

# Genome-wide SNP Discovery (*OryzaSNP*)<sup>a</sup> and Survey (*HaplOryza*)<sup>b</sup> in Rice



K.L. McNally, K.L. Childs, V. Ulat,  
R. Clark, R. Bohnert, G. Zeller,  
G. Rättsch, D. Weigel, D. Hoen,  
T. Bureau, R. Stokowski, D. Ballinger,  
K. Frazer, D. Cox, R. Bruskiewich,  
D. Mackill, C.R. Buell, R. Davidson,  
J. Leach, and H. Leung

*See Poster 1.23*

*Funded by:*

IRRI<sup>a</sup>



K.L. McNally, C. Billot, G. Droc,  
B. Courtois, A.A. Farouk, N. Ahmadi,  
G. Clément, J. Taillebois, B. Barry,  
G. Second, D. Brunel, A. Bérard,  
M. Lathrop, and M. Foglio

*See Poster 1.20*

# Project Rationales

## *OryzaSNP*

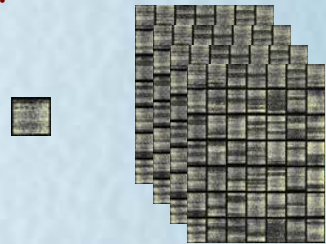
- Discover genome-wide SNPs for 100 Mb in 20 diverse rice varieties
- Understand extent of LD among variety groups and across the genome
- Develop set(s) of tag SNPs for genome scanning

## *HaplOryza*

- Genotype 1536 indica/japonica SNPs on 900 varieties
- Understand extent of LD in specific regions and across the genome
- Identify SNPs affiliated with domestication events

# OryzaSNP Platform

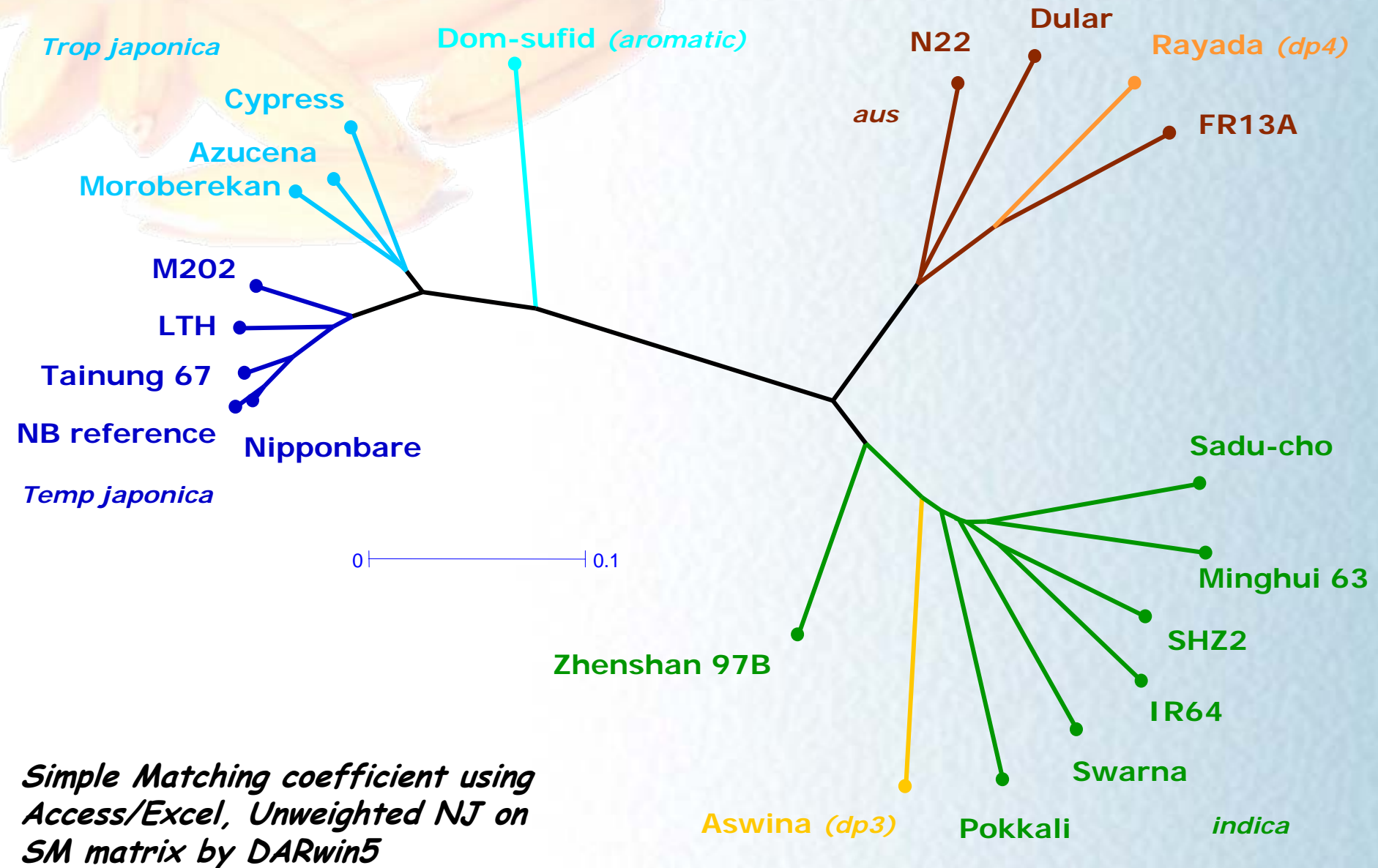
- 100 Mb of Nippon-bare reference genome
  - No repetitive DNA, fewer than 10 hits to genome for remaining segments, maximal coverage of annotated gene models (TIGR and RAP2)
- 20 diverse varieties chosen for utility, functional knowledge, diversity from all *Oryza sativa* variety groups
  - 1 reference/control, 3 temperate japonica, 3 tropical japonica, 1 aromatic, 2 deep-water, 3 aus/boro, 7 indica
- Perlegen re-sequencing technology using very high density oligomer arrays (*Affymetrix*)
  - 25-mers tiled with 1-base offset for both strands and full degeneracy at 13th base (8 oligos/base)
  - 1 chip + 5 wafers (5'x5') for 20 varieties
- LR-PCR amplification for preparation of targets
  - > 13,582 LR-PCR amplicons for each variety
- Hybridization, analysis, model-based prediction using Perlegen algorithms (259,721 SNPs called)



## On-going

- Annotation, quality assessment (recall and FDR rates), machine-learning calls via HMM-SVMs, LD analyses, tag SNPs, ...

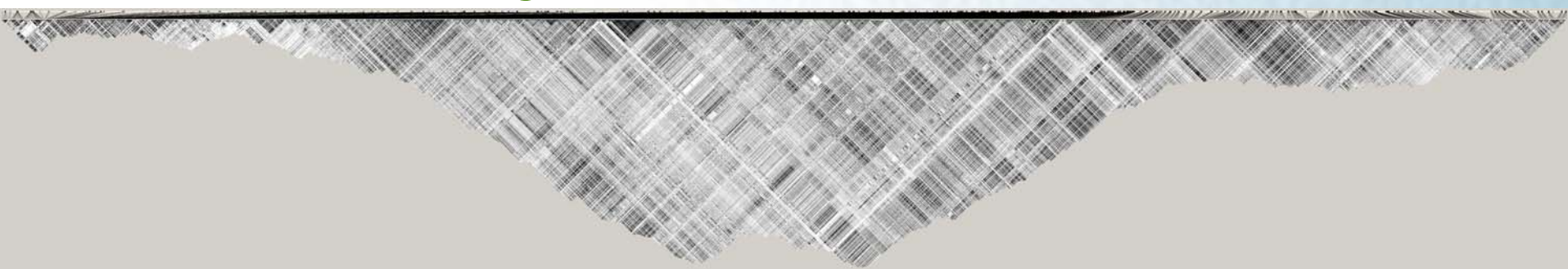
# Variety Groups from 259,721 SNP Sites



# Summary Statistics for Model-Based Calls

Chr	IRGSP r4 (bp)	Tiled_length	Non-repetitive	SNP_Sites	SNPs/kb	Total_calls	Clear	Ambiguous	2-allelic
1	45064769	15411868	13377445	36993	2.7653	739860	205653	140140	36156
2	36823111	12695346	11028257	27628	2.5052	552560	146965	101119	27012
3	37257345	13687037	12353186	28228	2.2851	564560	156412	110220	27506
4	35863200	10674305	9111561	22226	2.4393	444520	102068	96606	21615
5	30039014	9054116	7799270	20433	2.6199	408660	114697	89165	19842
6	32124789	9491255	8000505	25549	3.1934	510980	133891	103692	24945
7	30357780	8989073	7646272	14202	1.8574	284040	67925	82363	13787
8	28530027	8179660	6923419	16117	2.3279	322340	79672	85411	15625
9	23843360	6900725	5840458	15349	2.6280	306980	72399	72929	14955
10	23661561	6405335	5355995	12248	2.2868	244960	63611	69859	11847
11	30828668	8063435	6768998	21606	3.1919	432120	97304	106348	21031
12	27757321	7333621	6243138	19142	3.0661	382840	86771	84495	18622
<b>All</b>	<b>382150945</b>	<b>116885776</b>	<b>100448504</b>	<b>259721</b>	<b>2.5856</b>	<b>5194420</b>	<b>1327368</b>	<b>1142347</b>	<b>252943</b>
							0.2555	0.2199	0.9739

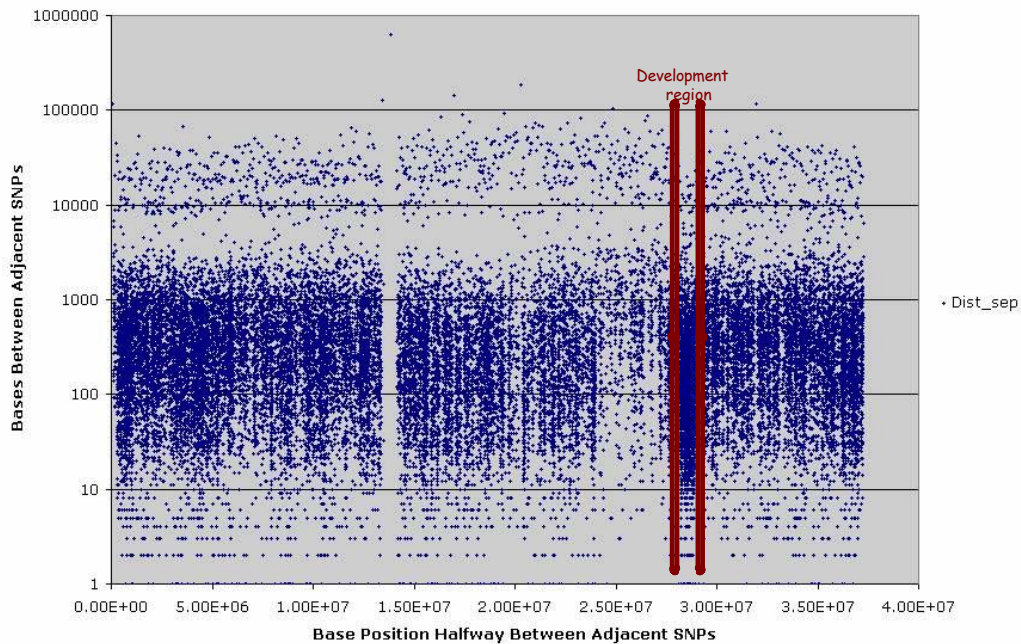
## Pairwise $r^2$ for 7 Mb region on Chr3 (5000 SNPs)



# Pairwise SNPs at 259,721 sites

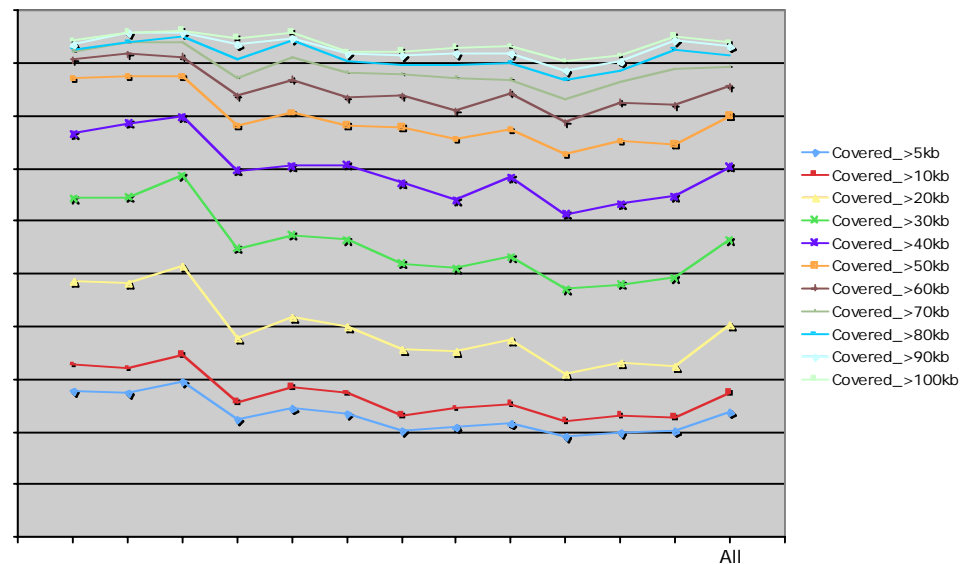
	NB_ref	Aswi	Azuc	Cypr	Doms	Dular	FR13A	IR64	LTH	M202	MH63	Morob	N22	Nippo
NB_ref		101198	32004	39028	55042	97271	100965	91543	20261	22267	91381	35712	88159	5425
Aswi	101198		75049	79185	74488	48230	48602	40154	85759	83349	38859	81300	58511	88377
Azuc	32004	75049		21667	41319	72775	77642	65784	25309	26325	69375	19346	64858	28067
Cypr	39028	79185	21667		45103	79468	83444	73741	33925	33246	74180	25113	70742	35634
Doms	55042	74488	41319	45103		68703	74143	75004	44760	49206	76805	44004	66946	49866
Dular	97271	48230	72775	79468	68703		32813	55471	81389	82436	53903	77023	33187	86207
FR13A	100965	48602	77642	83444	74143	32813		56045	85372	85546	54342	82599	42282	89117
IR64	91543	40154	65784	73741	75004	55471	56045		77888	76208	29479	74620	61567	79829
LTH	20261	85759	25309	33925	44760	81389	85372	77888		23367	78987	29125	70542	18787
M202	22267	83349	26325	33246	49206	82436	85546	76208	23367		75659	30376	73997	19342
MH63	91381	38859	69375	74180	76805	53903	54342	29479	78987	75659		75780	60531	80340
Morob	35712	81300	19346	25113	44004	77023	82599	74620	29125	30376	75780		69095	32101
N22	88159	58511	64858	70742	66946	33187	42282	61567	70542	73997	60531	69095		78607
Nippo	5425	88377	28067	35634	49866	86207	89117	79829	18787	19342	80340	32101	78607	
Pokka	90481	40935	68933	75425	72948	51921	51512	35705	78278	76580	36595	75890	56569	79083
Rayad	100630	51494	75633	83015	75379	35213	29588	59740	85158	85514	58743	78463	45992	89497
SaduC	97375	39631	71538	79058	78381	55042	56296	35511	83820	79261	32180	79888	62173	85631
SH22	81976	37787	63153	69012	68909	49960	51181	23228	71965	69489	28832	70117	56260	70492
Swarn	91879	37923	69513	76275	74283	55867	54218	33146	79560	74409	33383	77623	62329	79730
TNG67	12648	86556	27040	34764	50278	84198	87186	77517	18880	20594	77071	32559	76719	10286
ZS97	72123	47790	58590	65878	67720	57043	55533	38918	63768	61301	38035	66176	57676	61763

Chr03 SNP-gap Distribution



SNP distribution

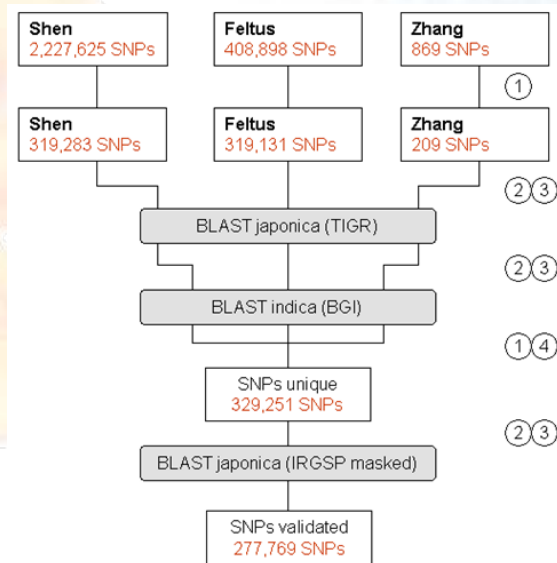
Genome coverage



# HaplOryza Platform

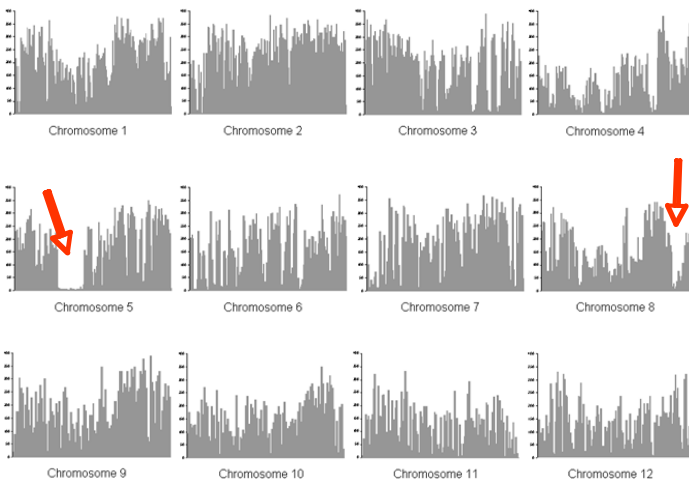
- 1536 SNPs chosen from Nipponbare (japonica) and 93-11 (indica) genome comparisons for
  - LD scan of the entire genome with 1 SNP/320 kb (76%)
  - A closer LD scan on specific regions with 1 SNP/50 kb (24 %)
    - test for introgression of different material:
    - two low density areas (chr7) and two normal density areas (chr 12)
  - all types of mutation (syn vs non-synonymous),
  - Best Illumina scores (> 0.9)
- 900 accessions
  - A good representation of the overall genetic diversity (GCP Composite)
    - "MiniGB" accessions (241), Complement from remaining 2757 accessions (231), and from previous study (46)
  - Accessions presenting putative patterns of introgressions between indica and japonica groups
    - Highland rice from Madagascar (132), Guinean rice (39), Chinese rice (12), Surinam rice (23), Breeding crosses: European (12), Brazilian (11)
  - 20 Reference accessions chosen for the OryzaSNP project 133 Wild accessions
- Illumina **GoldenGate** assay

# Rice SNP pipeline



- Only 1 polymorphism over a 50 bp window
- Only 1 hit to *japonica* or *indica* pseudomolecules (Excluding redundant hits)
- 100% identical sequence of 20 bp on either side of the polymorphism
- Only 1 SNP by location (Excluding redundant SNPs between the 3 datasets)

## SNP repartition in the genome



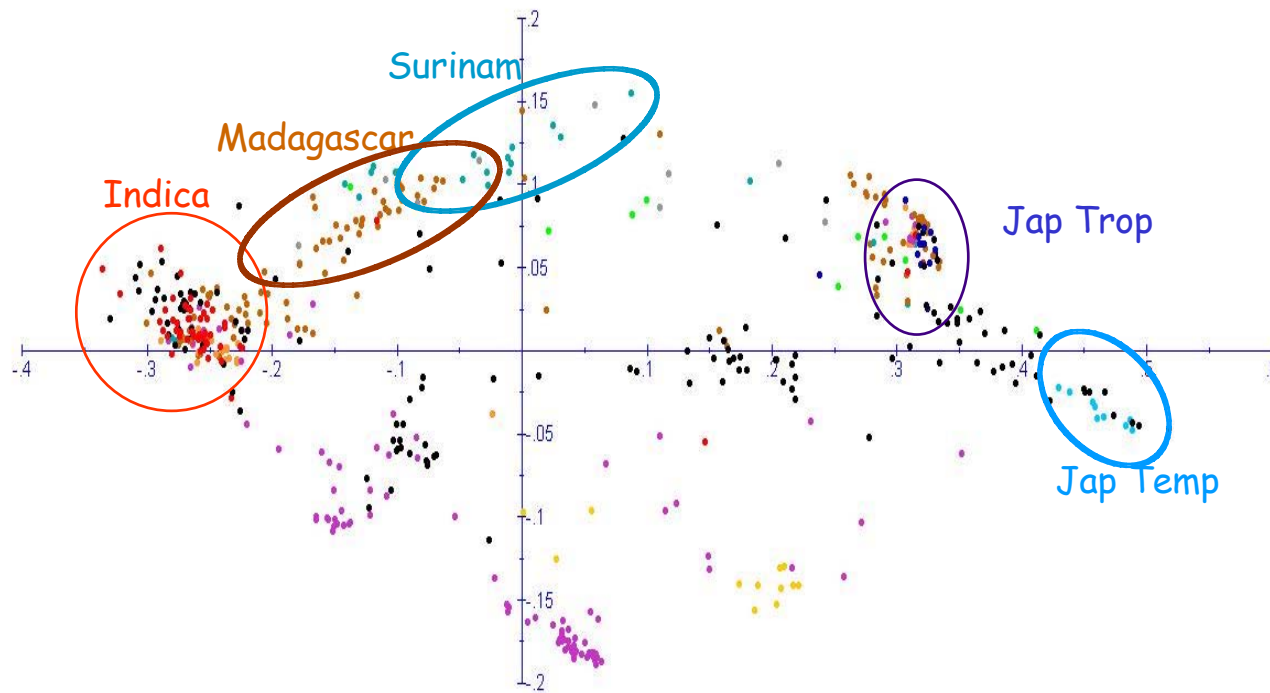
### Intergenic & genic

### Introns & exons

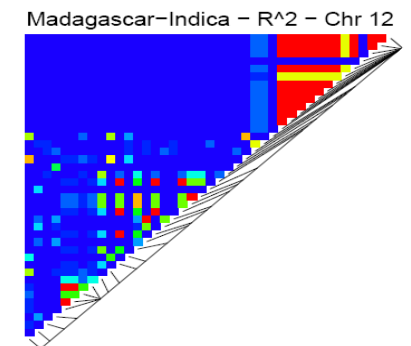
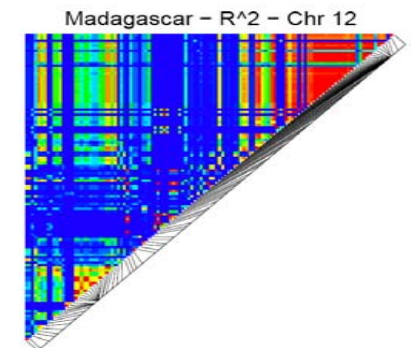
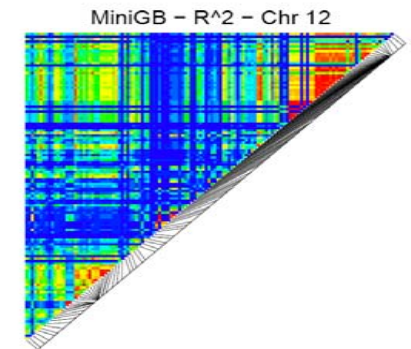
Chr.	SNP	Intergenic & genic		Introns & exons							
		Intergenic	%	Genic	%	UTR	%	Intron	%	Exon	%
1	40469	21917	54,2	18552	45,8	1385	7,5	10299	55,5	6868	37,0
2	37387	20848	55,8	16539	44,2	1294	7,8	9288	56,2	5957	36,0
3	31152	16335	52,4	14817	47,6	1561	10,5	8009	54,1	5247	35,4
4	19594	10544	53,8	9050	46,2	750	8,3	4866	53,8	3434	37,9
5	20873	11788	56,5	9085	43,5	784	8,6	4933	54,3	3368	37,1
6	20821	11928	57,3	8893	42,7	718	8,1	4688	52,7	3487	39,2
7	24401	13612	55,8	10778	44,2	770	7,1	5903	54,8	4105	38,1
8	18941	10885	57,5	8056	42,5	554	6,9	4374	54,3	3128	38,8
9	18010	10494	58,3	7516	41,7	440	5,9	4081	54,3	2995	39,8
10	15928	9010	56,6	6918	43,4	603	8,7	3607	52,1	2708	39,1
11	14251	8411	59,0	5840	41,0	341	5,8	2972	50,9	2527	43,3
12	15940	9201	57,7	6730	42,2	483	7,2	3630	53,9	2617	38,9
<b>Total</b>	<b>277769</b>	<b>154973</b>	<b>55,8</b>	<b>122774</b>	<b>44,2</b>	<b>9683</b>	<b>7,9</b>	<b>66650</b>	<b>54,3</b>	<b>46441</b>	<b>37,8</b>

# Results on a subset of 548 samples

- 44 monomorphic SNPs (2.8 %) over 548 samples
- 31 (2.1%) over 50% missing data
- 13 (2.1 %) samples over 50 % missing loci
- 535 samples x 1461 SNPs with 7.4 % missing data
- Diversity analysis - 2.1 % heterozygous



Axis 1: 51.9 %, Axis 2: 5.7 %



# Project Data Delivery, Impact, & Linkage

## *OryzaSNP*

- 259,721 non-redundant model-based SNPs
- Public release of annotation db (version 1)  
<http://www.oryzasnp.org>  
(October 16, 2007 at 12 pm Manila ST)
- Enables tag SNP assays for >95% genome
- Linkage to any rice mapping/gene discovery project

## *HaplOryza*

- 1536 indica/ japonica SNPs genotyped 548 varieties
- LD assessed in specific regions among certain types
- Remaining accessions genotyped in 2007
- Data delivered through TropGenes DB and OryzaSNP site

# The Future: OryzaSNP Phase 2

- **Develop High-Density Genotyping Arrays for WGS**
  - tag SNPs from Model-based and Machine-Learning calls
  - Design arrays with 4M features (new design Affy ala Arabidopsis):
    - 250 K tag SNPs oligos (allele specific)
    - 2 M tiling/expression oligos
    - 1 M CH<sub>3</sub>lation site oligos
  - Projected cost ~\$200/slide for 1500 slides
- **Genotype 500-1000 accessions over 3-5 yrs**
- **Phenotype for multiple traits**
- **Apply Association Genetics for allele discovery**
- **Partners under the Int.Rice Functional Genomics Consortium**

*Agropolis-CIRAD*

*Cornell*

*NYU*

*CNRRI*

*Yale*

*Academia-Sinica*

*NIG*

*RDA*

*+ IRRI, CSU, MSU (Buell), MPI-Tubingen ...*

*J.C. Glaszmann et al*

*S. McCouch, C. Bustamante*

*M. Purugganan*

*Q. Zhang*

*X.W. Deng*

*Y.I. Hsing*

*N. Kurata*

*Y.J. Park*