Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple

David Chagne´ [1], Ross N. Crowhurst [2], Michela Troggio[3], Mark W. Davey[4], Barbara Gilmore[5], Cindy Lawley[6], Stijn Vanderzande[4], Roger P. Hellens[2], Satish Kumar[7], Alessandro Cestaro[3], Riccardo Velasco[3], Dorrie Main[8], Jasper D. Rees[9], Amy Iezzoni[10], Todd Mockler[11], Larry Wilhelm[12], Eric Van de Weg[13], Susan E. Gardiner[1], Nahla Bassil[5], Cameron Peace[8]*

1 Plant and Food Research, Palmerston North Research Centre, Palmerston North, New Zealand, 2 Plant and Food Research, Mount Albert Research Centre, Auckland, New Zealand, 3 IASMA Research and Innovation Centre, Foundation Edmund Mach, San Michele all'Adige, Trento, Italy, 4 Laboratory for Fruit Breeding and Biotechnology, Department of Biosystems, Katholieke Universiteit Leuven, Heverlee, Leuven, Belgium, 5 USDA-ARS, National Clonal Germplasm Repository, Corvallis, Oregon, United States of America, 6 Illumina Inc., Hayward, California, United States of America, 7 Plant and Food Research, Hawke's Bay Research Centre, Havelock North, New Zealand, 8 Department of Horticulture and Landscape Architecture, Washington State University, Pullman, Washington, United States of America, 9 Agricultural Research Council, Onderstepoort, South Africa, 10 Department of Horticulture, Michigan State University, EastLansing, Michigan, United States of America, 11 The Donald Danforth Plant Science Center, St. Louis, Missouri, United States of America, 12 Oregon Health Sciences University, Portland, Oregon, United States of America, 13 Plant Breeding, Wageningen University and Research Centre, Wageningen, The Netherlands

Introduction

Understanding the link between "phenotypic variation" and their underlying "DNA variation"

Advances in genomic technologies (NGS, high throughput genotyping) – high resolution genetic characterization of plant germplasm possible

Genomics tools – potentials to increase genetic gains, more efficient and more precise breeding

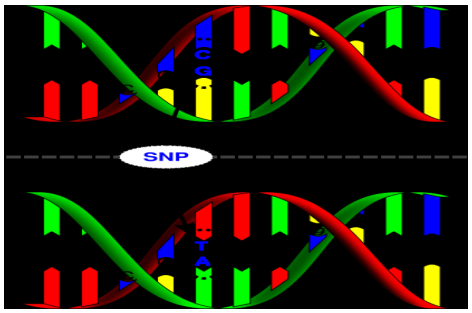
Genetic marker-based strategies (QTL, interval mapping, association mapping, genomic selection) powerly enabled by genomic technologies

SSR and SNP markers available in apple – low density– need to develop a higher density SNP markers for fine dissection of functional genetic variations

Development of medium-to-high throughput multiplexed SNP assays

SNP detection ⇒ verification ⇒ final selection

Genomics Research For Breeding Purposes



Sequencing

(*De novo*, re-sequencing)

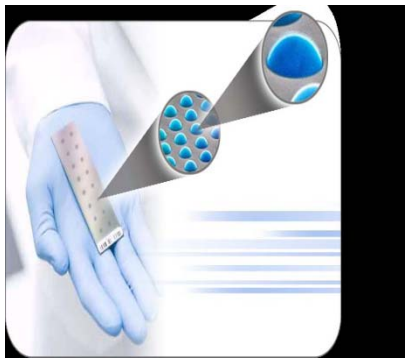
SNP Discovery

(and Validation)

Illumina HiSeq2000

Low-High Density SNP Chips

(96, 384, 1536, HD up to 1 millions SNPs)



Marker-Trait Association

(Gene and QTL tagging)

Molecular Breeding

(Marker-assisted selection, Marker-assisted back crossing)

Illumina iScan

Illumina's genetic platforms

Genome Analyzer HiSeq2000
(Solexa technology)

Next gen sequencing

- Whole-genome *de novo* → Whole-genome resequencing
- Transcriptome analysis
- Targeted (reduced representative) sequencing

HIGHEST OUTPUT

Capable of up to
600 GB per run

FASTEST DATA RATE
~ 25 Gb/day
7-8 days for 2 x 100 bp

iScan
(BeadArray Reader)

- GoldenGate: Sentrix (SAM) → 96, 384, 768, 1536 SNP genotyping
- Infinium BeadChip
  - iSelect → 3k-200k Custom SNP chips
  - Human → Up to 1 million SNPs/chip

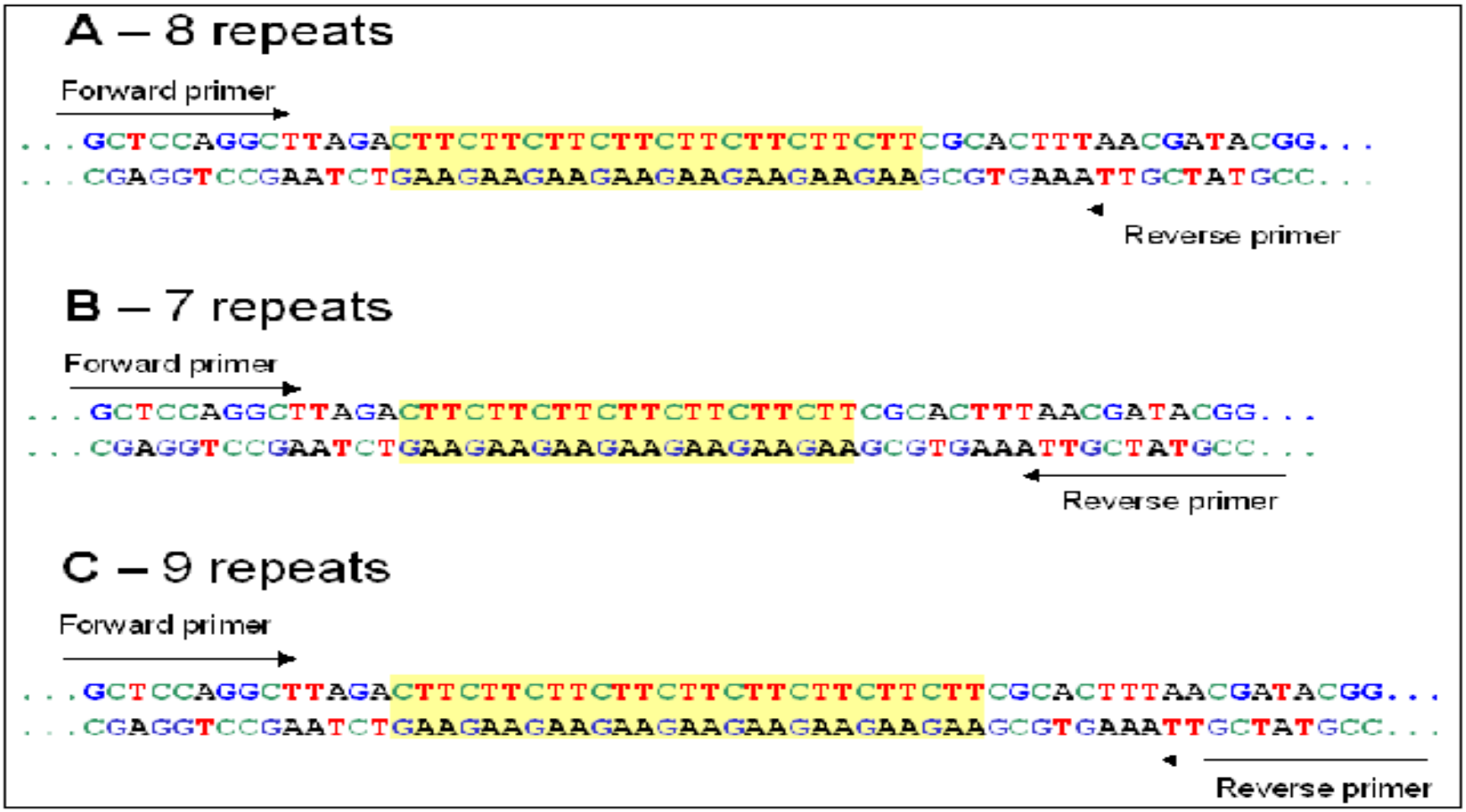
Commercial SNP Chips Available

Bovine SNP50 and HD
Ovine SNP50 and HD
Canine SNP50 and HD
Maize SNP50 and HD, etc.

DNA markers for breeding

- Simple sequence repeat (SSR)
- Single nucleotide polymorphism (SNPs)

Simple Sequence Repeat (SSR)

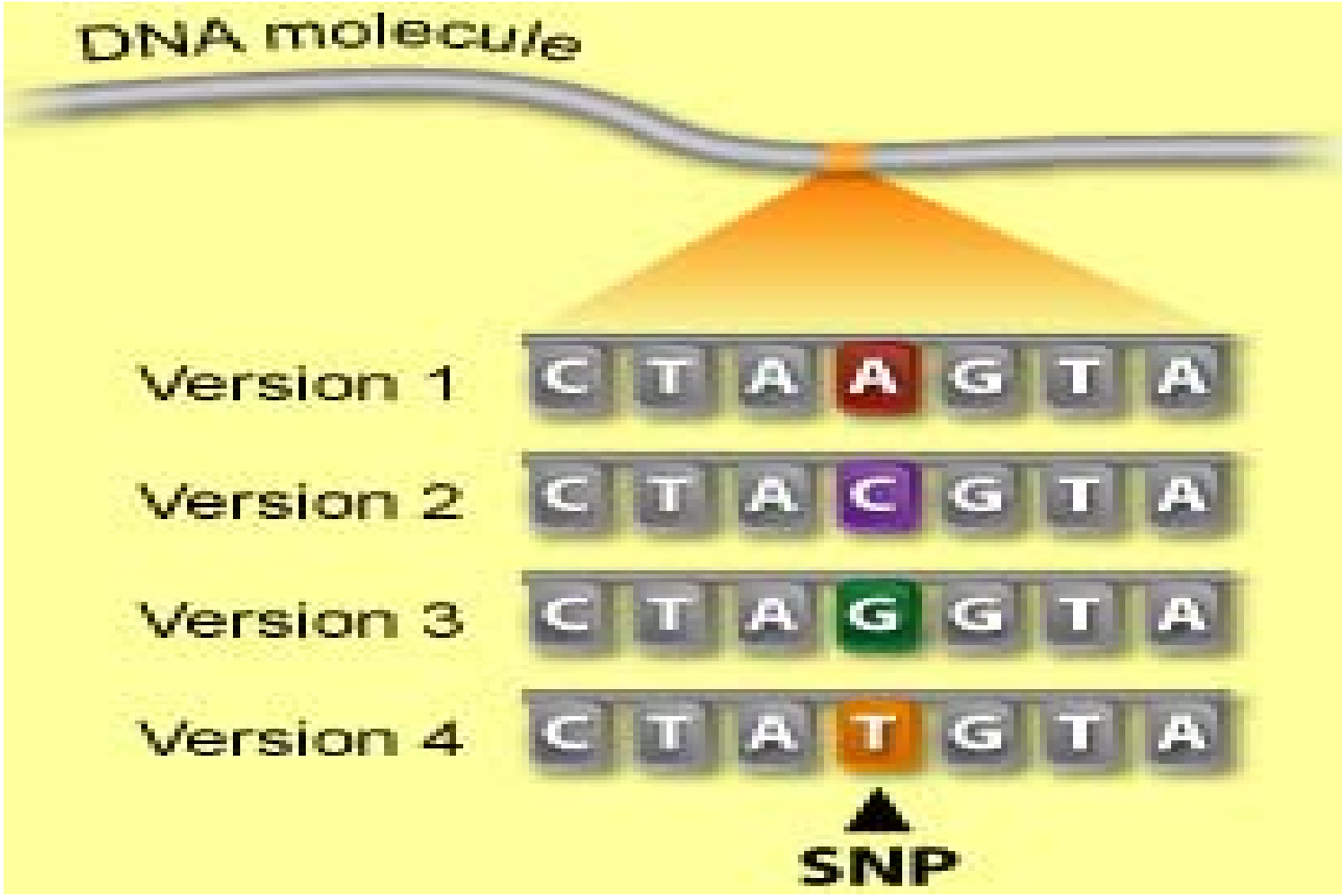


The number is in term of thousands SSR markers detected in the genome

SSR markers have some disadvantages

- High polymorphism rate, but having “many alleles” makes precise scoring difficult
- SSR data is “difficult to merge across labs and groups”
- Not easy to run in a high-throughput system due to “limitations in multiplex levels”

Single Nucleotide Polymorphism (SNP)



The number is in terms of millions SNP markers detected in the genome

SNP facts

- SNPs are found in
  - coding and (mostly) noncoding regions.
- Occur with a very high frequency
  - about 1 in 1000 bases to 1 in 100 to 300 bases.
- The abundance of SNPs and the ease with which they can be measured make these genetic variations significant.
- SNPs close to particular gene acts as a marker for that gene.
- SNPs in coding regions may alter the protein structure made by that coding region.

Soybean SNP detection

Type (alphabetical order)	***Count***	***Percent***
DOWNSTREAM	1,174,917	26.637%
EXON	95,154	2.157%
INTERGENIC	1,393,216	31.586%
INTRON	381,701	8.654%
NONE	16	0%
SPLICE_SITE_ACCEPTOR	537	0.012%
SPLICE_SITE_DONOR	606	0.014%
UPSTREAM	1,312,546	29.757%
UTR_3_PRIME	36,219	0.821%
UTR_5_PRIME	15,985	0.362%

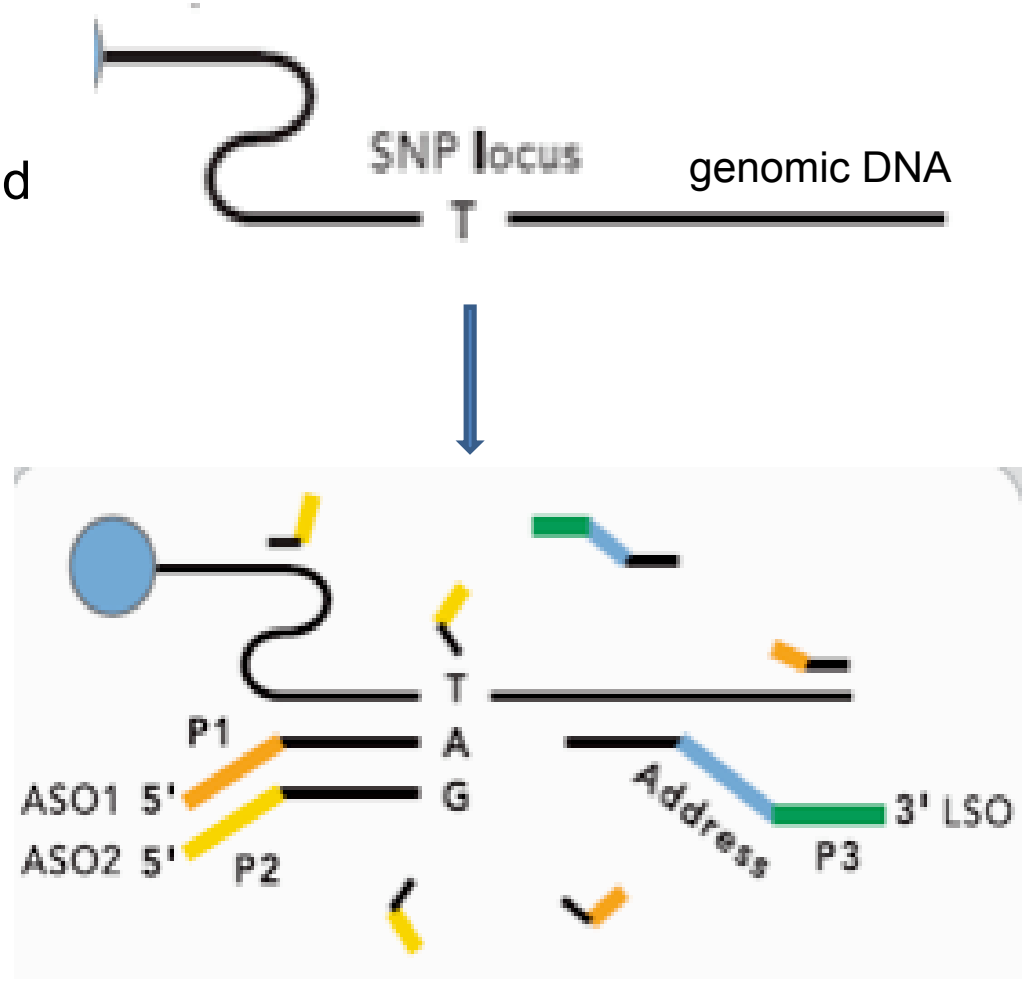
Chromosome	*Length*	*Changes*	*Change rate*
Gm01	55,915,595	164,050	340
Gm02	51,656,713	140,745	367
Gm03	47,781,076	182,784	261
Gm04	49,243,852	122,041	403
Gm05	41,936,504	102,278	410
Gm06	50,722,821	207,015	245
Gm07	44,683,157	137,522	324
Gm08	46,995,532	156,487	300
Gm09	46,843,750	152,358	307
Gm10	50,969,635	91,056	559
Gm11	39,172,790	115,306	339
Gm12	40,113,140	99,504	403
Gm13	44,408,971	168,210	264
Gm14	49,711,204	169,969	292
Gm15	50,939,160	218,143	233
Gm16	37,397,385	194,427	192
Gm17	41,906,774	164,865	254
Gm18	62,308,140	280,791	221
Gm19	50,589,441	148,527	340
Gm20	46,773,167	111,122	420
Total	972,068,482	3,150,869	308

Cocoa SNP detection

Chromosome	Length	Changes	Change rate
Tc00	108,886,888	835,679	130
Tc01	31,268,538	254,505	122
Tc02	27,754,001	232,265	119
Tc03	25,475,297	198,549	128
Tc04	23,504,306	166,738	140
Tc05	25,651,337	208,396	123
Tc06	15,484,475	138,079	112
Tc07	14,169,093	106,745	132
Tc08	11,535,834	111,952	103
Tc09	28,459,094	284,016	100
Tc10	15,164,258	151,245	100
Total	**327,353,121**	**2,688,169**	**121**

SNPs are now the marker of choice

- SNPs are “abundant” across the genome
- Large “pools of SNPs” can be used to identify sets of polymorphic markers
- SNP markers are “bi-allelic” making allele calling more simple
- SNP data from different systems or groups can be “easily merged in a database”
- SNP genotyping “can be automated”, allowing for rapid, high-throughput marker genotyping



SNP genotyping with allele specific oligos

Sequencing

“SNP discovery”

Why Sequencing?

(1). Elucidating the genome sequence of important crops and life-stock

(a) The “reference sequence” is the all important map where all subsequent research work is based upon.

(b) The genetic sequence is especially important in mapping out important genetic regions responsible for economically important traits.

Why Sequencing?

(2). SNP discovery

One of the main goals for sequencing is the discovery of SNPs common and specific to a species or between sub-species. The discovery and comparison of these SNPs improves the understanding of genes and markers responsible for a trait or phenotypic manifestation. These markers can then be used as proxies for Marker Assisted Selection (MAS)

(3). Discovering novel biochemical pathways via transcriptome analysis

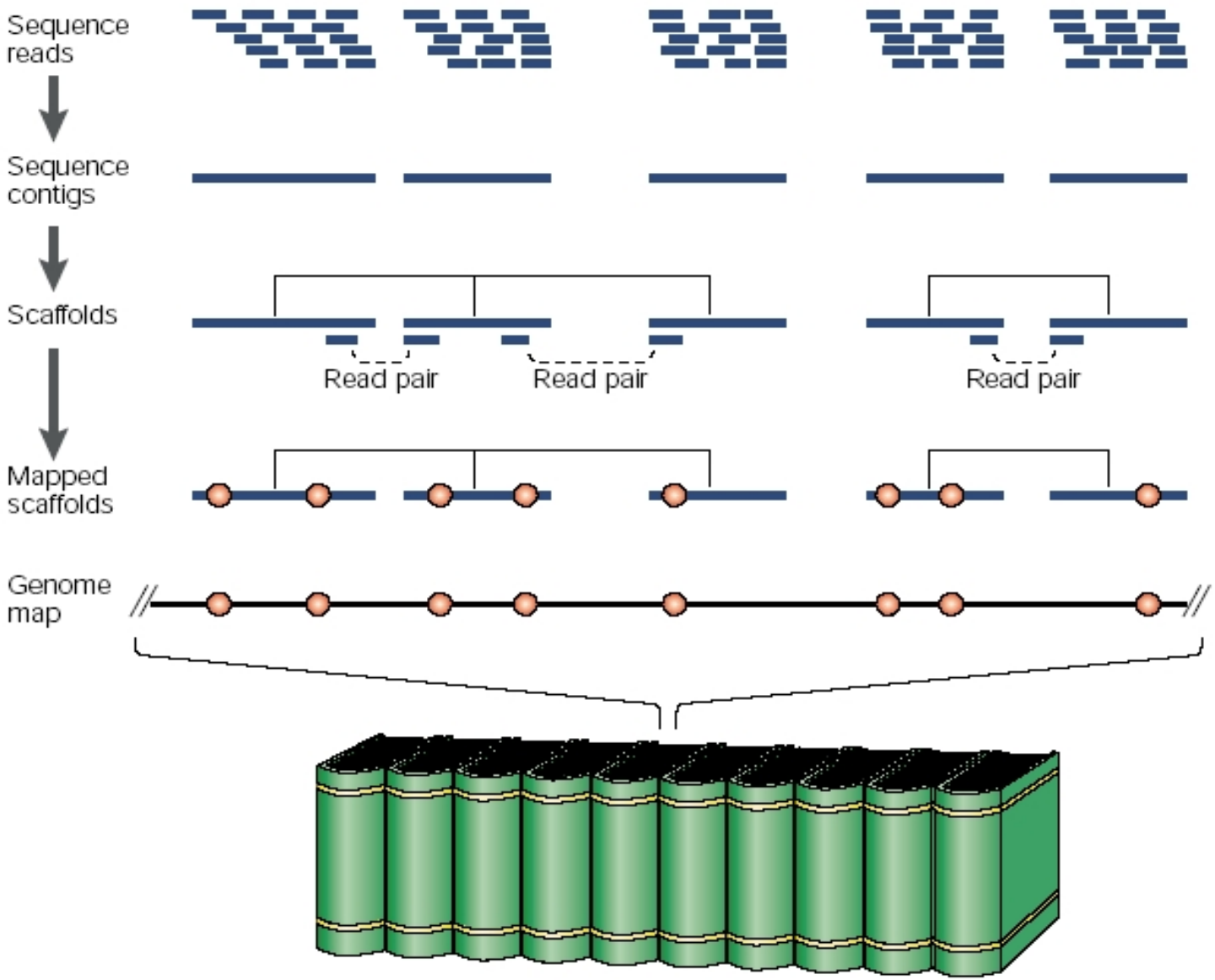
Transcriptome sequencing has been used to measure transcription profiles which could help correlate gene expression with function.

(4). Epigenetics

Methylation sequencing can be used to determine epigenetic control and elucidate how genes are transcriptionally regulated.

De novo sequencing
vs
Re-sequencing

De novo sequencing



Re-sequencing

- Reference genome sequence available
- Alignment of the re-sequence data to detect genome variation (e.g. SNP)

Biogen Soybean Resequencing Project

- *Resequencing* 5 genotipe kedelai: B3293, Davros, Grobogan, Malabar dan Tambora
- Sekuensing dilakukan dengan menggunakan alat Illumina HiSeq 2000
- Data sekuen rujukan genom kedelai (Ukuran genom 1,11 milyar basa, 950 juta diantaranya telah disekuensing) diperoleh dari:

http://www.phytozome.org/cgi-bin/gbrowse/soybean/

Plants and animals with the reference genome sequence available

Plants/Animals	Genome size (Mb)		Plants/Animals	Genome size (Mb)
Arabidopsis thaliana	125		*Vitis vinifera*	487
Oryza sativa	389		*Carica papaya*	32
O. glaberima	357		*Ricinus communis*	400
Zea mays	2 600		*Manihot esculenta*	770
Sorghum bicolor	770		*Jatropha curcas*	372
Glycine max	1 100		*Lotus japonicus*	472
Cucumis sativus	367		*Brachypodium distachion*	300
Solanum tuberosum (DH)	840		*Theobroma cacao*	500
S. lycopersicum	950		*Populus trichocarpa*	485
Medicago trancatula	500		*Bos taurus*	3000
Musa accuminata (DH Pahang; AA)	823		*Capsicum annuum*	3.500

Web-based crop genome reference sequences accessable to public

Soybean

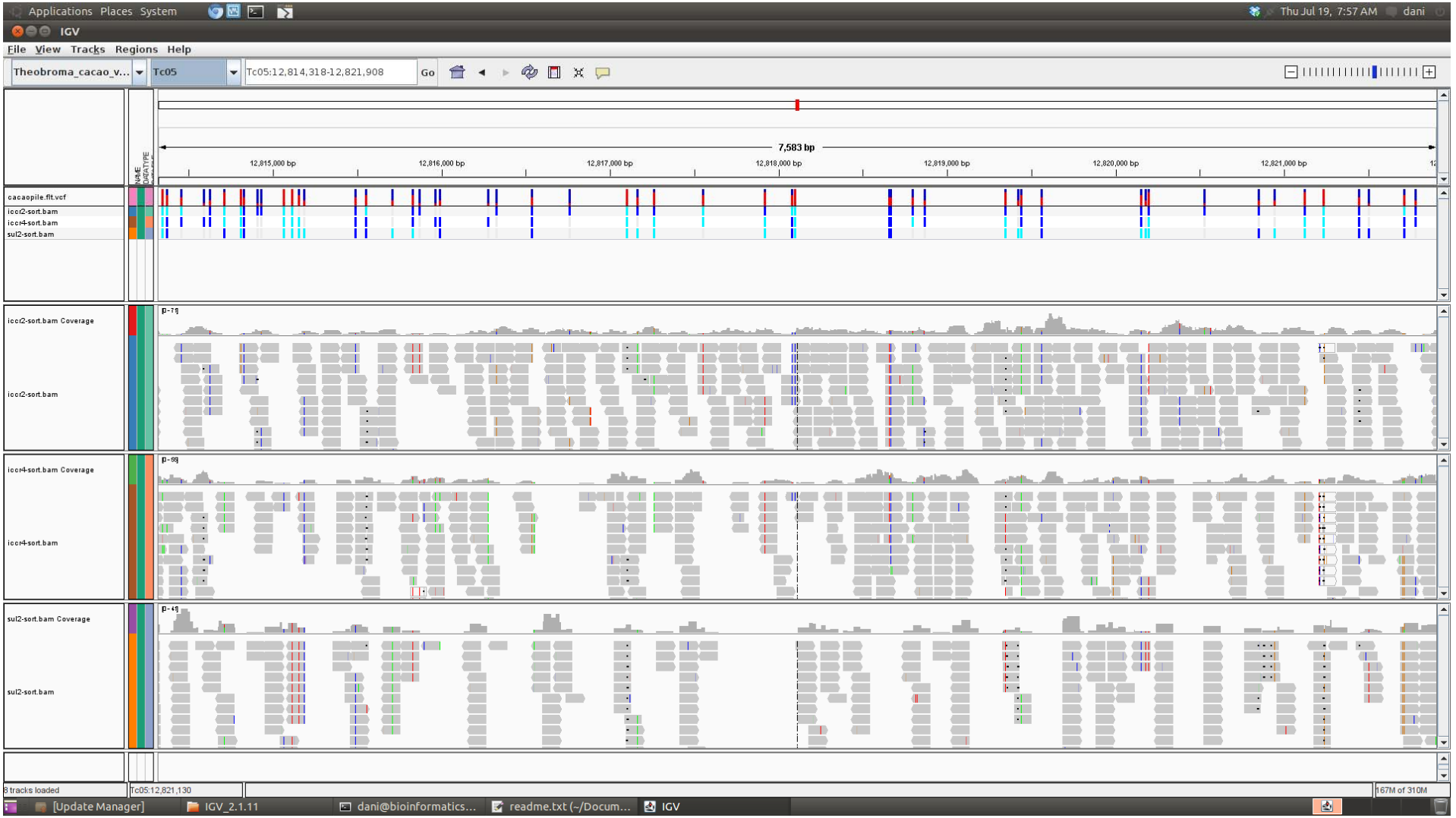
http://www.phytozome.org/cgi-bin/gbrowse/soybean/

Maize http://www.maizegdb.org/

Banana http://banana-genome.cirad.fr

Potato http://www.potatogenome.net

Perbandingan Hasil Resequencing 5 Genotipe Kedelai dengan Sekuen Genom Rujukan Kedelai Varietas Williams 82



SNP discovery, an example

OryzaSNP project:

Genome-wide SNP variation in landraces and modern rice varieties

Perlegen high density oligomer arrays on 20 varieties:

- Discovery of single nucleotide polymorphisms (SNPs)
- 5 in. X 5 in. arrays each interrogating 25 Mb using over 160 million separate features

OryzaSNP consortium:

PERLEGEN SCIENCES

Generation
Cultivating Plant Diversity for the Resource-Poor

CSHEES

Colorado State University

TIGR

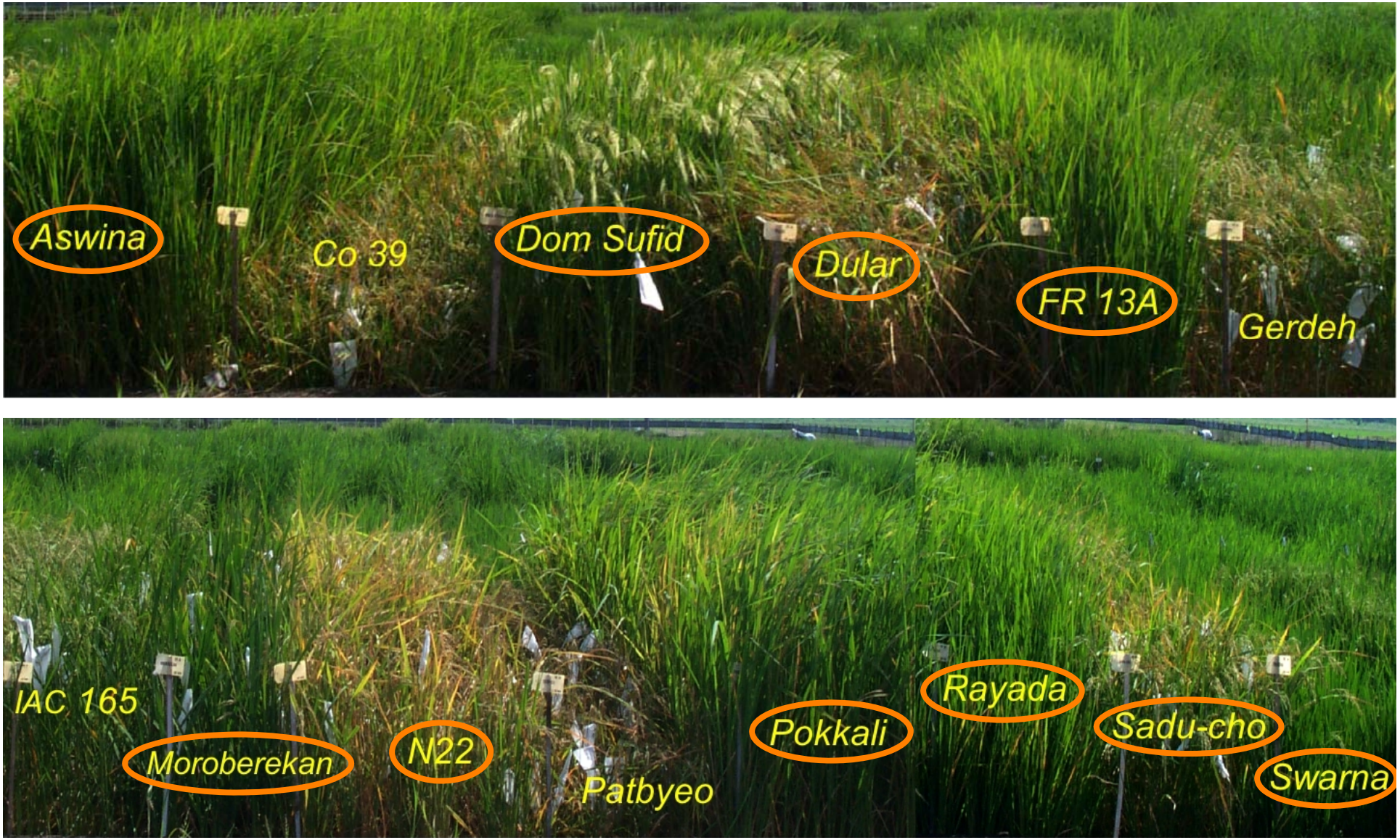
MAX-PLANCK-GESELLSCHAFT

CORNELL

McGill

MICHIGAN STATE UNIVERSITY

Wide range of plant types used in OryzaSNP set



What did they do in this study

A next-generation sequencing (NGS) was used to detect SNPs covering the apple genome.

The effort involved re-sequencing a small set of cultivars, ancestors, and founders, chosen to represent the pedigrees of worldwide apple breeding programs by RosBREED, a consortium established to enable marker-assisted breeding for Rosaceae crops (www.rosbreed.org).

They validated the SNPs detected and determined adequate filtering parameter values, using the Illumina GoldenGateH assay to screen a larger set of accessions from the international apple breeding germplasm.

Based on results of the GoldenGateH assay, they further refined SNP-filtering criteria to enable the development of an 8K InfiniumH II array and evaluated it, using a segregating population of apple seedlings.

Materials and Methods

Whole genome re-sequencing of apple breeding accessions

27 apple germplasm accessions was chosen for whole-genome, low coverage re-sequencing.

The accessions were founders, intermediate ancestors, or important breeding parents used extensively in apple breeding programs worldwide.

Table 1. Apple cultivars used for low coverage re-sequencing and subsequent single nucleotide polymorphism detection.

Accession	Number of reads	Type of reads	Estimated genome coverage (X)	Location of sequencing
'Braeburn'	34,699,040	Paired-end	6.9	PFR
'Co-op 15'	51,864,610	Single-end	5.2	ARC
'Cox's Orange Pippin'	35,120,961	Paired-end	7.0	PFR & ARC
'Crimson Crisp'	6,128,575	Paired-end	1.3	RosBREED
'Cripps Pink'	56,738,286	Single-end	6.1	ARC
'Delicious'	17,995,052	Paired-end	1.9	RosBREED
'Dolgo'	59,863,282	Single-end	6.4	ARC
'Duchess of Oldenburg'	17,762,643	Paired-end	3.8	RosBREED
F_2 26829-2-2	31,580,732	Paired-end	6.7	RosBREED
'Frostbite'	44,740,552	Single-end	4.5	ARC
'Fuji'	44,545,764	Single-end	4.8	ARC
'Geneva'	31,977,525	Paired-end	6.4	PFR
'Golden Delicious'	22,101,159	Paired-end	4.7	RosBREED
'Granny Smith'	36,112,204	Paired-end	7.2	PFR
'Haralson'	15,499,148	Paired-end	3.3	RosBREED
'Honeycrisp'	20,887,451	Paired-end	4.5	RosBREED
'Idared'	33,479,444	Paired-end	6.7	PFR
'James Grieve'	26,005,304	Paired-end	5.2	PFR
'Jonathan'	18,812,375	Paired-end	4.0	RosBREED
'McIntosh'	59,934,507	Paired-end	12.8	RosBREED
Malus sieversii PI613981	23,500,375	Paired-end	4.7	USDA-ARS
'Ralls Janet'	35,485,026	Paired-end	7.1	PFR
'Red Delicious'	25,993,982	Paired-end	5.2	PFR
'Red Dougherty'	32,718,666	Paired-end	6.5	PFR
'Rome Beauty'	23,220,046	Paired-end	5.0	RosBREED
'Splendour'	36,353,918	Paired-end	7.3	PFR
'Zestar'	54,404,792	Single-end	5.4	ARC

An approximate genome coverage was estimated for each accession using a genome size of 750 Mb. The source of cultivars and numbers of Illumina GA II reads obtained are indicated. ARC: Agricultural Research Council, South Africa; PFR: The New Zealand Institute for Plant & Food Research Ltd; RosBREED: U.S.-based international project.

doi:10.1371/journal.pone.0031745.t001

Detection and filtering of SNPs

SNPs were detected using SoapSNP (http://soap.genomics.org.cn/soapsnp.html) essentially as described by Wang et al. (2008) with modification.

Apple genome reference sequence of ‘Golden Delicious’ (GD)

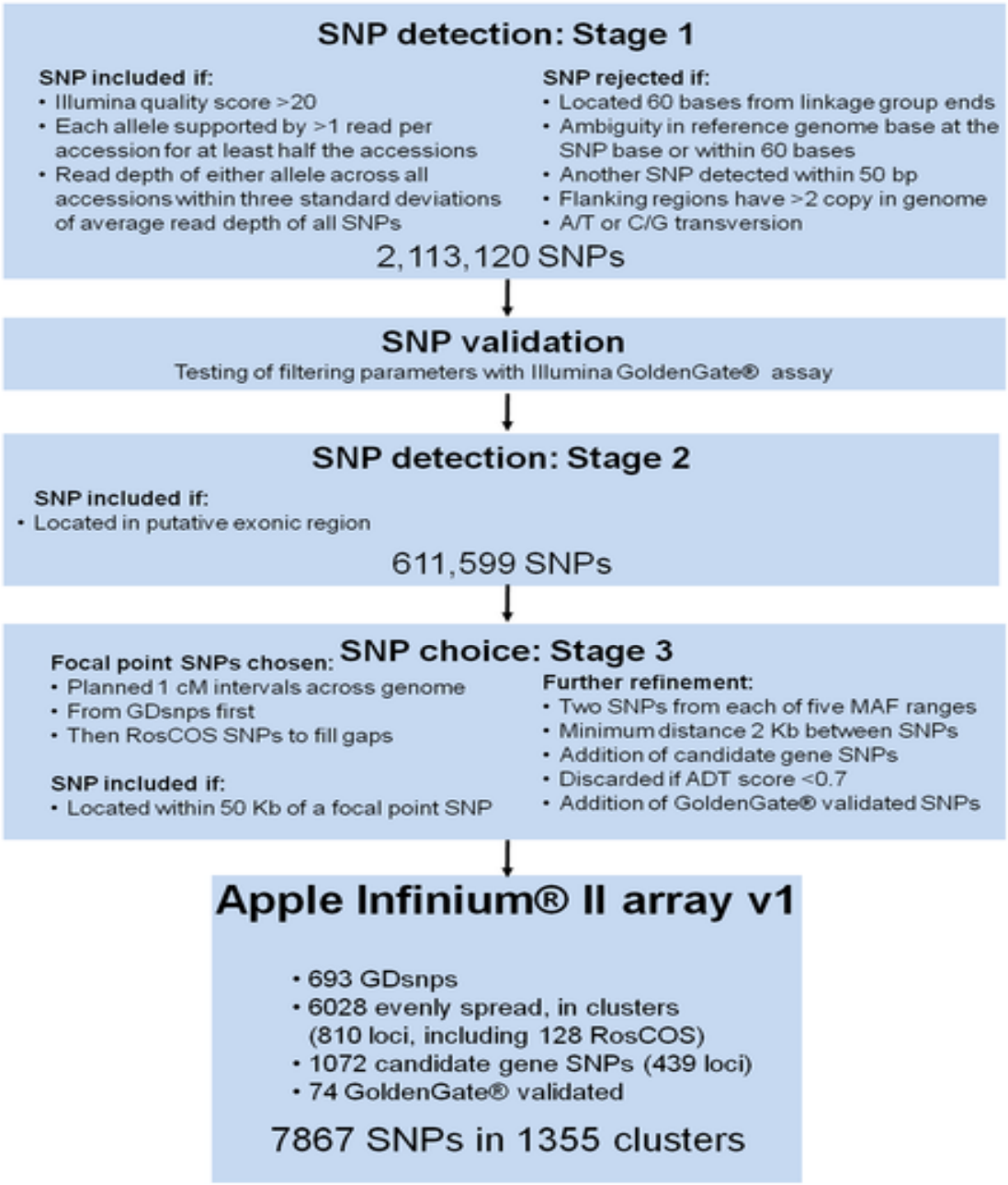


Figure 1. Workflow for single nucleotide polymorphism (SNP) detection, validation, and final choice employed for development of the IRSC apple 8K SNP array v1.

(Stage 1 filtering):

SoapSNP calls were discarded if

(1) the SNP call was within the first or last 60 bases of a linkage group (LG) sequence,
(2) the reference genome base was ambiguous at the SNP position or within 60 bases of it,
(3) Another SNP was detected within 50 bases,
(4) the average copy number of the SNP flanking region was more than two,
(5) the Illumina quality score of either allele was less than 20,
(6) the number of reads supporting either allele was less than two reads per accession for at least half the accessions,
(7) the number of reads for either allele across all accessions was greater than the average read depth of all SNPs plus three standard deviations, and
(8) if the call was an A/T or C/G transversion.
(9) This filtration yielded ‘‘Stage 1 SNPs’’.

(Stage 2 filtering):

Stage 1 SNPs were then subjected to a Stage 2 filter, whereby SNPs not located in a predicted exonic region were discarded.

For each LG, the exon space was defined by mapping all gene models (Velasco et al. 2010) and 396,643 cDNA sequences from the Plant & Food Research Malus Sequence Database libraries (Newcomb et al. 2006)

(‘AAAA’,‘AABA’, ‘AACA’, ‘AADA’, ‘AAEA’, ‘AAMA’, ‘AAFA’, ‘AAFB’,‘AAGA’, ‘AAHA’, ‘AAIA’, ‘AAJA’, ‘AAKA’, ‘AALA’, ‘AALB’,‘AANA’, ‘AAOA’, ‘AAPA’, ‘AAQA’, ‘AARA’, ‘AASA’, ‘AAUA’,‘AAVB’, ‘AVBC’, ‘AAWA’, ‘AAXA’, ‘AAYA’, ‘AAZA’, ‘AATA’,…….

SNP validation with GoldenGate® assay

A subset of 144 SNPs was chosen to validate the efficiency of SNP detection and fine-tune the filtering parameters (Figure 2).

(1). Evenly spread SNPs across the 17 apple chromomosomes (100)

(2). SNPs associated with malic acid content in fruit, *Ma* locus (20)

(3). SNPs within exons of candidate genes (27)

(4). Previously identified Golden Delicious (GD) SNPs (8)

(5). Accession specifics (28)

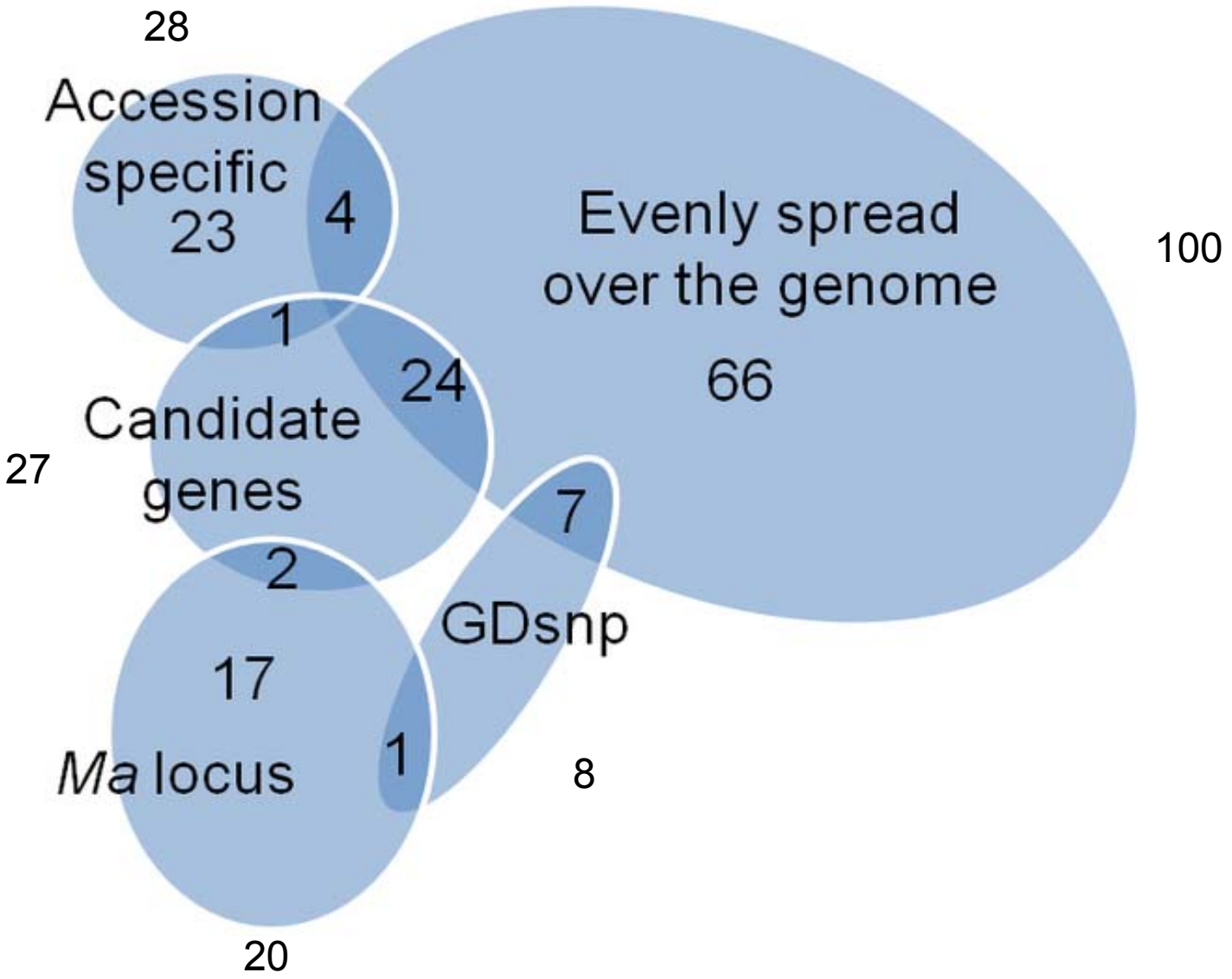


Figure 2. Classification of the 144 apple single nucleotide polymorphisms (SNPs) used for validation using the Illumina GoldenGate® assay.

SNP final choice for 8K Infinium® II array

A clustering strategy was devised that would evenly span the genetic map of apple with clusters of exonic SNPs, in order to provide a final SNP genome scan with the capability of determining SNP haplotypes at distinct loci.

IASMA-FEM ‘Golden Delicious’X‘Scarlet’ reference genetic map (http://genomics.research.iasma.it/cgi-bin/cmap/viewer),

Rosaceae Conserved Orthologous Set (RosCOS; [23]) loci to fill genetic gaps between GDsnps, according to physical map location [24].

SNP final choice for 8K Infinium® II array (continued)

(1). The design featured focal points at approximately 1 cM intervals with 4–10 SNPs clustered at each focal point, minimal distance between SNPs was set at 2 Kb

(2). Chromosome ends, the first 200 Kb of each LG according to the apple draft genome pseudo-chromosomes [11], were also targeted

(3). SNPs were binned according to MAF (bins of 0.1, 0.2, 0.3, 0.4, and 0.5 corresponding to MAF ranges of 0.01–0.1, 0.101–0.2, and so on).

(4). Additional SNP clusters were developed from candidate genes for fruit quality, tree architecture, and flowering that were chosen using the available literature [25,26,27]. Up to 4 SNPs were chosen within a candidate gene

SNP array evaluation and cluster file development

A set of populations from various crosses, as well as accessions from the apple germplasm, were used to evaluate the Apple 8K Infinium® II array.

A ‘Royal Gala’X‘Granny Smith’ F1 population of 186 seedlings [28], seven controlled F1 comprise 1313 of individuals [29], a set of 117 accessions from the Plant & Food Research germplasm collection

‘Royal Gala’, ‘Granny Smith’, and ‘Golden Delicious’ were used as controls

SNP array evaluation and cluster file development (continued)

200 ng gDNA were used as template for the reaction, following the manufacturer's instructions (Illumina Infinium Assay)

SNP genotypes were scored with the Genotyping Module of the GenomeStudio Data Analysis software (Illumina Inc., San Diego, CA).

Individuals with low SNP call quality (p50GC,0.54), as well as seedlings putatively resulting from an unintended pollination, were removed from the analysis.

SNPs with a GenTrain score 0.6 were retained and those with scores ranging between 0.3 and 0.6 were visually checked for accuracy of the SNP calling.

Clusters were manually edited when the parent-offspring segregation was not correct, or when the number of missing genotypes was greater than 20.

SNP array evaluation and cluster file development (continued)

"Two trios" with both parents and one seedling were used to test the usefulness of SNP clusters for identifying haplotypes:

'Royal Gala' X 'Braeburn' → 'Scifresh' and
'(Royal) Gala' X 'Splendour' → 'Schiros'

SNPs were coded using A and B alleles and haplotypes were inferred using FlexQTLTM (www.flexqtl.nl) for each cluster of SNPs.

Results and Discussion

NGS re-sequencing of apple breeding accessions and SNP detection

(1). A total of 67 Gb DNA sequence from 898 millions 75-base reads were generated for the 27 apple accessions (Table 1).

(2). The total sequences obtained among the accessions ranged from 6 to 60 millions (due to differences in cluster density)

(3). A total of 10,915,756 SNPs was detected using *SoapSNP* of which 2,113,120 SNPs (19.5%) passed the filtering criteria for Stage 1 detection (Table 2; Figure 1)

(4). The average SNP frequency was one per 288 bp. Of these, 611,599 (28.9%) were predicted located in the exonic regions and passed the Stage 2 filter (Table 2; Figure 1).

Table 1. Apple cultivars used for low coverage re-sequencing and subsequent single nucleotide polymorphism detection.

Accession	Number of reads	Type of reads	Estimated genome coverage (X)	Location of sequencing
'Braeburn'	34,699,040	Paired-end	6.9	PFR
'Co-op 15'	51,864,610	Single-end	5.2	ARC
'Cox's Orange Pippin'	35,120,961	Paired-end	7.0	PFR & ARC
'Crimson Crisp'	6,128,575	Paired-end	1.3	RosBREED
'Cripps Pink'	56,738,286	Single-end	6.1	ARC
'Delicious'	17,995,052	Paired-end	1.9	RosBREED
'Dolgo'	59,863,282	Single-end	6.4	ARC
'Duchess of Oldenburg'	17,762,643	Paired-end	3.8	RosBREED
F_2 26829-2-2	31,580,732	Paired-end	6.7	RosBREED
'Frostbite'	44,740,552	Single-end	4.5	ARC
'Fuji'	44,545,764	Single-end	4.8	ARC
'Geneva'	31,977,525	Paired-end	6.4	PFR
'Golden Delicious'	22,101,159	Paired-end	4.7	RosBREED
'Granny Smith'	36,112,204	Paired-end	7.2	PFR
'Haralson'	15,499,148	Paired-end	3.3	RosBREED
'Honeycrisp'	20,887,451	Paired-end	4.5	RosBREED
'Idared'	33,479,444	Paired-end	6.7	PFR
'James Grieve'	26,005,304	Paired-end	5.2	PFR
'Jonathan'	18,812,375	Paired-end	4.0	RosBREED
'McIntosh'	59,934,507	Paired-end	12.8	RosBREED
Malus sieversii PI613981	23,500,375	Paired-end	4.7	USDA-ARS
'Ralls Janet'	35,485,026	Paired-end	7.1	PFR
'Red Delicious'	25,993,982	Paired-end	5.2	PFR
'Red Dougherty'	32,718,666	Paired-end	6.5	PFR
'Rome Beauty'	23,220,046	Paired-end	5.0	RosBREED
'Splendour'	36,353,918	Paired-end	7.3	PFR
'Zestar'	54,404,792	Single-end	5.4	ARC

An approximate genome coverage was estimated for each accession using a genome size of 750 Mb. The source of cultivars and numbers of Illumina GA II reads obtained are indicated. ARC: Agricultural Research Council, South Africa; PFR: The New Zealand Institute for Plant & Food Research Ltd; RosBREED: U.S.-based international project.

doi:10.1371/journal.pone.0031745.t001

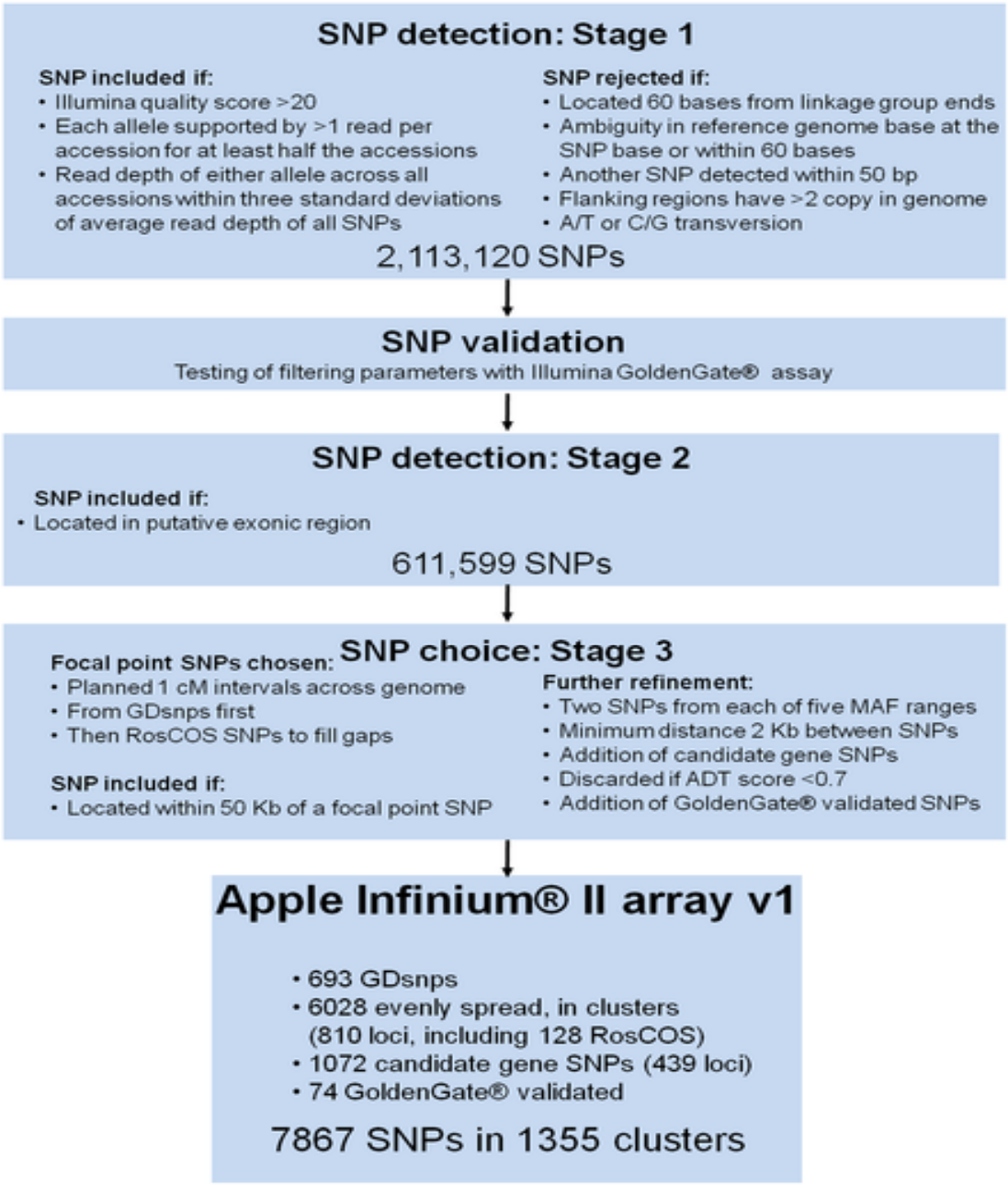


Figure 1. Workflow for single nucleotide polymorphism (SNP) detection, validation, and final choice employed for development of the IRSC apple 8K SNP array v1.

Table 2. Single nucleotide polymorphisms (SNPs) detected across the apple genome by re-sequencing 27 apple accessions.

Chromosome	Sequence used for SNP detection (bp)	SNPs examined from *SoapSNP*	SNPs passing "Stage 1" filtering	Average distance between SNPs (bp)	Exonic SNPs
1	36,084,648	556,642	102,825	350	31,678
2	40,172,783	793,583	144,715	277	42,496
3	39,907,579	700,701	134,595	296	36,005
4	25,411,901	474,397	91,212	278	27,540
5	37,603,833	661,644	134,672	279	38,400
6	30,670,413	514,972	97,014	316	29,450
7	31,181,013	546,126	117,652	265	33,155
8	35,800,717	625,020	114,173	313	31,946
9	37,514,065	737,124	146,472	256	41,935
10	38,388,612	687,313	146,102	262	43,522
11	40,097,013	751,789	146,718	273	42,681
12	36,276,268	664,326	121,739	297	35,456
13	39,686,055	720,281	158,912	249	43,699
14	34,156,235	607,215	114,259	298	32,712
15	55,775,419	919,506	159,368	349	46,218
16	23,462,870	421,197	85,530	274	24,357
17	27,122,502	533,920	97,162	279	30,349
Total	609,311,926	10,915,756	2,113,120	288	611,599

doi:10.1371/journal.pone.0031745.t002

SNP validation

(1). In GoldenGate® validation assay, 148 apple accessions in the test panel gave good quality scores (call rate > 0.8, and 10% GC score > 0.5 and 12 accessions were failed due to low DNA quality

(2). 73 (50.7%) SNPs were polymorphic, 46 (31.9%) SNPs failed reactions, 25 (17.5%) were monomorphic (MAF<0.05 or A/B frequency <0.1) (Table 3)

(3). 8 GDsnps, evenly spread SNPs, SNPs at *Mla* locus, SNPs within candidate genes showed a good to very good results

(4). 28 SNP chosen based on accession specificity, 14 (50%) had a MAF < 0.05; of which 11 were not polymorphic

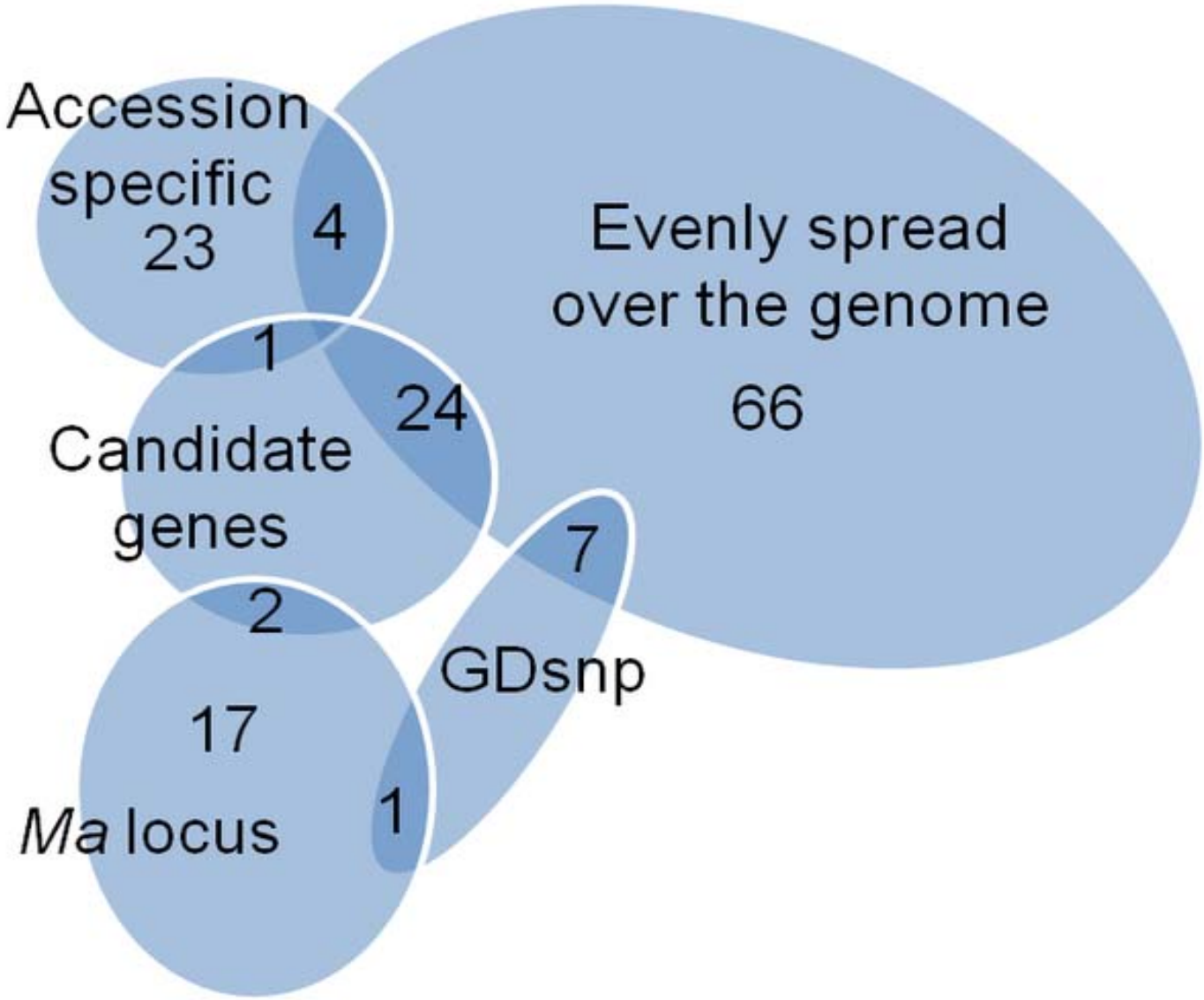


Figure 2. Classification of the 144 apple single nucleotide polymorphisms (SNPs) used for validation using the Illumina GoldenGate® assay.

Table 3. Results from a GoldenGate® assay of 144 single nucleotide polymorphisms (SNPs) screened over 160 apple accessions (Table S1).

		Proportion of SNPs			
SNP type	Total	Failed	Mono-morphic	Polymorphic (MAF<0.05)	Polymorphic (MAF>0.05)
Evenly spread	100	0.32	0.08	0.04	0.56
Accession-specific	28	0.25	0.11	0.39	0.25
Candidate genes	27	0.22	0.11	0.04	0.63
Ma locus	20	0.40	0.00	0.05	0.55
GDsnp	8	0.00	0.00	0.00	1.00
ADT score <0.8	40	0.35	0.00	0.05	0.60
ADT score 0.8–0.9	35	0.40	0.06	0.03	0.51
ADT score <0.9	69	0.26	0.12	0.17	0.45
Mean ADT score (s.d.)	0.85 (0.13)	0.83 (0.12)	0.94 (0.04)	0.92 (0.10)	0.83 (0.14)
Total	144	0.32	0.07	0.10	0.51

doi:10.1371/journal.pone.0031745.t003

SNP final choice

(1). Third stage filtering: focal points 1 cM intervals across the apple genome, choosing SNPs in 100 kb windows initially based on 712 GDsnp markers (Figure 3) for a total of 842 focal points

(2). 6074 SNPs based on focal points + 712 GDsnp + 1652 candidate gene SNPs = 8438 SNPs chosen by this stage

(3). Using ADT scores reduced the pool to 7793 SNPs: 693 GDsnps, 6028 SNPs arround focal points, 1072 SNPs within canditate genes (Figure 1)

(4). 74 SNP validated from GoldenGate® assay were also included for a ground total of 7867 SNPs in 1355 clusters for construction of the final Infinium® II SNP array, officially named the "International RosBREED SNP consortium (IRSC) apple 8K SNP array v1".

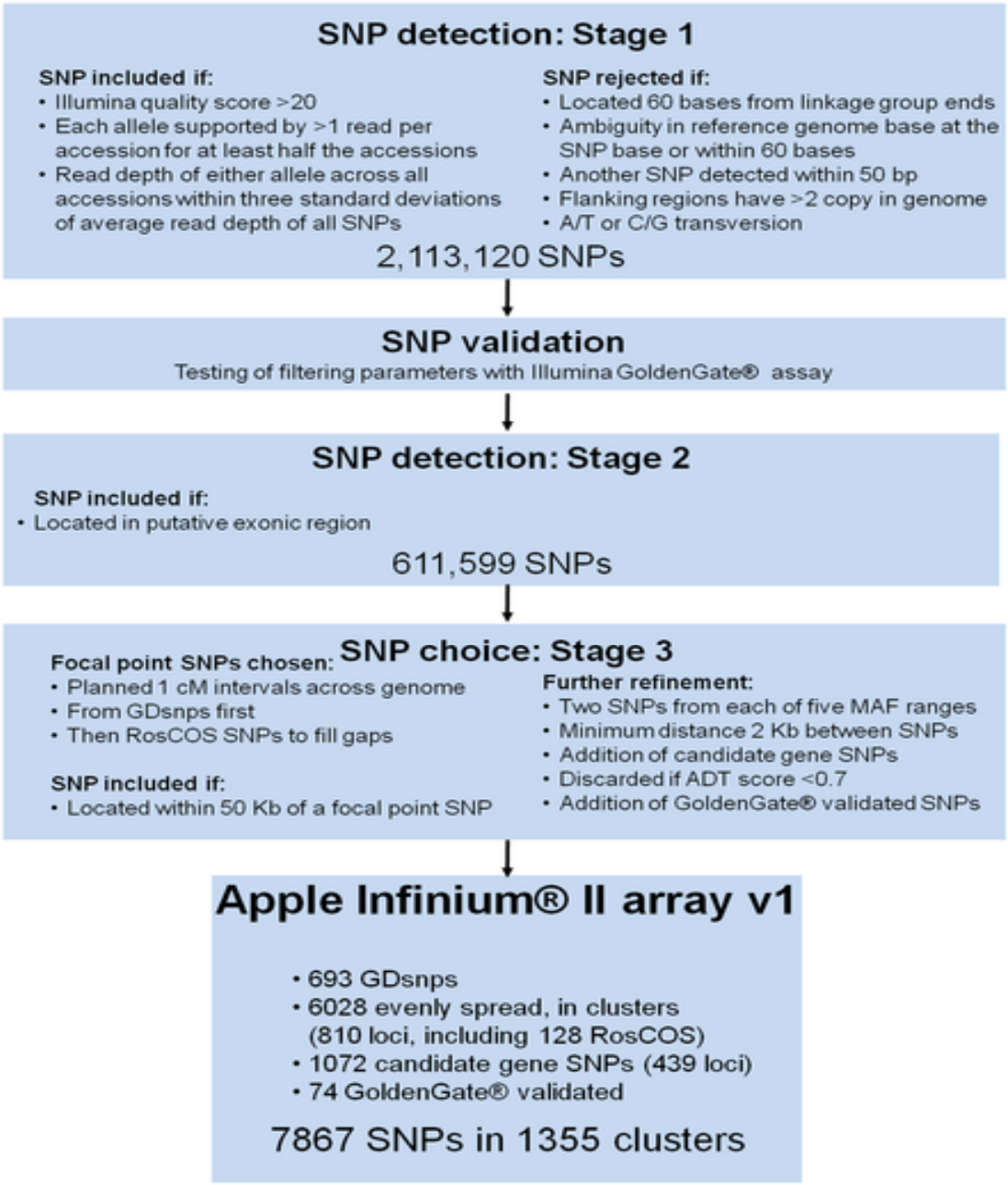


Figure 1. Workflow for single nucleotide polymorphism (SNP) detection, validation, and final choice employed for development of the IRSC apple 8K SNP array v1.

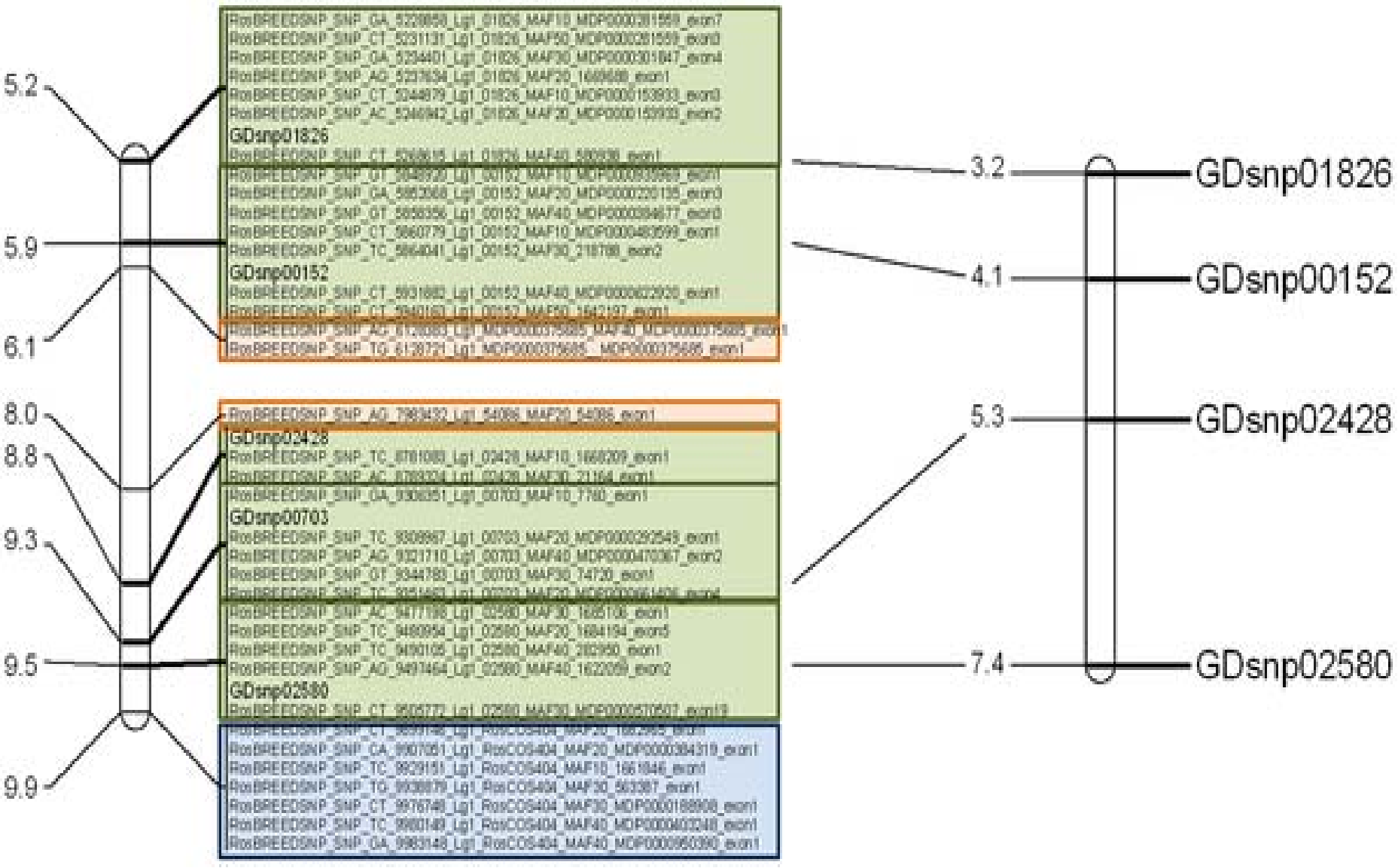


Figure 3. Detailed view of a genomic region at the top of Linkage Group 1 showing SNPs chosen for the International RosBREED SNP Consortium (IRSC) apple 8K SNP array v1. Physical map locations of SNPs (left; in megabases) are compared with known genetic map locations of SNP markers developed from 'Golden Delicious' (GDsnp in centiMorgans). SNP clusters around focal points are boxed, with green boxes denoting GDsnps, blue boxes denoting Rosaceae Conserved Orthologous Set markers (RosCOS), and orange boxes denoting candidate genes.

The IRSC apple 8K SNP array v1

(1). The 7867 SNPs were uniformly represented on the 17 apple LGs (Table 4). The number of SNP clusters ranged from 64 to 113.

(2). The average of 5.8 SNPs per SNP cluster.

(3). The average of physical distance between SNP clusters ranged from one cluster every 316.2 kb to 538.3 kb in LG 16 and 3, respectively.

(4). Overall cluster density on the 'Golden Delicious' x 'Scarlet' reference genetic map was one cluster every centiMorgan, with only five gaps between clusters larger than10 cM.

Table 4. Apple Infinium® II array v1 content and evaluation in the Plant & Food Research dataset. The number of attempted SNPs and clusters is indicated per linkage group (LG), as well as their density based on the apple genome assembly in kilobases (Kb) and the 'Golden Delicious' genetic map (cM).

	Attempted SNPs on IRSC Apple Infinium II array v1				Array evaluation in the Plant & Food Research dataset						
LG	No. of SNPs	No. of clusters	Average physical distance between focal points (kb)	Average genetic distance between focal points (cM)	No. of successful beadtypes	No. of poly-morphic SNPs	% poly-morphic SNPs	No. of poly-morphic SNPs with MAF>0.05	MAF>0.05, 50 GC>0.4, call rate >0.95	No. of poly-morphic clusters	% poly-morphic clusters
1	443	81	443.7	1.10	434	299	68.9	250	180	75	92.6
2	695	113	355.1	0.75	684	519	75.9	436	313	108	95.6
3	499	74	538.3	1.19	487	350	71.9	288	212	72	97.3
4	392	65	388.7	1.00	386	292	75.6	235	201	63	96.9
5	492	82	456.9	1.10	486	360	74.1	308	239	76	92.7
6	351	64	473.2	1.17	340	233	68.5	194	117	63	98.4
7	345	71	438.5	0.83	340	215	63.2	181	129	62	87.3
8	405	68	519.9	1.07	399	276	69.2	229	182	62	91.2
9	490	80	449.2	0.91	477	364	76.3	311	232	76	95.0
10	554	83	446.6	1.12	531	377	71.0	271	224	79	95.2
11	466	80	500.7	0.97	456	322	70.6	277	177	73	91.3
12	469	85	426.5	0.87	459	338	73.6	258	210	81	95.3
13	433	81	489.6	0.94	423	299	70.7	229	178	76	93.8
14	388	72	472.2	1.02	374	266	71.1	215	151	64	88.9
15	629	106	525.6	1.09	621	440	70.9	365	253	101	95.3
16	355	74	316.2	0.76	347	271	78.1	215	158	69	93.2
17	461	76	356.7	0.99	448	333	74.3	231	215	70	92.1
Total	7867	1355	446.2	1.00	7692	5554	72.2	4493	3371	1270	93.7

The Apple Infinium® array was evaluated by screening 1398 progeny in 8 segregating populations and 117 individuals from the Plant & Food Research apple germplasm collection.

doi:10.1371/journal.pone.0031745.t004

IRSC apple 8K SNP array v1 evaluation

(1). Evaluation of the IRSC apple Infinium® II 8K SNP array using 1619 individuals (accessions and segregating populations) yielded 7692 successful beadtypes (97.7%) of which 5554 (72.2%) were polymorphic (Table 4).

(2). The remaining 2138 SNPs (27.8%) exhibited poor quality genotype clustering or monomorphic. Number of polymorphic SNPs/LG ranged btw 519 and 215 on LG 2 and 7, respectively.

(3). Polymorphic SNPs were observed for 1190 (93.7%) of the attempted clusters.

(4). The Genomic Studio® *GeneTrain* scores for polymorphic SNPs ranged from 0.35 to 0.92 and the call frequency ranged from 0.73 to 1.

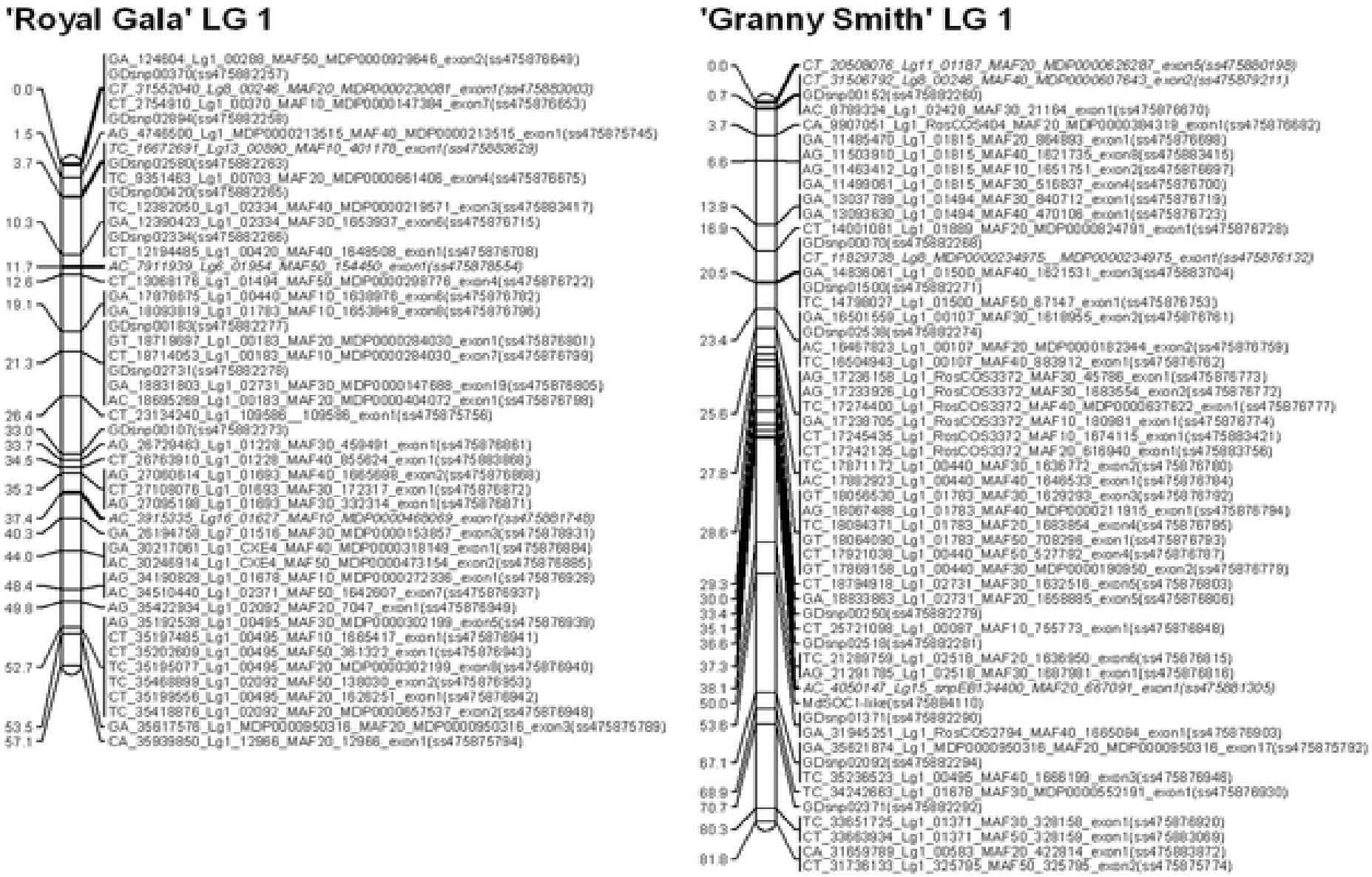


Figure 4. Genetic maps of linkage group (LG) 1 constructed using the 'Royal Gala' × 'Granny Smith' segregating population. Marker names indicate the type of SNP, physical location on the 'Golden Delicious' genome assembly (in base pair), cluster marker, minor allelic frequency (MAF) in the 27 re-sequenced accessions, gene model, position in the predicted gene model, and NCBI dbSNP accession number (in parentheses).

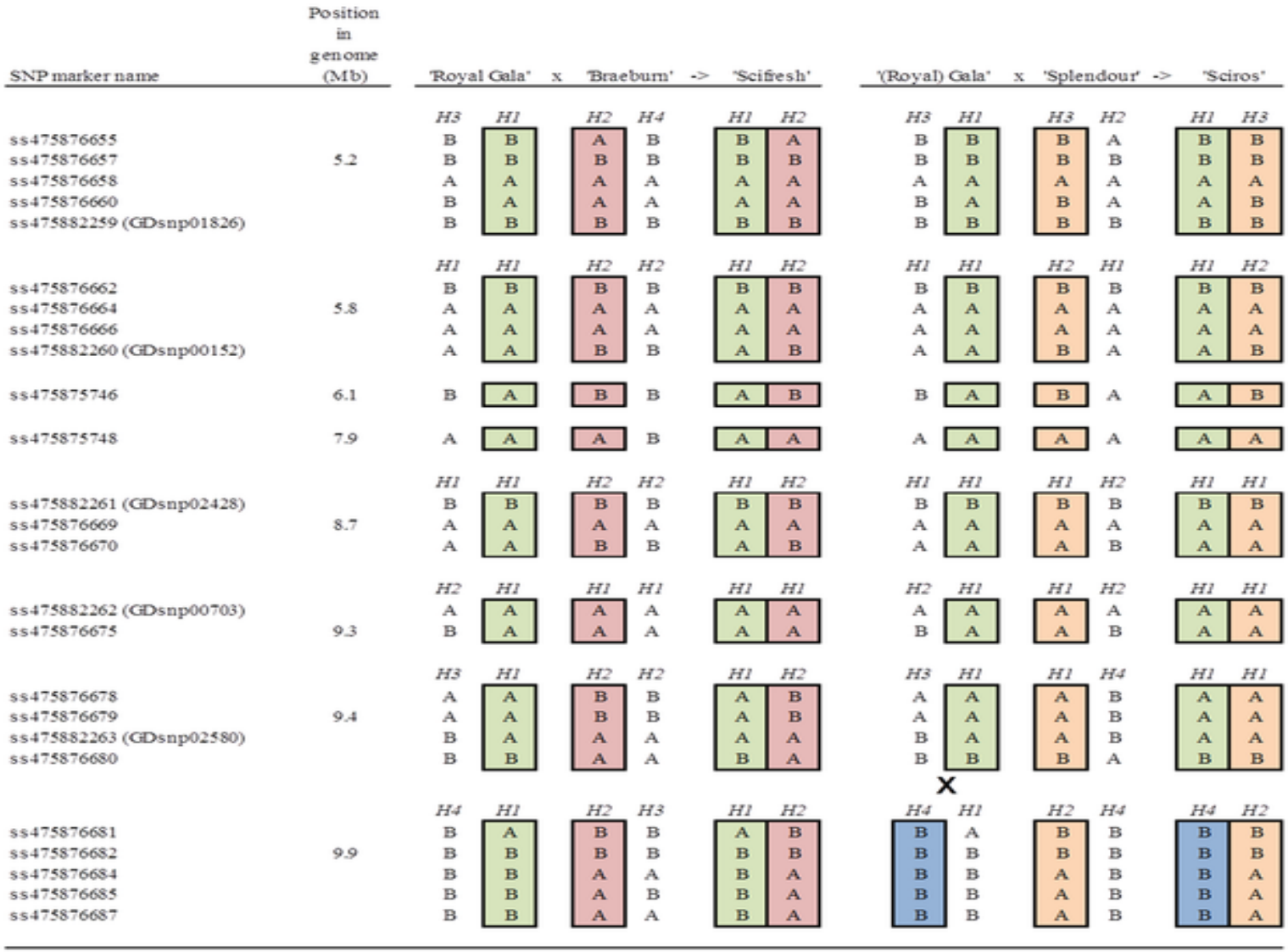


Figure 5. Example of the usefulness of the International RosBREED SNP Consortium (IRSC) apple 8K SNP array v1 for developing haplotypes. SNP markers were the same as represented in the example of Figure 3. Haplotypes inferred using FlexQTLTM for each cluster of SNPs are numbered from H1 to H4). Haplotypes inherited by the progeny from the parents are boxed and colored coded for each parent. A putative recombination is indicated by a cross for the '(Royal) Gala'x'Splendour' → 'Sciros' trio.

Conclusion

The International RosBREED SNP Consortium apple 8K SNP array v1 has been developed for public use by apple geneticists worldwide.

The design and evaluation of the array has indicated that it will be effective for a wide range of germplasm and applications such as high-resolution genetic mapping, QTL detection and characterization, marker-assisted introgression, and genomic selection.

Thank you..