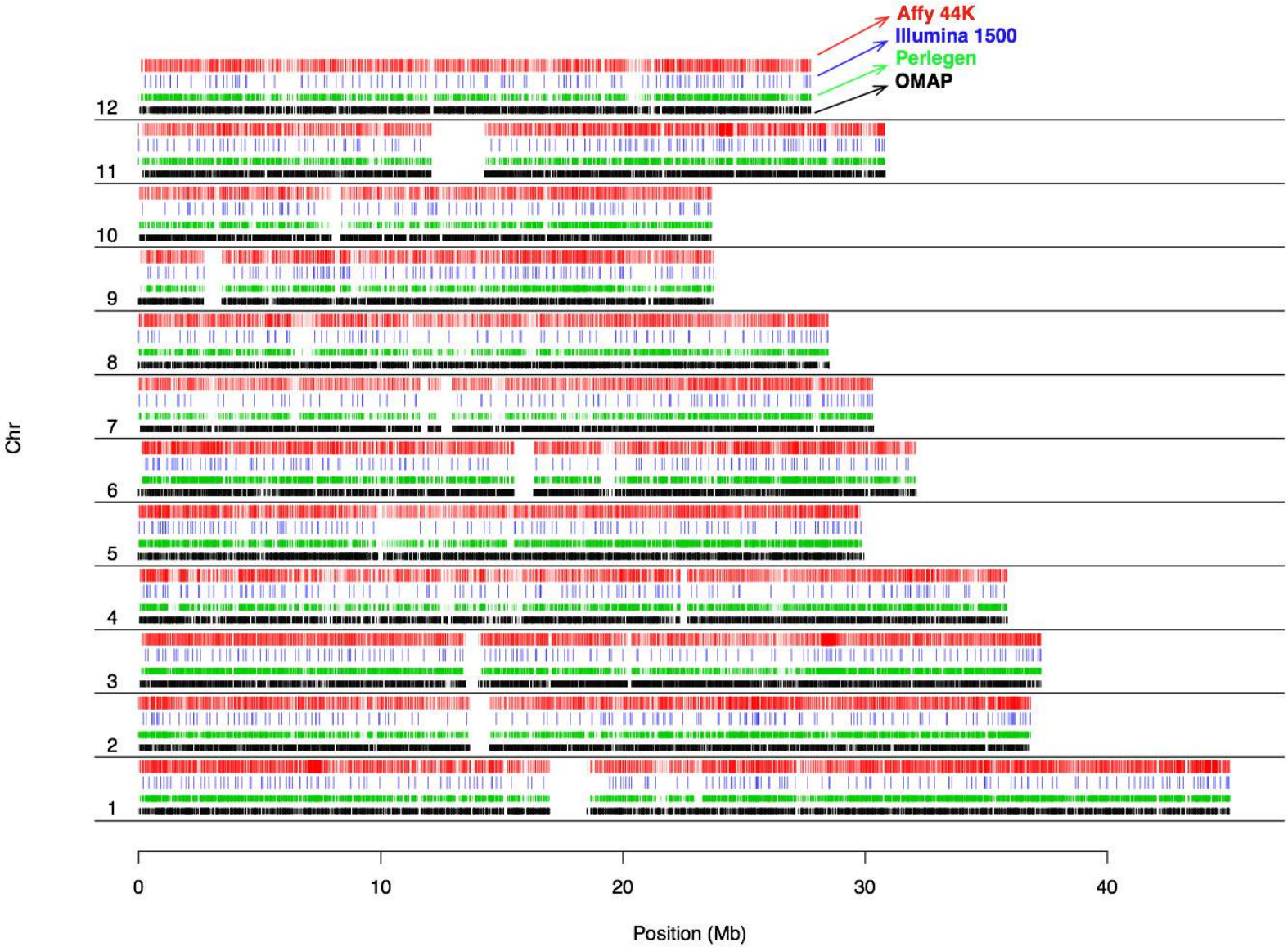
2) No SNPs within 60bp on either side of target SNP

3) None of the SNPs were in perfect LD with any other SNP within 500kb

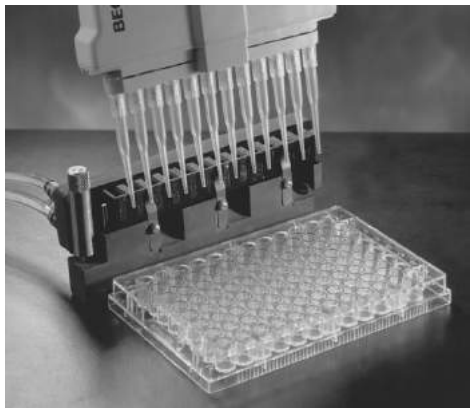
4) High frequency of polymorphism within major subpopulations of *O. sativa*

5) Good distribution throughout the genome

SNP distribution

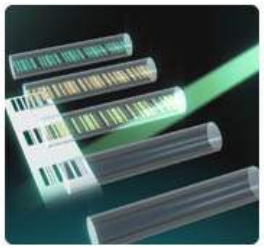


Illumina GoldenGate Assay
48 to 1536 loci

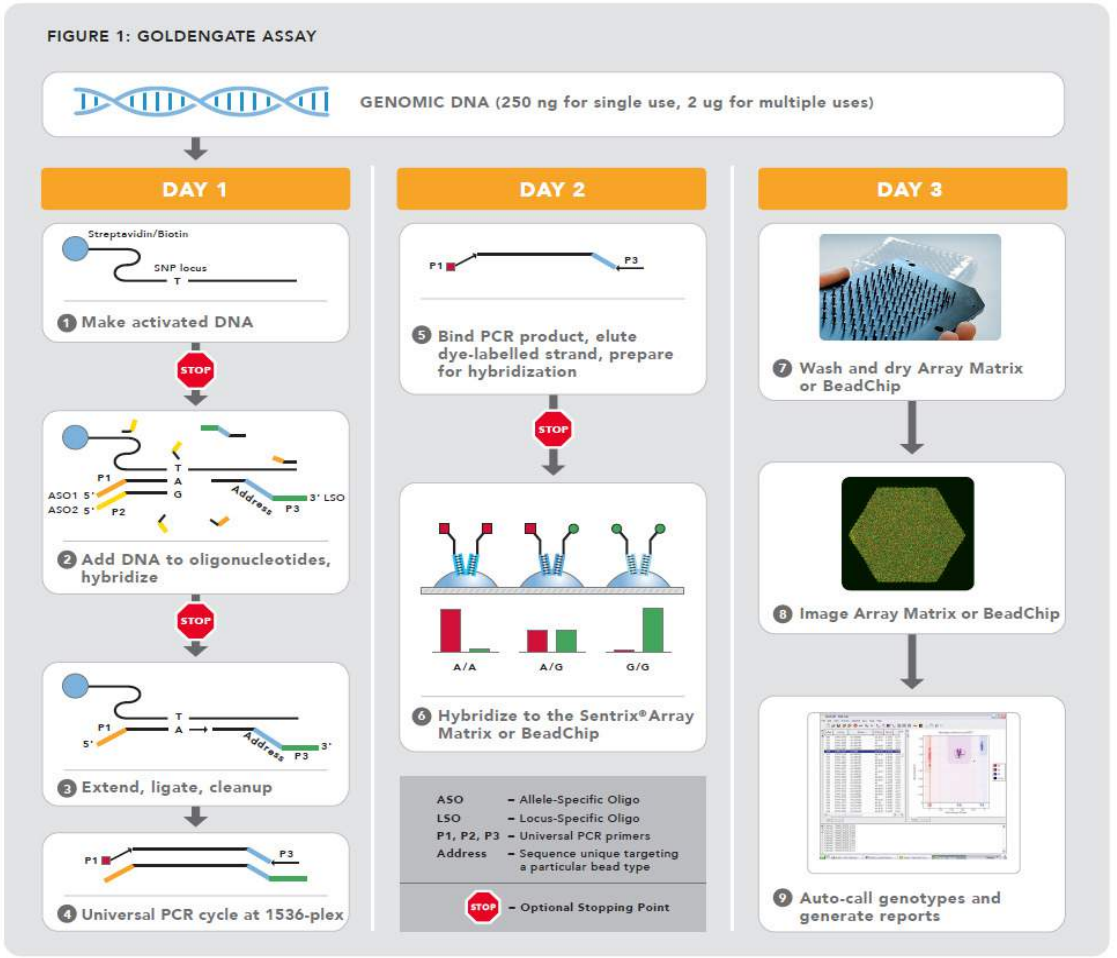


250ng of genomic DNA from samples

GoldenGate
12 – 96 samples
96-1536 SNP loci



VeraCode
96 samples
48-384 SNP loci



illumina®
making sense out of life

Quality control of SNP genotypes

- Individual sample call rate
  - 97% SNPs called
- SNP call rate
  - 99% SNPs called reliably across all samples being assayed
- Quality scores (per SNP)
- Concordance
  - Between replicates of same sample (99%)
  - Between genotyping platforms
- Validation
  - Genotyping data analysis

SNP verification using Re-sequencing data

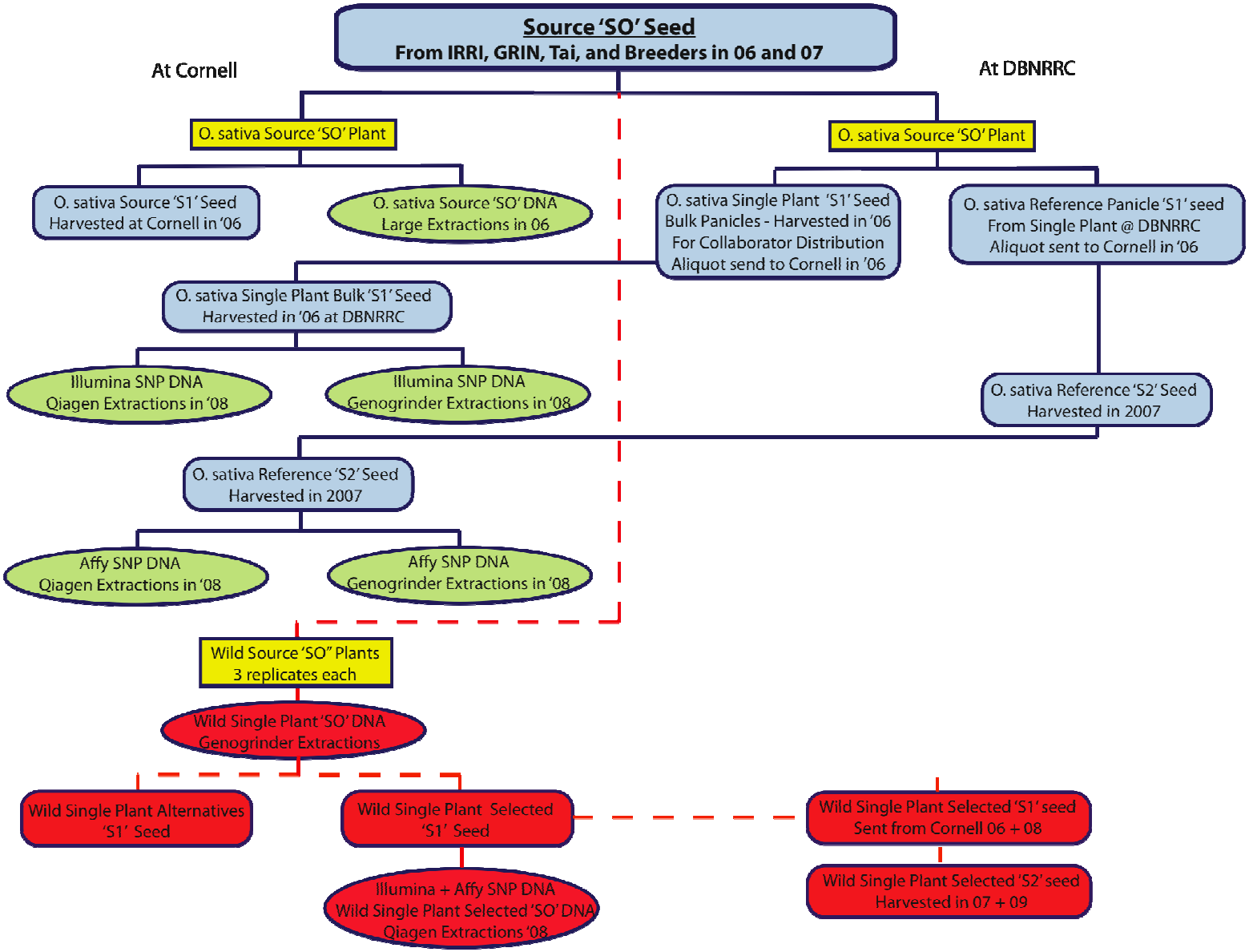
Sample/Chip	NSF-TV #	Accession Name	Same Source ?	Solexa Coverage	Accuracy
090818-A03	19	Black Gora	Yes	5.1X - 60bp pe	99.4%
090414-B05	29	Chau	Yes	4.6X - 60bp pe	99.3%
090209-A04	43	Dee Geo Woo Gen	Yes	2.9X - 36bp se	99.3%
090209-B05	94	Koshihikari	No	19.8X - 33bp se	99.1%
090415-D05	163	163	Yes	7.8X - 88bp se	99.3%
090226-A01	173	nipponbare	Yes	3.4X - 88bp se	99.6%
090324-A01	173	nipponbare	Yes	3.4X - 88bp se	99.5%
081216-D11	173	nipponbare	Yes	3.4X - 88bp se	99.6%
090209-A01	173	nipponbare	Yes	3.4X - 88bp se	99.6%
090204-H09	173	nipponbare	Yes	3.4X - 88bp se	99.6%
090204-H10	173	nipponbare	Yes	3.4X - 88bp se	99.6%
081215-A02	173	nipponbare	Yes	3.4X - 88bp se	99.7%
081216-E02	173	nipponbare	Yes	3.4X - 88bp se	99.7%
090302-E09	317	Dj 123	Yes	6.4X - 88bp se	99.3%
090331-G03	341	Shirkati	Yes	4.3X - 60bp pe	99.3%
090415-F10	621	Lagrue	No	6.8X - 53bp pe	97.6%
081215-C07	647	Cypress	No	6.2X - 53bp pe	98.2%
averages					**99.3%**

Mark Wright (2009) In preparation

Rice diversity panel

- *O. sativa* – 410 accessions (landraces, elite cultivars) from 80 countries
- *O. rufipogon* – 100 accessions from 14 countries
- 50 agronomic traits are phenotyped

Seed purification and management



Genotype reference samples

- 410 landraces, elite cultivars was genotpyed by 1536 SNPs assay (Displayed by Flapjack graphic viewer)

Genotype F1 hybrid

Genetic similarity

Measure pair-wise genetic distance

- the number of the SNP that are the same
- IBS (identical by state)
- 0, 1, 2 allele

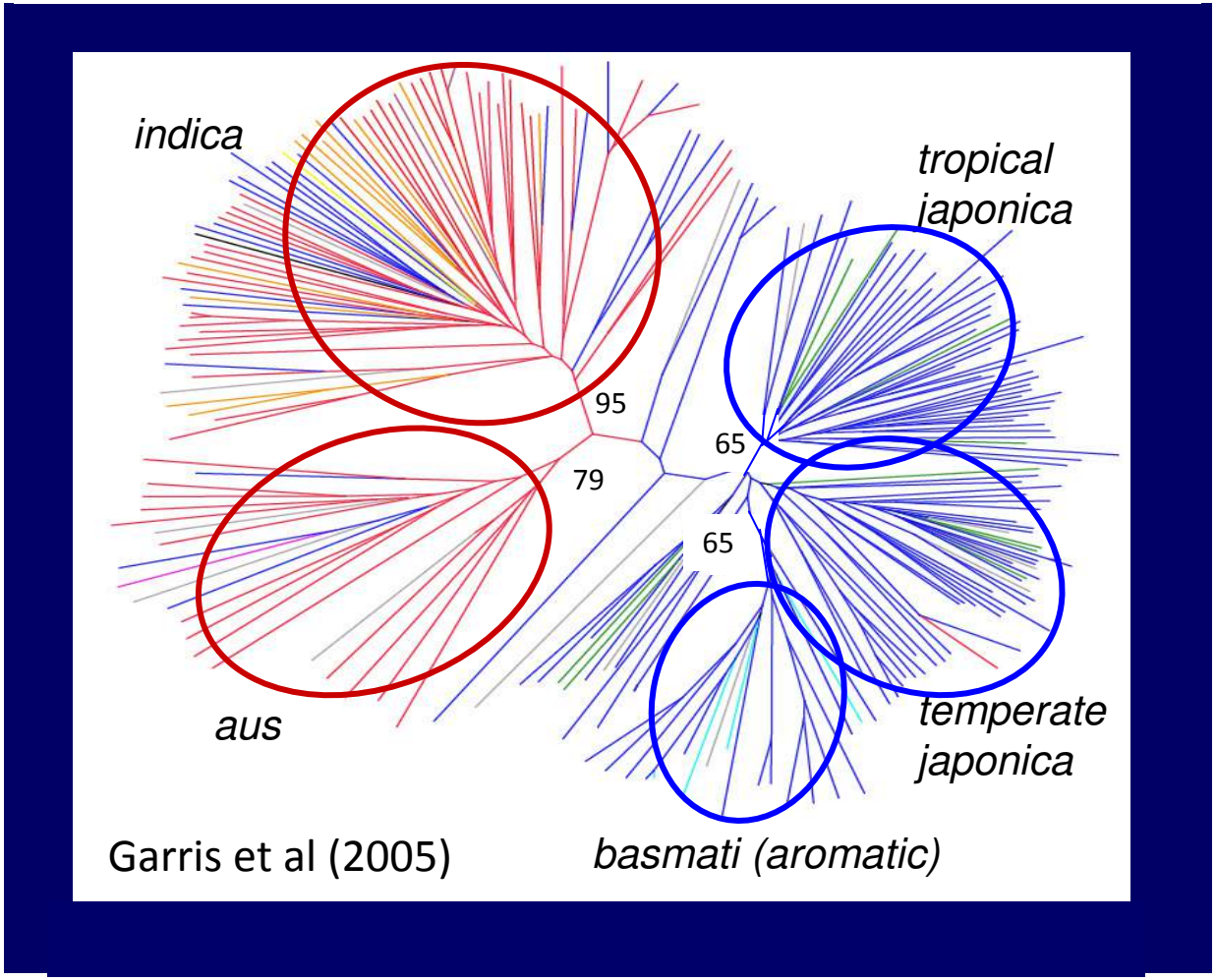
FID1	IID1	FID2	IID2	DST	IBS0	IBS1	IBS2	subpopulation
1	081215-A05	118	090324-A04	0.923701	70	1	853	Temperate Jap
1	081215-A05	355	090331-H04	0.92299	63	10	810	Temperate Jap
1	081215-A05	230	090331-E05	0.922708	72	1	865	Temperate Jap
1	081215-A05	179	090209-C06	0.90694	88	1	862	Temperate Jap
1	081215-A05	282	090105-C03	0.897667	96	1	846	Temperate Jap
1	081215-A05	180	090209-C07	0.897043	97	1	849	Temperate Jap
1	081215-A05	374	090302-G10	0.896902	96	1	839	Temperate Jap

FID1	IID1	FID2	IID2	DST	IBS0	IBS1	IBS2	Subpopulation
5	081215-A08	53	090325-A12	0.9617	36	1	916	Aromatic
5	081215-A08	16	081215-A12	0.96166	36	1	915	Aromatic
5	081215-A08	640	081216-E07	0.942122	53	2	878	Aromatic
5	081215-A08	93	090209-B04	0.941774	54	1	881	Aromatic
5	081215-A08	640	081215-C01	0.941176	56	0	896	Aromatic
5	081215-A08	45	090325-A09	0.935313	61	0	882	Aromatic
5	081215-A08	191	090209-D02	0.934668	62	0	887	Aromatic
5	081215-A08	373	090302-G09	0.914286	26	74	635	Aromatic
5	081215-A08	124	090414-C10	0.912037	71	10	783	Aromatic
5	081215-A08	160	090324-C08	0.902105	92	2	856	Aromatic

Accessions similarity measured by # of matching alleles/ # of comparisons (Simple statistic by Flapjack viewer)

Population structure analysis

1536 SNP and 44K SNP genotyping assays



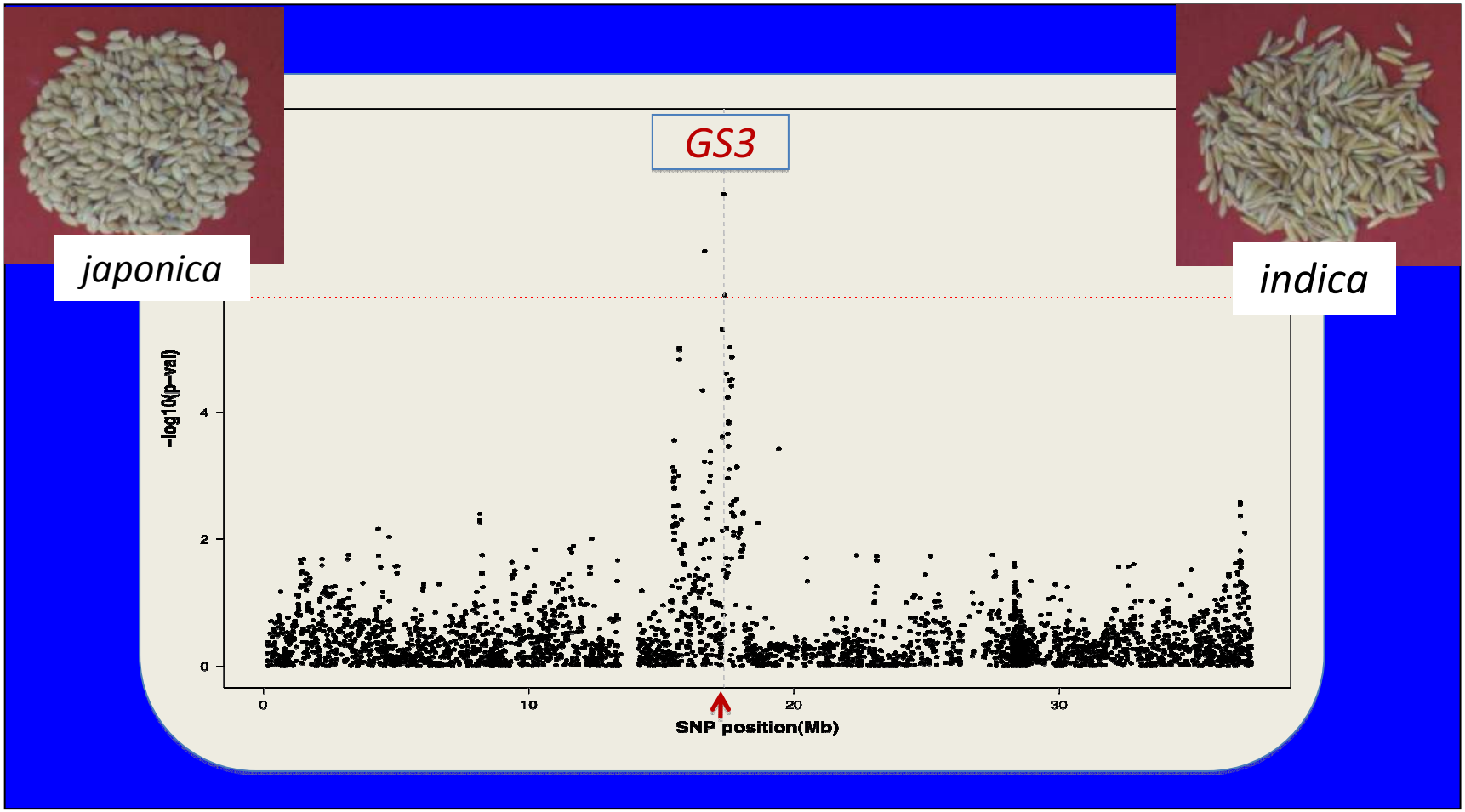
5 sub-populations

- *indica*
- *aus*
- *tropical japonica*
- *temperate japonica*
- *aromatic (basmati)*

SNP, SSR and isozyme analyses recognize similar sub-populations

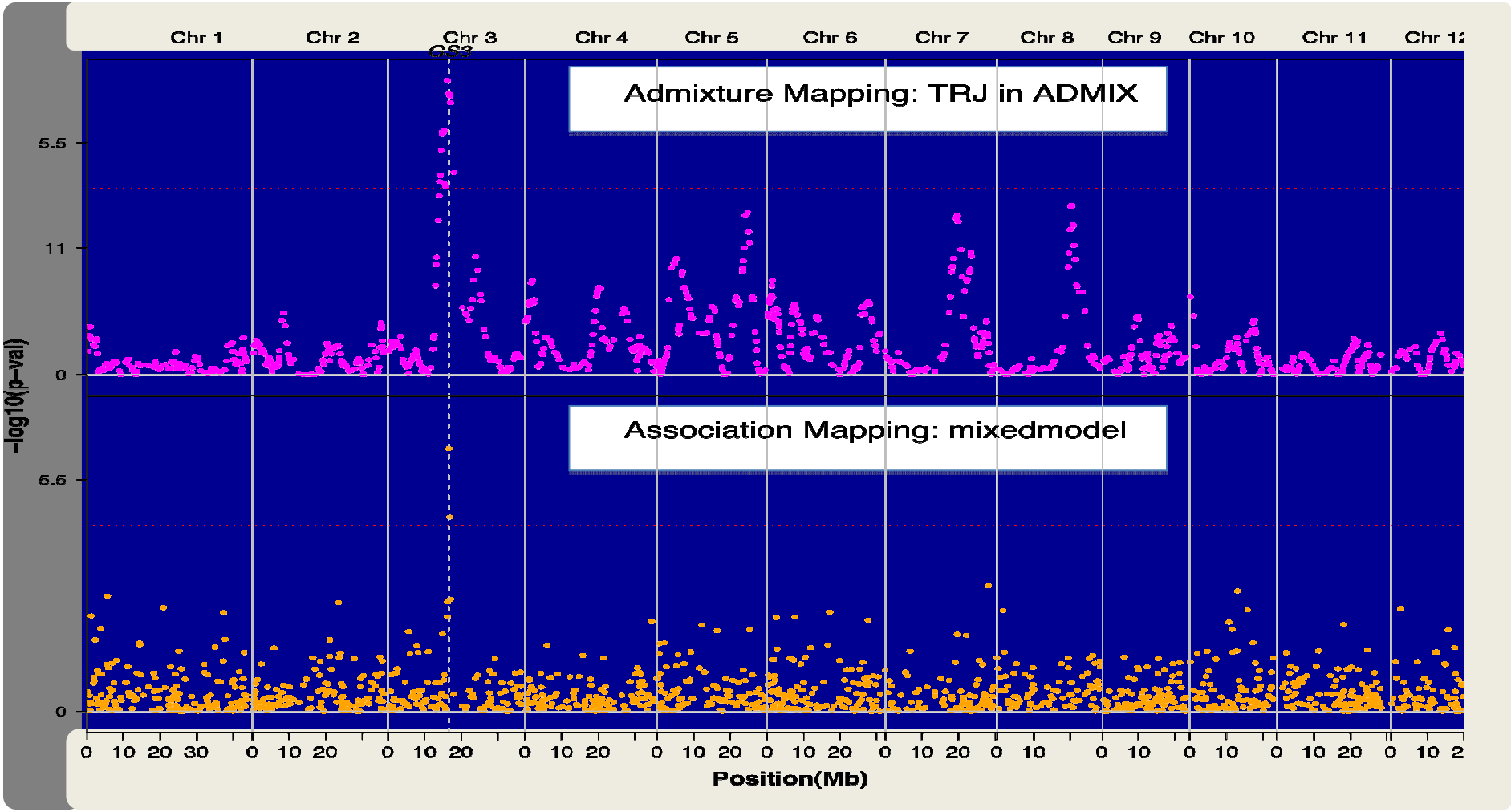
Phenotype-Genotype association using the 44K SNP chip

Genome wide association mapping for grain length using the subpopulation component matrix Q (K=5) as a cofactor in the model identifies the *GS3* region as most significant



Grain size gene GS3 also detected using 1536 SNP assays

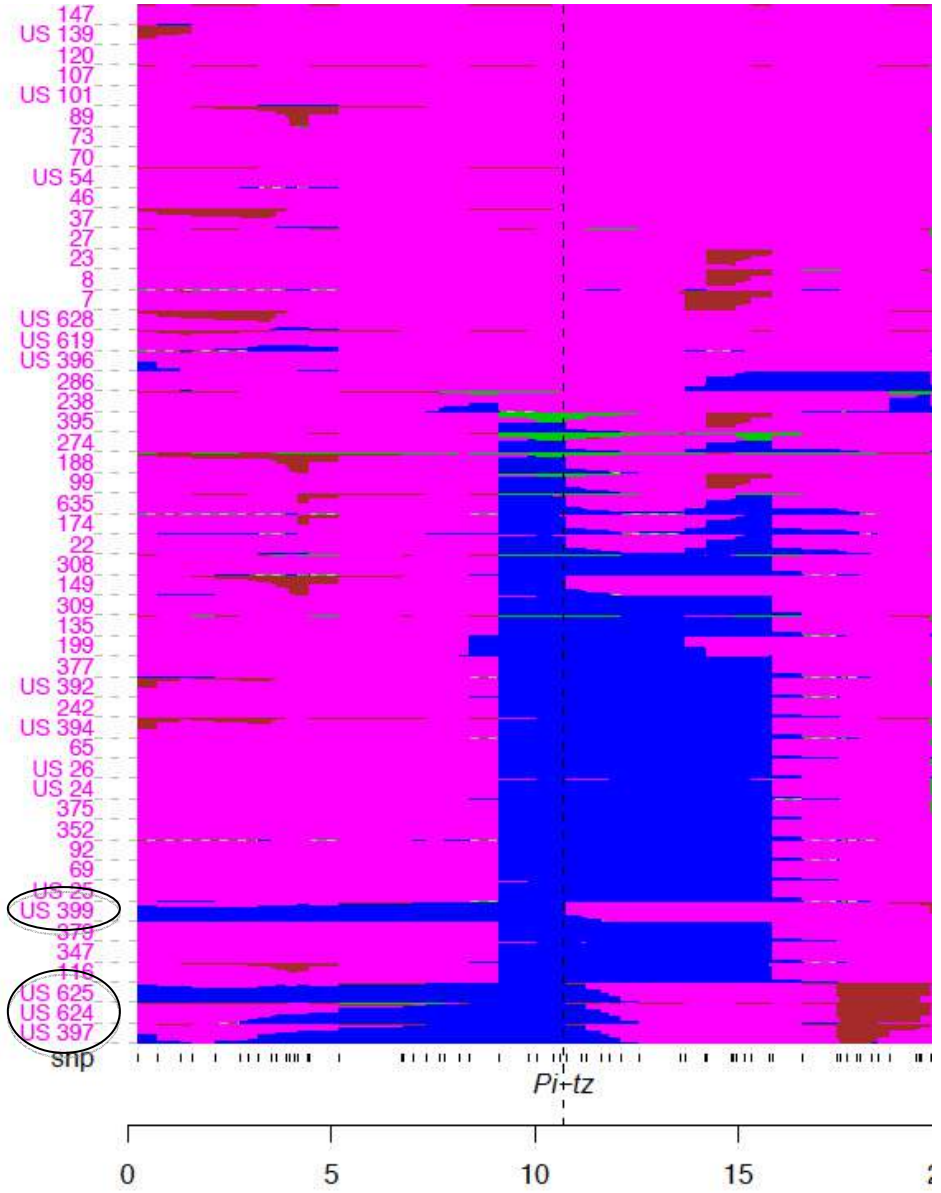
Admixture mapping requires fewer SNPs to identify significant phenotype-genotype associations than association mapping



Blast disease resistance associated with SNP on chr. 12 (Introgression analysis)

***Pi-ta*: blast resistant gene**

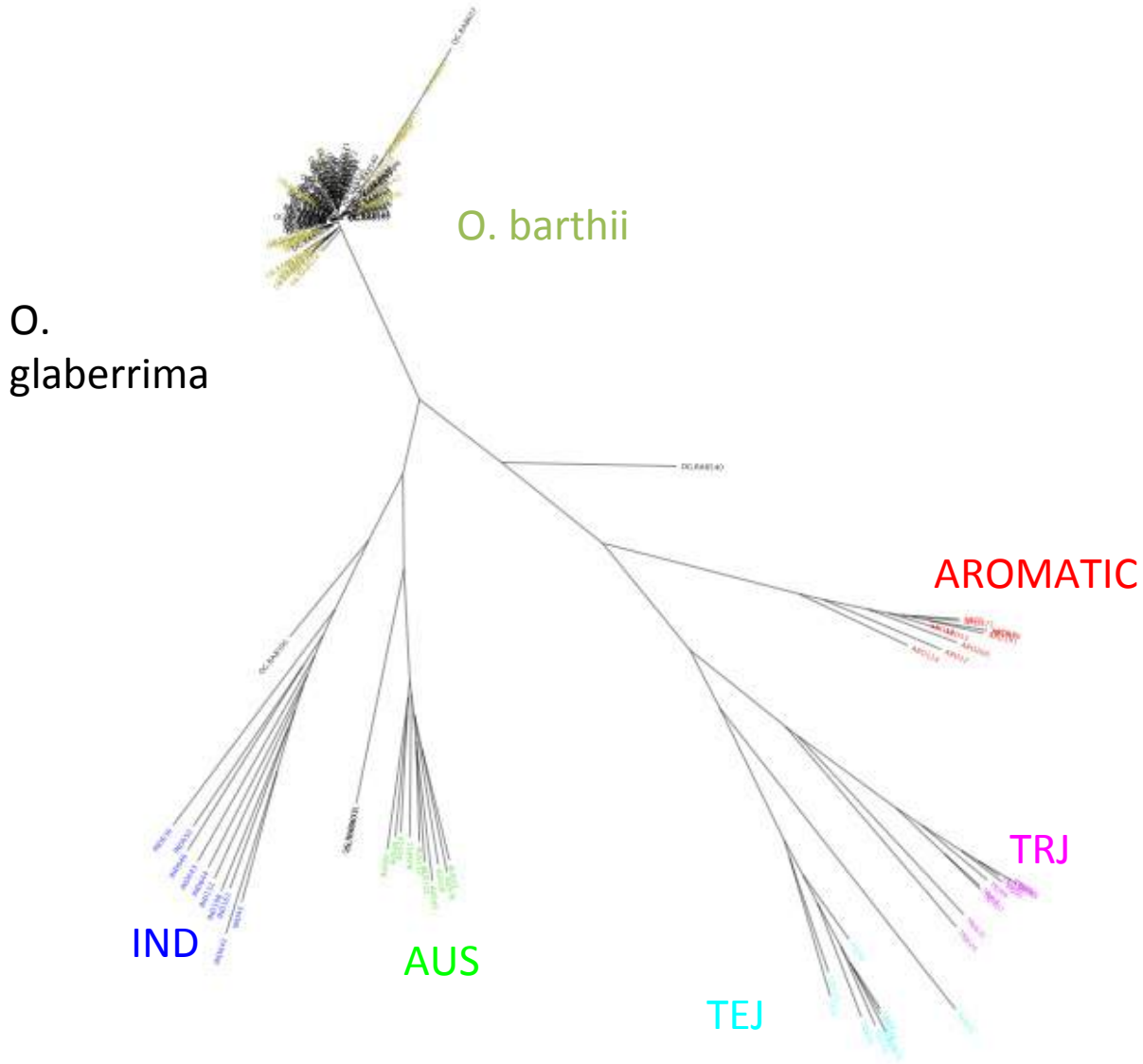
- Originated from *indica* variety
- Introduced into some US rice varieties
- Our SNP assays confirm the blast-resistant varieties in the US carry *pi-ta* functional allele



Genotype African material

OB - *O. barthii*
OG - *O. glaberrima*

50 *O. Sativa* controls
(10 each subpop)



Breeder's SNP chips

Illumina 384-plex VeraCode technology

- In collaboration with IRRI
- Based on SNP polymorphisms between:
  - *Indica* x *indica*
  - *Japonica* x *japonica*
  - *Indica* x *japonica*
  - Wilds x *japonica*
  - Wilds x *indica*
  - 5 Subpopulations (Fingerprinting)

Genotyping NIL, RIL, F2, CSSL population using SNP arrays

Re-sequencing more rice varieties

- Cornell University, IRRI, JGI-JBE, USDA and some external resources
- Illumina GA II technology (7x to 20X)
- *Indica*, *tropical japonica*, *temperate japonica*, Aus, aromatic, *O. rufipogon*
- Gene Discovery
- SNP pool for generating 600K SNP array

Summary

Our SNP genotyping assays can:

- Reveal genetic similarity between accessions
- Identify F1 hybrid genotype
- Assign subpopulations
- Trace origin of introgression
- Discover SNPs contribute to agronomic traits
- Detect samples mix-up

Strategy of applying high-throughput SNP assays for rice authentication

“Reference” varietal profiles

- The group that wants to authenticate varieties must develop reference varietal profiles defining the range of variations that will be classified as “variety X”
- Because samples with same variety name or accession id may be genetically different, deep sampling is required to establish reference profiles
- Our SNP platforms offer cost-effective, efficient genotyping assays that can be easily used to establish reference profiles


27
B2008-001
T0070/2008
CHAHORA 144
PAKISTAN
GSOR #: 301025
NSFTVID: 27
Acc. 2786
Chahora
(Pakistan)
27869
123 07_27_1_223_a 37
SPA = SINGLE PANICLE
Absent
Short & partly awned
Short & fully awnded
Long & partly awned
Long & fully awned
PANICLE
1 Compa
3
5 Interme
7
9 Open
9
98
PLANT HEIGHT measur
RC 2007 single panicle harvest
MG-002-160
CHAHORA 144
NRP - 3 seeds - 1/19/04
RA4934
CHAHORA 144
Tai#160
5/31/02
O.sativa
source:Tom Tai
Cornell McCouch Lab, Tai set
027
Chahora 144
IRGC 27869
O. sativa
NTS
SPA
DB-NRRC
PI 584557 OR02AR SD
Crop: RICE
Name: Chahora 144
Order: 205608
NSF-TV SET
NSF ID# 27
IRRI ID: IRGC 27869
GRIN ID: PI 584557
GSOR# 301025
ACCESSION NAME:
Chahora 144
GSOR 301025
CORE PHOTO
05/18/2007

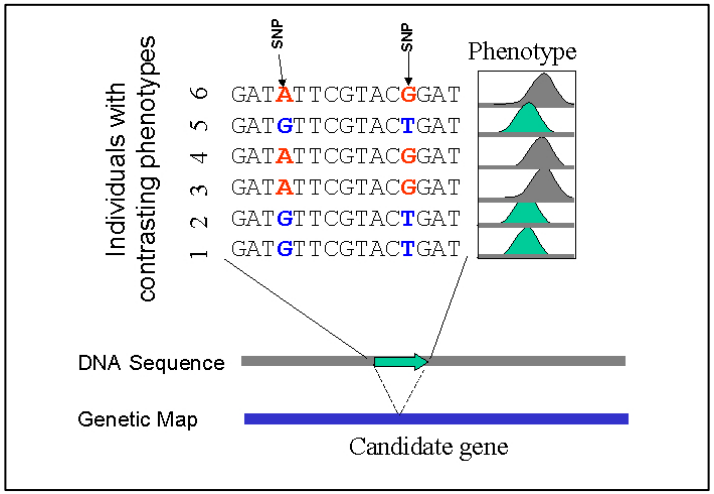
Sample sources are different, but having same name

Rice varietal authentication

- Collect reference samples
  - Multiple samples per variety from difference sources
  - Designated authority recognizes “true to type”
- Genotype reference samples
  - SNP platforms are cost-effective
  - SNP data easy to database
- Identify key phenotypic attributes
  - Identify genes/SNP haplotypes associated with key traits (Association mapping can help)
- Develop custom SNP arrays
  - Targeting genes/haplotypes/subpopulation/germplasm of interest

Rice SNP/Diversity Database

- Assemble a collection of functional SNPs and haplotypes associated with key phenotypic attributes
- Assemble a collection of varietal SNP profiles
- 1536 and 44K SNP genotype data will be released to public



Example

Authentigene company in UK < http://www.authentigene.com/>

Basmati SNP profile

RICE VARIETY LINEAGE	Basmati 370 Pure Bred	Derhadun Pure Bred	Kernal Pure Bred	PK370 I Pure Bred	Pak 385 Hybrid	Pusa 1 Hybrid	Sherbati Non-Bas	Super Hybrid	Taraori Pure Bred
	GA	GG	AA	GG	GG	AA	GG	GG	GA
	AA	AA	AA	AA	TT	AA	TT	AA	AA
	GG	GG	GG	GG	GG	GG	CC	GG	GG
	TT	TT	TT	TT	TT	CC	CC	TT	TT
	GG	GG	GG	GG	GG	CG	CC	GG	GG
	GG	GG	GG	GG	GG	GG	CC	GG	GG
	GA	GA	GA	Ga	GA	GA	GA	GA	GA
	GG	GG	AA	GG	GG	AA	GG	GG	AA
	TC	TC	TC	TC	CC	CC	CC	TC	TC
	GG	GG	GG	GG	CC	GG	GC	GG	GG
	AG	AA	AA	GA	GG	GG	GG	AA	AA

Are 11 SNPs enough to distinguish all Basmati and non-Basmati varieties?

Comparison of SSR and SNPs

Marker	Advantage	Disadvantage
Microsatellites (SSR)	• Highly informative (large number of alleles • Low ascertainment bias • Easy to isolate • Easy to set up	• High mutation rate • Homoplasy • Difficult to automate • Cross-study comparisons require special preparation • Not easy to database
SNPs	• Very stable • Very abundant • Easy to score and database • High throughput • Lots of analytical approaches • Easy to make cross-study comparisons	• Cost to develop • Ascertainment bias • Low information content of a single SNP (bi-allelic nature), need 5-10x SNPs for same resolution, in terms of diversity. • Require statistical and informatics support

Acknowledgement

- NSF-TV project # 0606461 (Cornell University)
- USDA-RiceCAP, DBNRRC
- Japan-NIAS
- Affymetrix company
- IRRI
- OMAP project
- Plant Bioinformatic Group, SCRI