Improving the association mapping pipeline in a loblolly pine population with a complex pedigree through increased marker coverage validation using resistance data

Tania Quesada, Christopher Dervinis, and Gary Peter

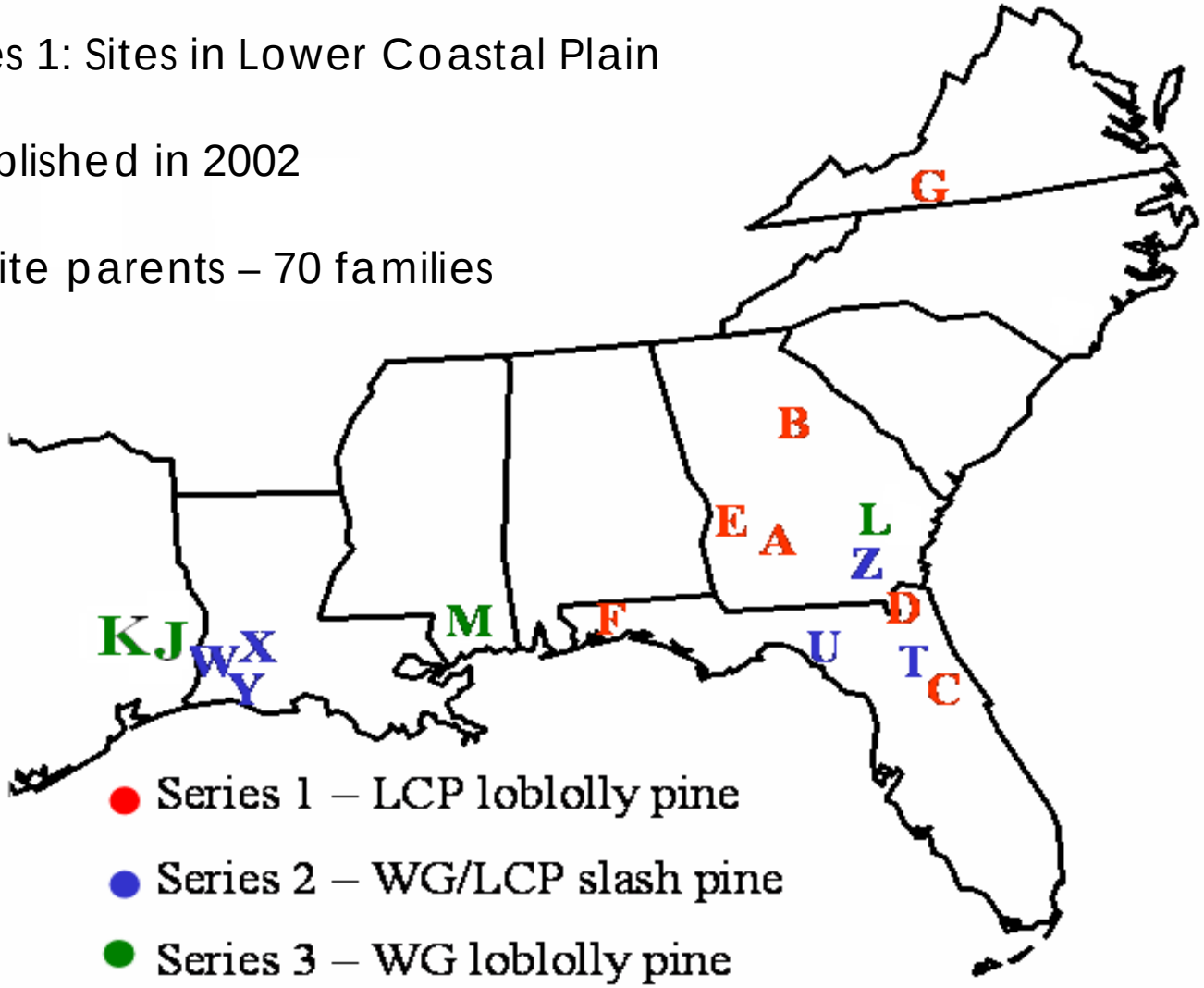
34th Southern Forest Tree Improvement Conference
Melbourne, FL June 19-22, 2017

CCLONES: Comparing Clonal Lines ON Experimental Sites

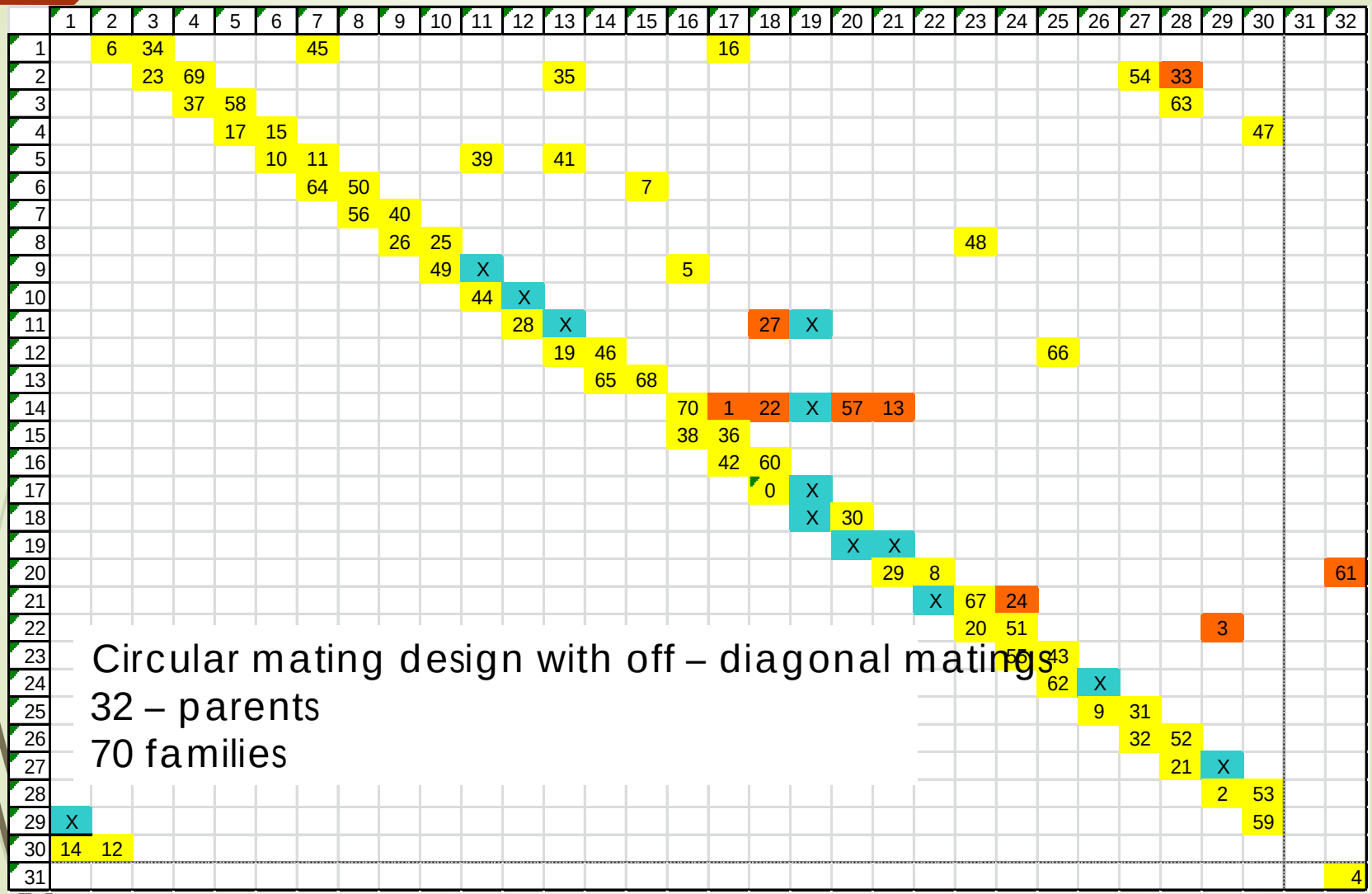
Series 1: Sites in Lower Coastal Plain

Established in 2002

32 Elite parents – 70 families



CCLONES mating design



CCLONES phenotypic data

Category	Trait	Ages measured
Growth	Height, DBH, survival	1-6, 8, 10, 12
Crown	Crown width, height to live crown	2,6
Disease - greenhouse	Fusiform rust & pitch canker	1
Disease- field	Pitch canker lesion length, fusiform rust score	2-6
Branch	Diameter, angle	6
Shoot	# flushes, flush length, # parastichies	2
Phenology	Shoot initiation, cessation, duration	2
Root	Root number, dry weight	1
Water use	Carbon isotope	3
Wood properties	Wood density, stiffness, lignin, diterpenoid content	3-8
Resin production	Oleoresin flow, resin canal number	6,7

CCLONES genotypic data

Project	Funding	Genotyping Platform	N SNPs	SNP location
ADEPT 1	NSF	Illumina Golden Gate	43	Known genes
ADEPT 2	NSF	Illumina Infinium	7,216	ESTs
PINEMAP	USDA-NIFA	Capture-Seq	67,387	Genome-wide

From probe design to SNP genotyping

Probe design	Probe selection	Probe testing / SNP genotyping
Unique for loblolly pine Genomic, EST unigenes and transcriptome 120 base pairs Exome sequences (target region of ~ 9.6 Mbp) High repeatability	Subset of population High-quality reads Good depth Adequate # SNPs Genic and intergenic	Total population SNP calling: • Polymorphic • Biallelic • Min. quality =10 • Minimum depth = 3 • Max. missing data = 0.4 • Minor allele freq. = 0.01
40K-80K probes	**20,000 probes**	**18,241 probes** **67,637 SNPs**

Probe selection : Selecting pilot population

Selection Criteria:

- **24 clones**
- **All provenances represented**
- **No duplicated families**
- **All 32 parents represented**
- **Sufficient DNA available**

Provenance	Clone	Family	Female	Male
ACCxFL	40042	400	17	18
ACCxFL	40204	402	29	28
ACCxFL	40321	403	29	22
ACCxLG	40414	404	31	32
ACCxLG	41124	411	7	5
FLxFL	41236	412	30	2
ACCxACC	41321	413	19	21
FLxFL	41505	415	6	4
ACCxACC	41648	416	17	1
ACCxACC	41928	419	12	13
FLxFL	42559	425	8	10
ACCxFL	42933	429	20	21
ACCxACC	43156	431	25	27
ACCxFL	43204	432	32	82
ACCxFL	43503	435	13	2
ACCXFL	43839	438	15	16
LGxLG	44052	440	7	81
ACCxFL	44423	444	11	10
ACCxACC	44609	446	12	14
FLxLG	44916	449	81	10
FLxFL	45378	453	28	30
ACCXFL	45531	455	24	23
ACCxACC	45845	458	3	5
FLxLG	46446	464	7	6

Probe selection : Selecting 20,000 probes

80,000 → **39,978** → **15,555** Genic + **24,423** Intergenic

Initial probes

Selected probes

Excluded:

- Probes with low or high depth
- Probes with no SNPs or too many SNPs

Do these probes align with ADEPT2 EST contigs?

15,555 Genic → **Yes:** 10,375 **No:** 5,180

24,423 Intergenic → **Yes:** 8 **No:** 24,415

Probe selection : Selecting 20,000 probes

Align with ADEPT2 EST contigs

15,555 Genic

Yes: 10,375
No: 5,180

Probes that aligned with one or more contigs. Only includes dataset with selected probes within genes.

Contigs matched to probe	Probes
1	10,341
2	28
3	5
4	1

Probes that aligned with one of more contigs. Only selected probes within genes were considered. There were 4,518 contigs with only one matching probe, and 2,509 contigs with two or more matching probes.

Matching Probes	Contigs
1	4518
2	1837
3	508
4	126
5	32
6	6

24,423 Intergenic

Yes: 8
No: 24,415

Number of intergenic selected probes with a given number of SNPs

SNP Number	N Probes
1	3569
2	3676
3	3459
4	3063
5	2647
6	2324
7	1825
8	1497
9	1317
10	1046

Probe selection : Summary

Selection criteria	Number Probes	Sum
Genic probes that don't align with ADEPT2 ESTs	5180	
Probes aligning with one contig	4518	9698
Contigs with 2 matching probes <4 SNPs*	1581	11279
Contigs with 3 matching probes <4 SNPs*	470	11749
Contigs with 4 matching probes <4 SNPs*	121	11870
Contigs with 5 matching probes <4 SNPs*	32	11902
Contigs with 6 matching probes <4 SNPs*	6	11908
Excluded genic monomorphic probes (no SNPs)	11217	23125
Excluded genic probes (alignment cutoff) - low read number	705	23830
Intergenic selected probes with 1-4 SNPs	13767	37597
Intergenic selected probes in large scaffolds or LG	TBD	37597

*Selected only one probe of the multiple matching a same contig

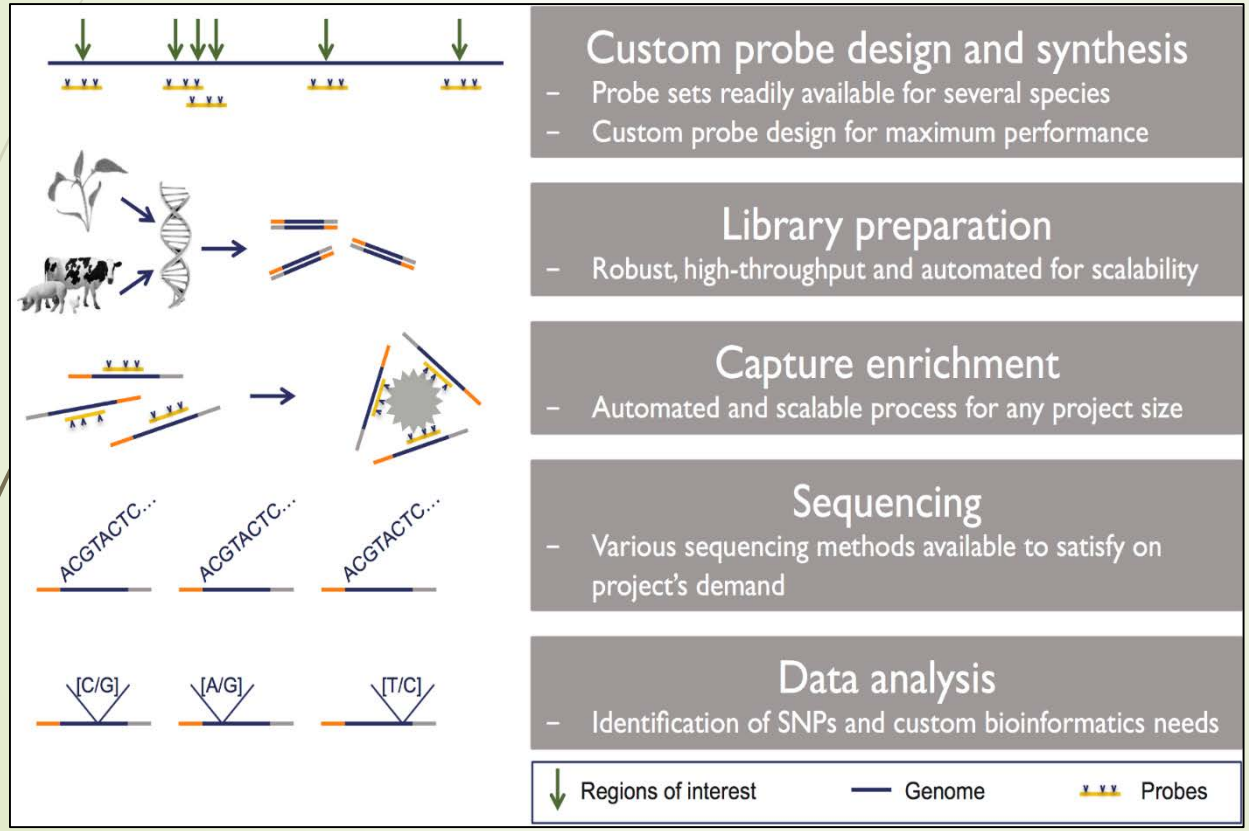
Matching Probes	Contigs
1	4518
2	1837
3	508
4	126
5	32
6	6

Probe selection : Final Selection

Final Selection	N Probes
Selected: Genic polymorphic probes with high quality	15455
Selected: High-quality genic monomorphic probes with medium read number that align to genes. Excludes ADEPT2 contigs. Selected randomly from 5,663 probes.	1697
Selected: monomorphic genic probes with high quality and low read number	140
Selected: Randomly-selected intergenic high-quality probes with less than 4 SNPs.	2708
Total	20000

Probe testing / SNP genotyping

Sequence capture method



Source: RAPiD Genomics
(http://www.rapid-genomics.com/technology/)

Probe testing / SNP genotyping

Total population: **920 clones**

SNP calling:

- Polymorphic
- Biallelic
- Min. quality =10
- Minimum depth = 3
- Max. missing data = 0.4
- Minor allele freq. = 0.01

18,241 probes

67,637 SNPs - PINEMAP

Association mapping for disease resistance

Pitch canker

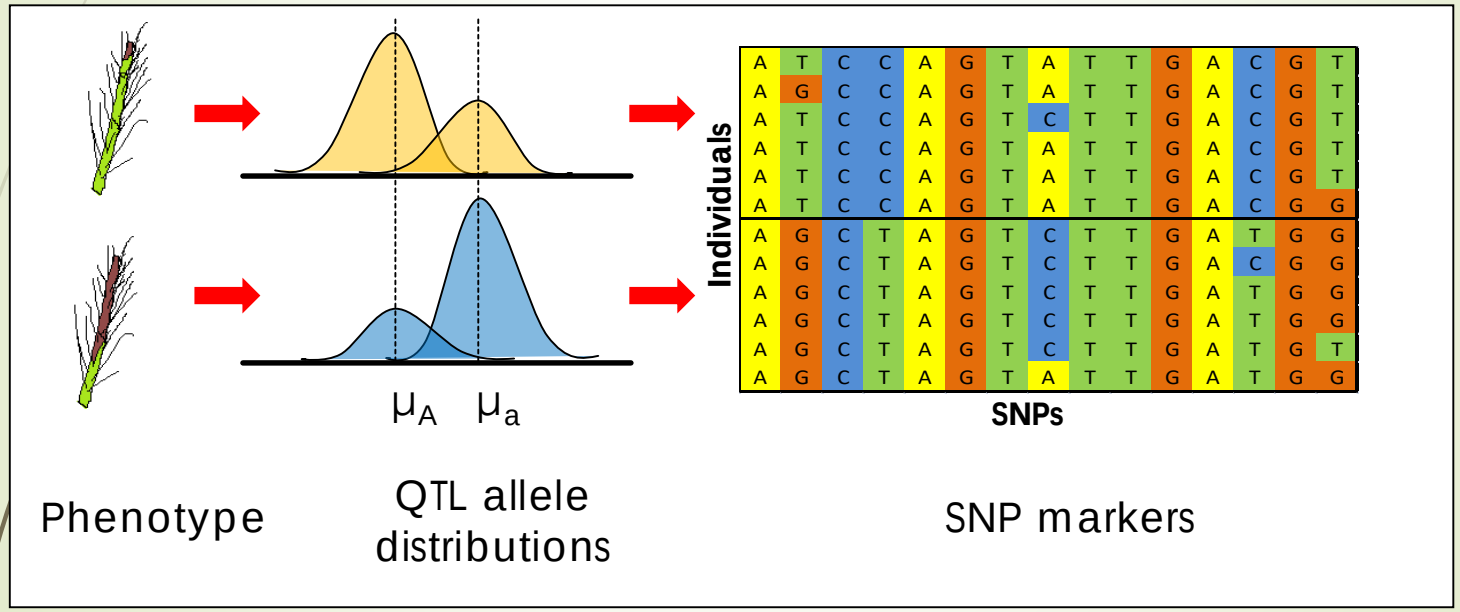
- Caused by *Fusarium circinatum*
- Episodic,Broad host range
- Causes resinous lesions
- Resistance quantitative and heritable

Fusiform rust

- Caused by *Cronartium quercuum* f.sp. *fusiforme*
- Causes stem galls, high seedling mortality
- Resistance due to R genes

Using SNP genotypes for association mapping

What is association mapping?



Adapted from Rafalski, 2010 Curr Opin Plant Biol 13: 174-180 and Jannink et al., 2001Trends Plant Sci 6: 337-342

Molecular markers associated to a trait will show a bias in their allelic frequency with respect to the phenotype

Association Model

- BAMD (Bayesian Association with Missing Data):
  - solves for all SNP effects simultaneously,
  - performs multiple imputations for missing data

(Li et al., 2012)

Trait Genotypic effects Error

$$y = X\beta + Z\gamma + \varepsilon$$

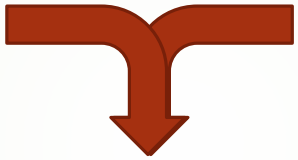
Phenotype **Structure matrix** **SNP matrix**

	clone	Predicted _Value	cluster _1	cluster _2	cluster _3	cluster _4	cluster _5	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	108A	1.9096	0.0118	0.043	0.6356	0.1086	0.201	AG	AC	AG	AT	TT	AG
2	121C	1.0752	0.02	0.039	0.7617	0.0423	0.137	AA	AA	GG	TT	AT	AG
3	125B	1.2011	0.0354	0.7823	0.0298	0.0275	0.125	AA	AC	AG	TT	TT	AG
4	128A	0.966	0.6652	0.0423	0.0618	0.1588	0.072	AG	AC	AG	TT	AA	AG
5	131B	0.9658	0.2197	0.7248	0.0189	0.0182	0.0184	AG	AA	AG	AT	AA	AA
6	151C	1.1784	0.0127	0.0318	0.2616	0.3562	0.3377	AG	CC	GG	AT	AT	GG

Integrating ADEPT 2 SNPs with PINEMAP SNPs

ADEPT2 EST Sequences

BLASTn



P. Taeda genome sequence v1.01

- Best Hits
- E-value < 0.001
- > 90 % identity

Genome scaffold
Matching interval

PINEMAP SNPs
- Genome scaffold
- SNP position

PINEMAP SNPs

No Match		Hits		Total
24,419	+	**42,968**	→	**67,387**

Results ADEPT 2 – Pitch canker

Association analyses revealed 10 significant SNPs

SNP_ID	% Diff in clonal variance	Best Hit (Expect < 1e-10)	Predicted SNP location	Effect on a.a. sequence
0_15227_01_160	0.924	ATP binding protein, lectin-like protein kinase	Coding region	Synonymous substitution
0_15382_01_104	3.832	geranylgeranyl transferase type I beta subunit	Coding region	V to A change
0_2234_01_128	1.745	putative long-chain acyl-CoA synthetase	Coding region	D to Y change
0_6323_01_248	0.889	DELLA protein	Coding region	Synonymous substitution
0_9288_01_372	0.288	No hits found	Coding region	Synonymous substitution
1_3327_01_116	2.187	No hits found	Coding region	C to Y change
2_4484_02_622	1.101	hexose transporter	Coding region	C to Y change
2_6181_02_400	1.129	hexokinase	Non-coding, Putative 3'UTR	N/A
2_8946_02_437	1.031	Cucumber peeling cupredoxin	Non-coding, Putative 3'UTR	N/A
CL4336Contig1_01_180	1.519	Unknown [Picea sitchensis]	Coding region	Synonymous substitution

(Quesada et al., 2010)

Significant SNPs altogether account for about 13% of the clonal variance

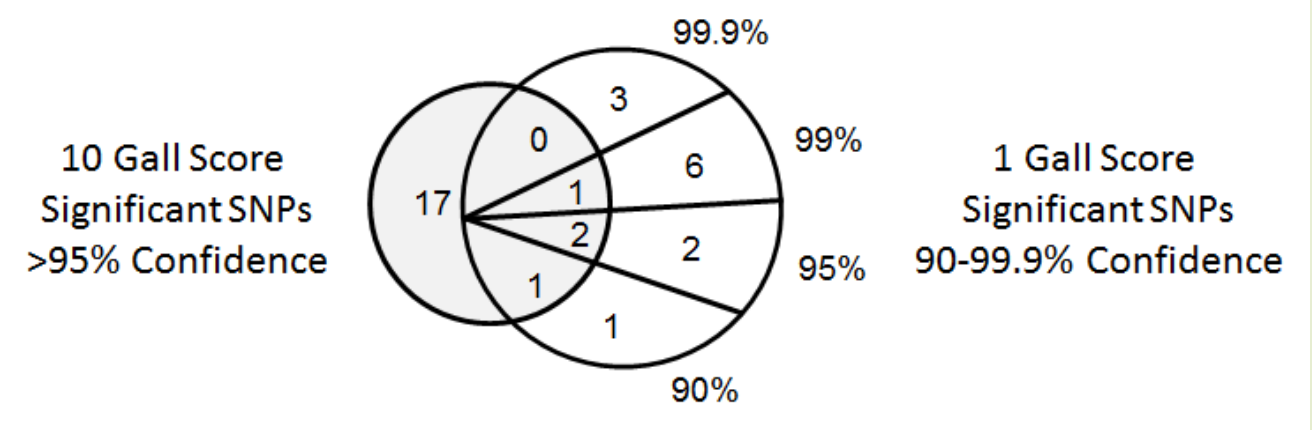
Ignoring annotation for SNP selection allowed detection of associations that would have been unnoticed

Results PINEMAP – Pitch canker

SNP	Conf. Int.	Scaffold	Position	Ref	Alt	Query acc.ver	Merge Marker ID	LG	SNP in Contig	Prot_desc
V18500	90	scaffold433234.2	19299	A	T	no match	N/A	N/A		
V22500	90	scaffold562427	135000	A	C	2_5516	N/A	N/A	1	unknown
V27650	90	scaffold739754	2698	T	C	no match	N/A	N/A		
V31916	90	scaffold857972	22985	G	C	no match	N/A	N/A		
V39665	90	tscaffold1240	157547	T	C	0_668	2_4205	4	0	unknown
						2_4205	2_4205		1	Ribosomal protein S27
V57839	90	tscaffold614	364426	C	T	0_10930	0_13765	5	0	ADR1-like 1
						0_13765	0_13765		1	
						CL1Contig163	0_13765		0	
						CL1Contig247	0_13765		0	
V19320	95	scaffold460829	31462	C	G	CL642Contig1	CL642Contig1	10	1	NA
						UMN_CL135Contig1	CL642Contig1		0	
V21018	95	scaffold513999	210140	A	G	no match	N/A			
V63202	95	tscaffold781	451675	A	G	0_11052	0_12447	8	0	ARM repeat superfamily protein
						0_12447	0_12447		0	
						0_16663	0_12447		0	
						0_1957	0_12447		0	
						2_4321	0_12447		0	
						CL2506Contig1	0_12447		0	
						CL465Contig1	0_12447		1	unknown
V57445	99	tscaffold602	32618	T	C	UMN_3547	N/A	N/A	0	

Results ADEPT 2 – Fusiform rust

Significant SNPs for rust resistance between 10 gall score data and 1 gall score.



Of the total of 21 significant SNPs that were detected in the 10 gall test, 4 were also significant for the 1 gall inoculum at varying levels of significance, while 12 were only significant for the 1 gall inoculum.

Results PINEMAP – Fusiform rust

1 Gall Score

SNP	MeanEffect	Confidence Interval	LG	ScaffoldID	Aligned Genome Coord./ Position	Matching ADEPT2 Contig	Prot. Desc.
0-12681-01-532	0.080	99	2	scaffold849546	7365--8810	0_12681	IQ-domain 19
0-9379-01-192	-0.075	95	-	-	-	0_9379	Metal ion binding
V7580	0.065	95	-	scaffold131375.2	85657	0_1133	unknown
0-881-01-114	-0.062	95	10	scaffold188103.2	217708--216609	0_881	unknown
V37258	-0.060	95	-	scaffold901548	93494	CL3858Contig1	unknown
V1911	0.053	95	-	C32066054	15870	no match	-
V58021	0.063	90	-	tscaffold6184	1447	CL2762Contig1	unknown
V62668	0.059	90	-	tscaffold765	521569	CL1888Contig1	PRED: ATP-dependent Clp protease proteolytic subunit 4
V11359	0.055	90	-	scaffold227162	25710	no match	-
CL258Contig1-01-98	-0.057	90	-	-	-	CL258Contig1	60S ribosomal protein L21
CL133Contig2-06-75	-0.056	90	2	tscaffold2704	130877--130073	CL133Contig2	Serine-rich protein-related protein
V7563	0.049	90	-	scaffold131057.3	112755	no match	-
V42882	-0.045	90	-	tscaffold2079	109999	CL429Contig1	Glucose-6-phosphate/phosphate translocator 1
0-4172-02-205	-0.042	90	9	tscaffold2607	57489--58078	0_4172	EPS15 domain 1
V28046	-0.042	90	-	scaffold75403	114031	CL351Contig1	unknown

Merging ADEPT 2 SNPs with PINEMAP SNPs

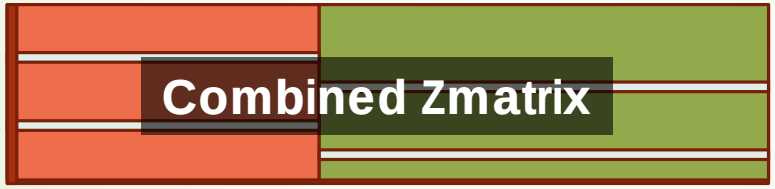
Zmatrix ADEPT2

7,216 SNPs

Zmatrix PINEMAP

67,387 SNPs

Merge by CloneID – Keep all



74,604 SNPs

Remove monomorphic SNPs

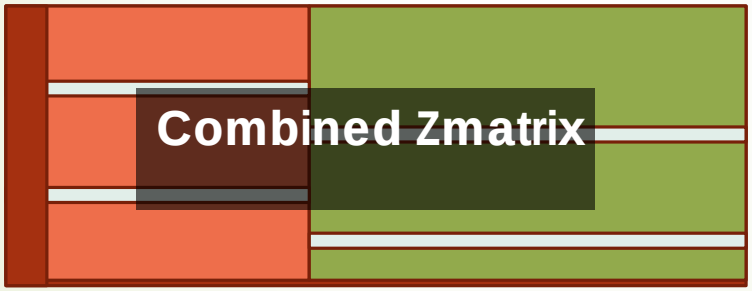
Merge with list of phenotyped clones (subset of Combined Zmatrix)

74,505 SNPs

SNP Preprocessing prior to Association Analysis

Remove monomorphic SNPs

74,412 SNP



Merge with phenotypic dataset (Clone, rep, phenotype)

Pre-processing ANOVA

Full Model: $Y_{ij} = \mu + SNP_k + Rep_j + SNP*rep_{jk} + e_{ij}$

Reduced Model: $Y_{ij} = \mu + Rep_j + e_{ij}$

Rank P-values; select top 400 SNPs → BAMD software

Results Combined – Fusiform rust

1 Gall Score

SNPID	Dataset	Mean Effect	CI	Scaffold	Position	Ref	Alt	Merge Marker ID	LG	Seg 10-5	Putative Function
V26963	PINEMAP	0.098	99	scaffold716874	10733	G	T	No match	--	--	--
V10922	PINEMAP	0.087	95	scaffold217054.2	13242	G	A	No match	--	--	--
V26435	PINEMAP	-0.084	95	scaffold698284.2	8460	G	A	No match	--	--	--
V56765	PINEMAP	-0.087	95	tscaffold5830	60979	T	C	No match	--	--	--
V40021	PINEMAP	-0.112	95	tscaffold1321	480399	C	A	0_5350	12	Yes	Transcription factor IIA/ alpha/beta subunit
V49998	PINEMAP	0.096	90	tscaffold3868	129244	C	T	No match	--	--	Glycine hydroxymethyltransferase
V51272	PINEMAP	0.087	90	tscaffold4233	3079	A	G	0_14705	5	Yes	Alcohol dehydrogenase transcription factor Myb/SANT-like family protein
V28961	PINEMAP	0.079	90	scaffold780931	122446	A	G	No match	--	--	--
V27641	PINEMAP	0.077	90	scaffold739441.1	94309	C	A	No match	--	--	--
V48793	PINEMAP	0.068	90	tscaffold3579	165544	A	G	0_5262	2	Yes	--
V53524	PINEMAP	0.068	90	tscaffold490	594068	C	T	0_15530	5	No	Tubulin alpha-2 chain
V8695	PINEMAP	-0.087	90	scaffold158435.2	1430	G	C	No match	--	--	--
V31656	PINEMAP	-0.109	90	scaffold854030.2	100001	A	G	No match	--	--	--
V19178	PINEMAP	-0.252	90	scaffold456449	78210	T	C	0_11406	7	Yes	--

Summary

- CCLONES → best characterized study in the FBRC
  - > 25 phenotypic traits
  - > 70,000 genotypic markers

- New genotyping process allows higher SNP density
- Association analyses with new SNP dataset produced equal or higher number of significant SNPs.
- When ADEPT2 SNPs and PINEMAP SNPs were analyzed together, some significant SNPs from ADEPT2 were also significant in the new dataset.
- Further characterization of significant SNPs is under way

Acknowledgments

FBRC