



Cortex_var as a novel tool for variant calling

Junyu Li & Matt Kendzior
Supervisor: Liudmila Mainzer

GitHub: <https://github.com/ncsa/SPIN-cortex-var>

Soybean Genome

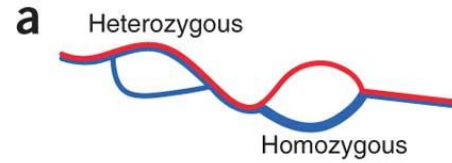
- Allotetraploid
 - 4 sets of chromosomes relative to haploid number, resulting from both sets a parental chromosomes being present in the gametes
- Highly Duplicated
 - Two major duplication events
 - 75% of genes present in multiple copies (Schmutz et al. 2010)
- Majority of genome composed of retrotransposons and DNA transposons
 - 42% long terminal repeat retrotransposons and 17% DNA transposons (Schmutz et al. 2010)
- Structure
 - Blocks of duplicated genes involve many chromosomes
 - Very high retention of homologues (Schmutz et al. 2010)
 -

Samples and Reference Used

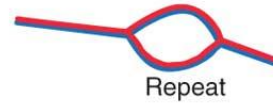
- 4 low coverage samples: for shorter computation time
 - Magellan: 3.49x
 - Maverick: 5.15x
 - PI398_881: 9.61x
 - PI574_486: 8.38x
- Latest Soybean Reference Genome: GCF_000004515.4_Glycine_max_v2.0

Cortex uses colored De Bruijn graph

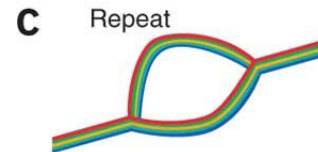
(a) **Bubble calling:** Identify variants by comparing **sample** vs. **reference**



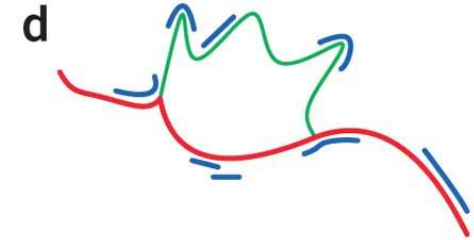
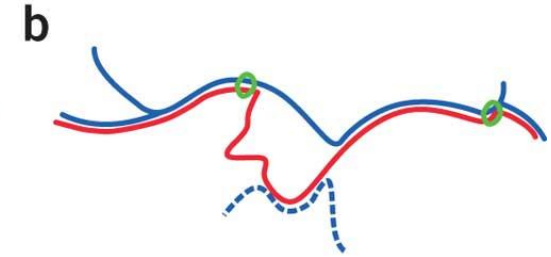
(b) **Path divergence:** Find homozygous variants by tracking **reference** vs. **sample**



(c) **Multiple-sample analysis:** Distinguish repeats from variants



(d) **Genotyping:** Calculate genotype likelihood from allele coverage



Step 1: Create binary graphs for each strain

Step 2: Pool and clean sequencing errors

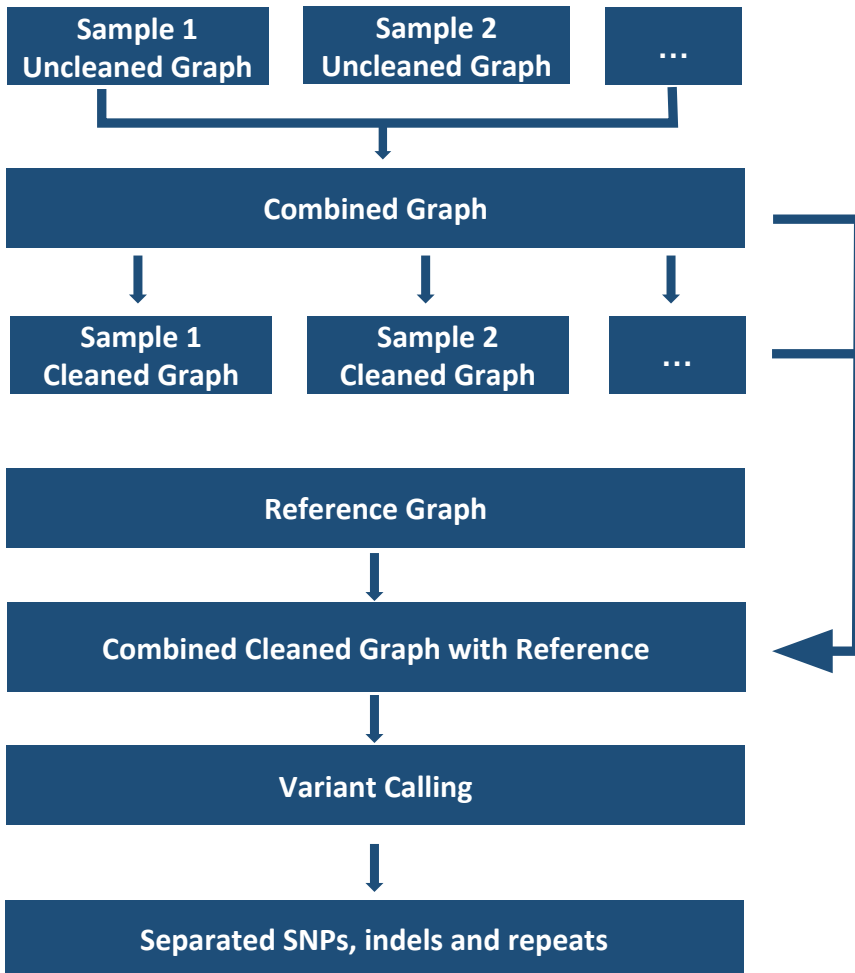
Step 3: Clean the errors in individual samples

Step 4: Create a graph for reference

Step 5: Combine reference graph with sample graphs and cleaned pool

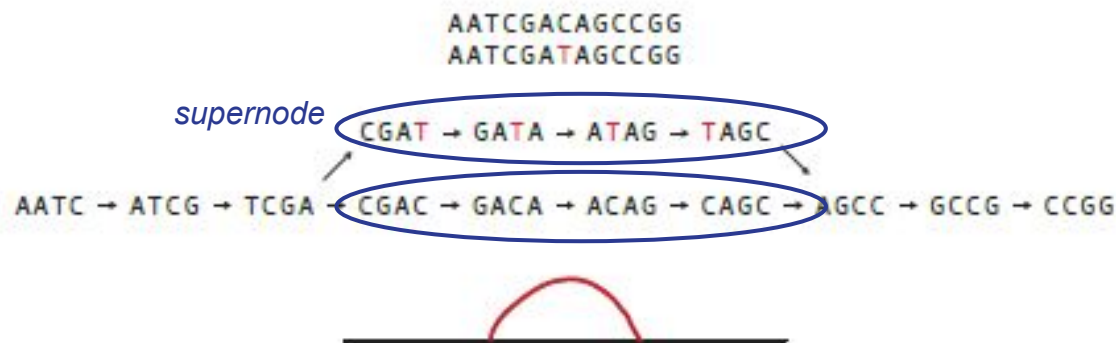
Step 6: Call variants by *Bubble Caller* or *Path Divergence*

Step 7 : Classify variants and generate VCF

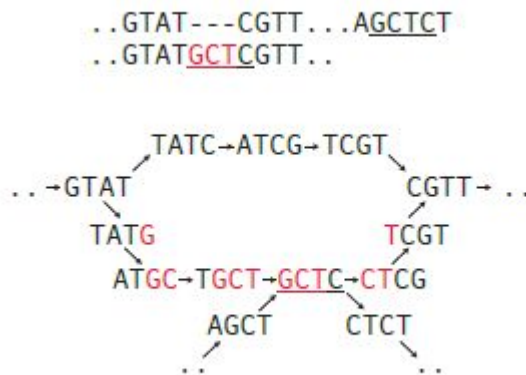


Bubble Caller

- Extends in 2 directions until a junction
- Only finds clean bubbles: alleles do NOT touch other parts of the graph
- Can be reference free



Clean bubble of a SNP



Confounded bubble

Bubble Caller results: Complex Variants

Needleman-Wunsch algorithm is used to differentiate SNPs from indels in variant clusters

```
1 4747331
Identity = 73.214 percent
Score = 670
TATTTTTTAATTCTCAATTAAATTTATATTTTTTATCTCTCAAATTCAAAAATATTTTTTAATCTCTCAAATTCAAAAATATTTTTTTAGTAAGATAGTATAACAATTCTA
T TT TTTAATTCTCAATT A A A A T TT ATATTTTTTATCTCTCAAATTCAAAAATATTTTTTTAGTAAGATAGTATAACAATTCT
TCTTCTTTAATTCTCAATT---A---A-A-----A---T---TT-----ATATTTTTTATCTCTCAAATTCAAAAATATTTTTTTAGTAAGATAGTATAACAATTCTT
```

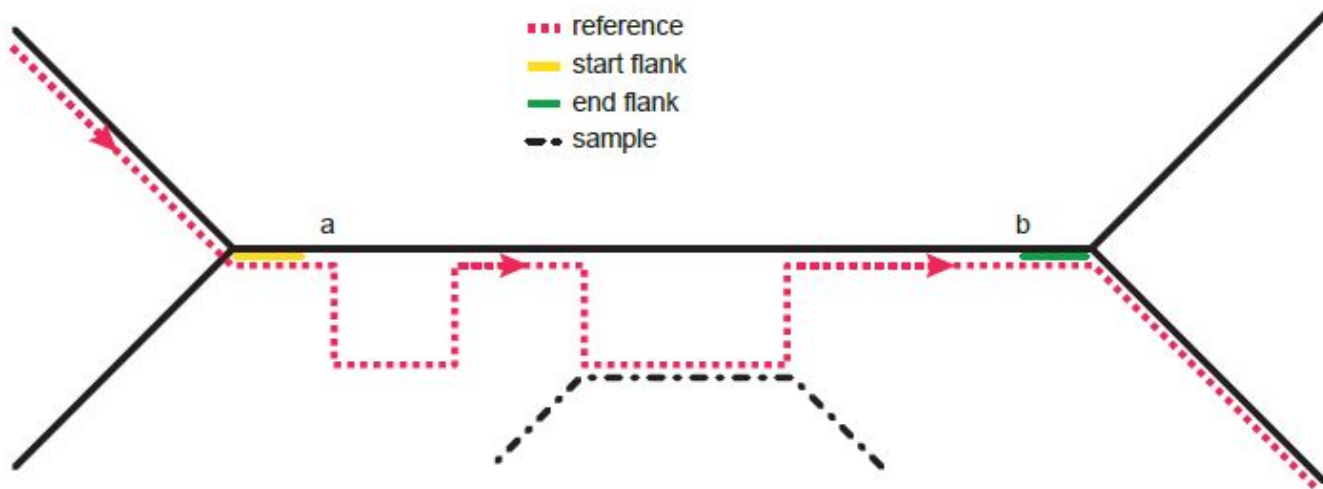
Decomposed VCF

```
1 Netw 4747332 var_66872_sub_snp_1 A C . PASS
SVTYPE=SNP_FROM_COMPLEX;SVLEN=0 COV 3,15 1,0 2,0 0,9
0,6
```

```
1 OneD 4747393 var_66872_sub_indel_1 ATCTCTCAAATTCAAAAATATTTTTTA A
. PASS SVTYPE=INDEL_FROM_COMPLEX;SVLEN=27 COV 3,15 1,0
2,0 This PC 0,9 0,6
```

Path Divergence Caller

- Follows reference path through the joint graph
- Both breakpoints on the same supernode for the sample
- Can generate haplotypes much longer than read-length or insert size



Classification

Variants are classified into one of the models:

Variation

Binomial(2,x) $x \sim$ population allele frequency

Repeat

Beta B(2,2)

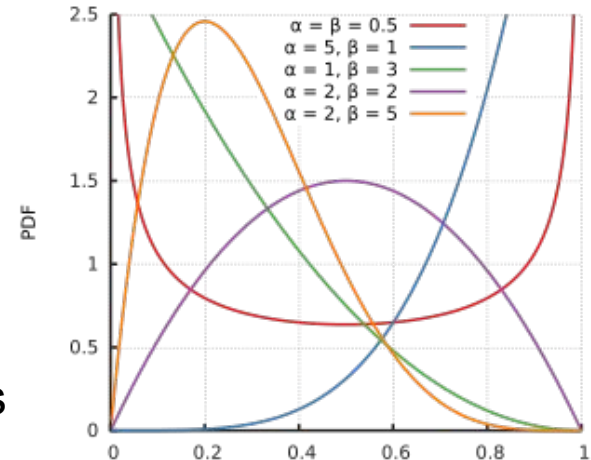
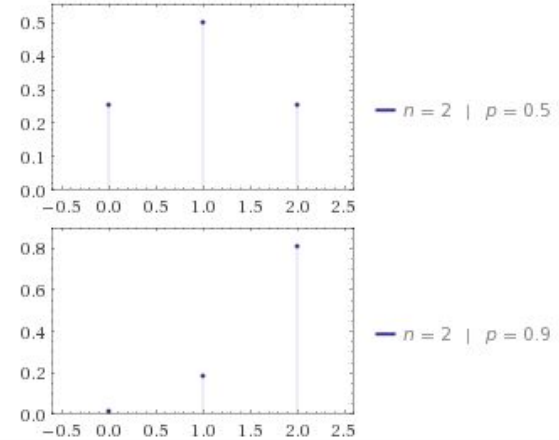
Error

Beta B(100,100 ϵ) $\epsilon \sim$ sequencing error rate

Models are compared by using Bayes Factors

Log-likelihood of the result model $\geq 10 \times$ the others

Plots for typical parameters:



What information does each file contain:

Bubble & Path Divergence Caller output:

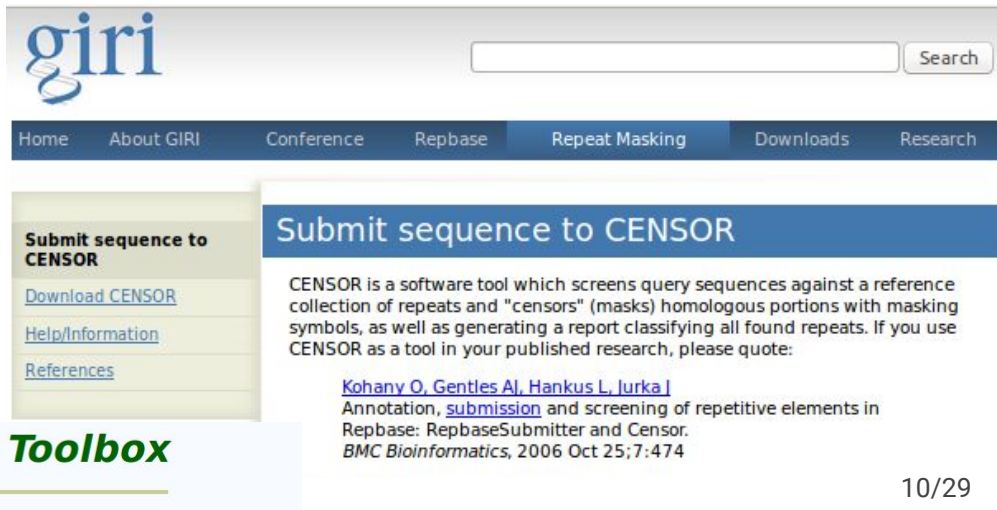
- ✓ Variant ID
- ✓ Flanking & branch sequences
- ✓ Coverage in each color

VCF:

- ✓ Variant ID
- ✓ Flanking & branch sequences
- ✓ Location

Classified repeats:

- ✓ Variant ID
- ✓ Confidence score



Submit sequence to CENSOR

CENSOR is a software tool which screens query sequences against a reference collection of repeats and "censors" (masks) homologous portions with masking symbols, as well as generating a report classifying all found repeats. If you use CENSOR as a tool in your published research, please quote:

[Kohany O, Gentles AJ, Hankus L, Jurka J](#)
Annotation, [submission](#) and screening of repetitive elements in
Repbase: RepbaseSubmitter and Censor.
BMC Bioinformatics, 2006 Oct 25;7:474



We read variant IDs from the classified variant file, and look up the corresponding sequences in the bubble caller output.

```
var_1    variant 0.8577763
var_2    variant 4.088314
var_3    variant 0.2867406
var_4    variant 0.4844286
var_5    variant 1.041722
var_6    variant 0.1751015
var_7    variant 1.902281
var_8    repeat 0.07640206
var_9    repeat 0.02359119
```

[illegible]

Using Soybase to find variant locations

- Copy & past the 5' flanking sequence from the top it in the variant classified file of PI398881 relative to reference into soybean BLAST

```
>Gm13
      Length = 45874162

      Score = 327 bits (165), Expect = 3e-88
      Identities = 165/165 (100%)
      Strand = Plus / Minus

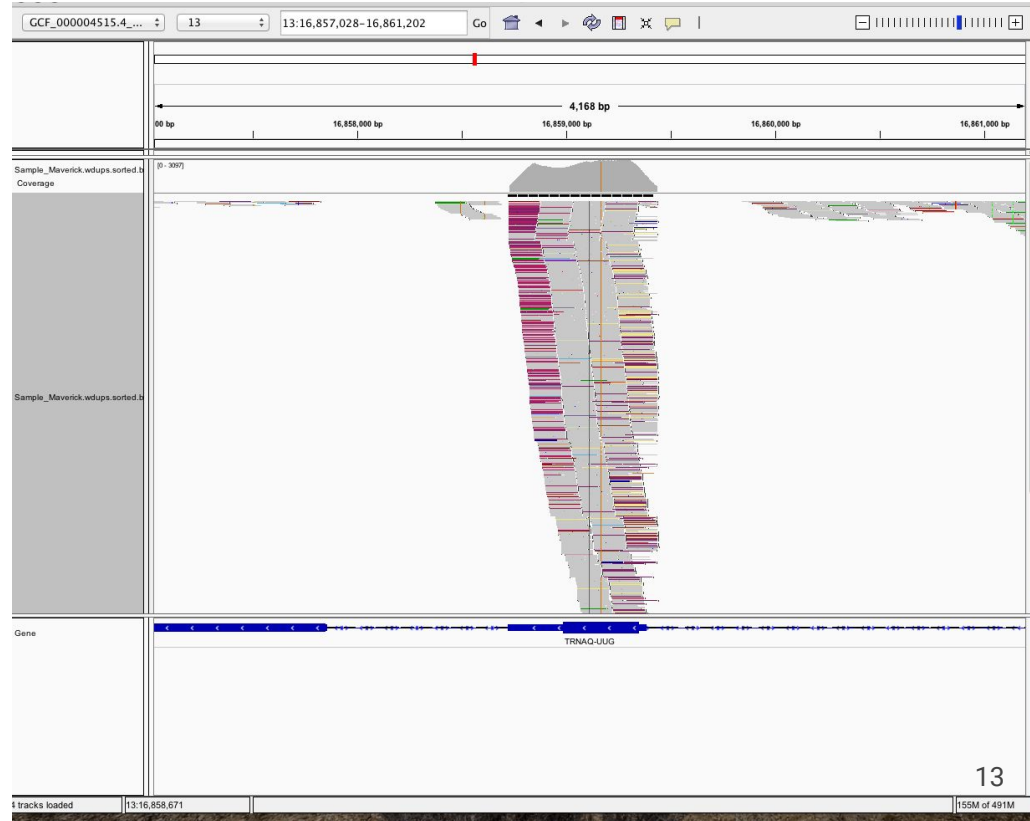
Query: 1          atgaaatgctccactagtagtactcttcttatctacttttcatcctacataatctgaatagga 60
                |||
Sbjct: 16859660 atgaaatgctccactagtagtactcttcttatctacttttcatcctacataatctgaatagga 16859601

Query: 61          atacc caactaagcttataggtgcacccttaggataccacaatcctacttttaatgtacc 120
                |||
Sbjct: 16859600 atacc caactaagcttataggtgcacccttaggataccacaatcctacttttaatgtacc 16859541

Query: 121         cttaatgtactttaaaaaattctcttaacaactatttaa atgagattt 165
                |||
Sbjct: 16859540 cttaatgtactttaaaaaattctcttaacaactatttaa atgagattt 16859496
```

Using IGV to view potential repetitive regions

- Very large read pile up
- Annotation track shows gene36014
- PI398881 showed a large read pile up also
- What is causing the read pile up?
 - LTR retrotransposon
 - A gene with many paralogues?



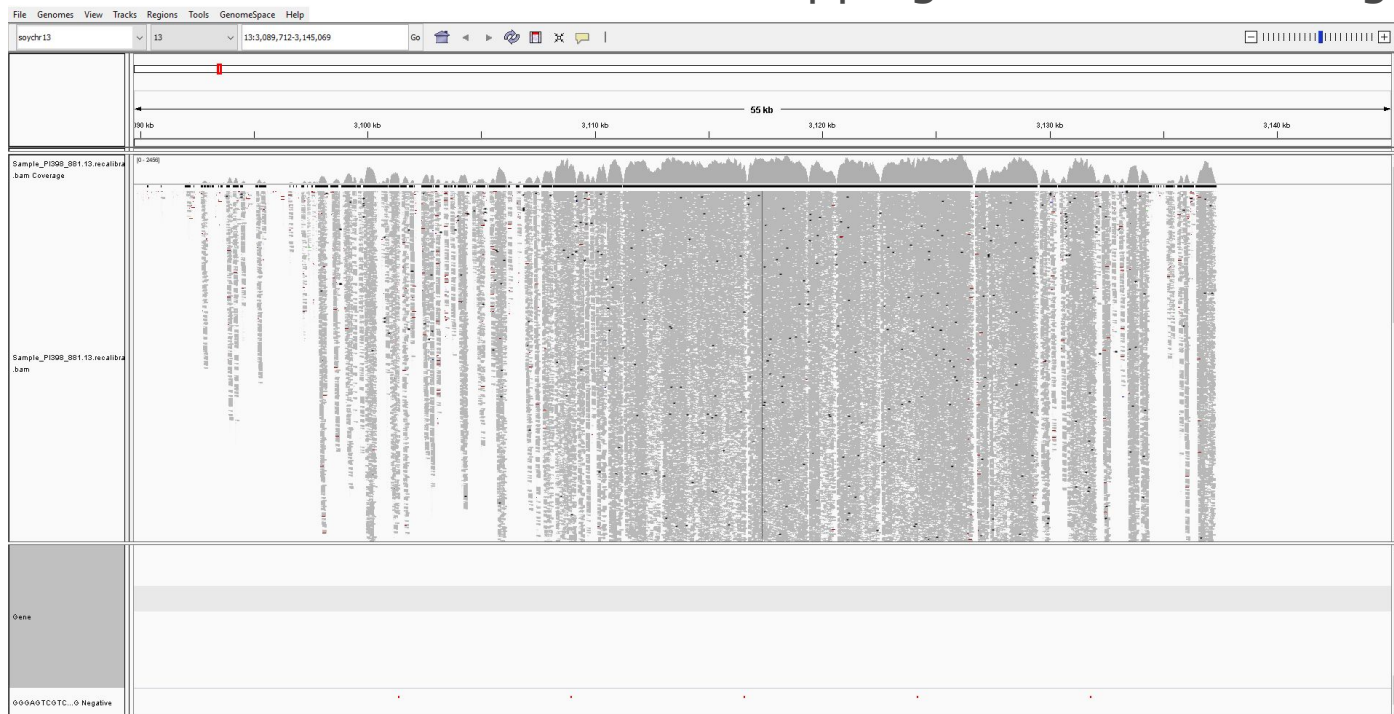
Next steps

- Extract the sequence of interest from IGV
- Repbase
 - LTR retrotransposon from the Chickpea genome
- We then had the idea to started calling samples relative to themselves in Cortex instead of relative to the reference and pool
 - To find much more instances paralogs and tandem repeats

Cortex run with PI398881 relative to itself

- Top hit in classified variant output file
 - Var_26780 repeat 171.7938
- Soybase BLAST showed strong coverage in chromosome 13
 - Maps to a 5-time tandem gene repeat
 - Sequence similarity to ubiquitin-protein transferase activatory activity
- 5' flanking region mapped 5 times in tandem on reference (shown in next slide)

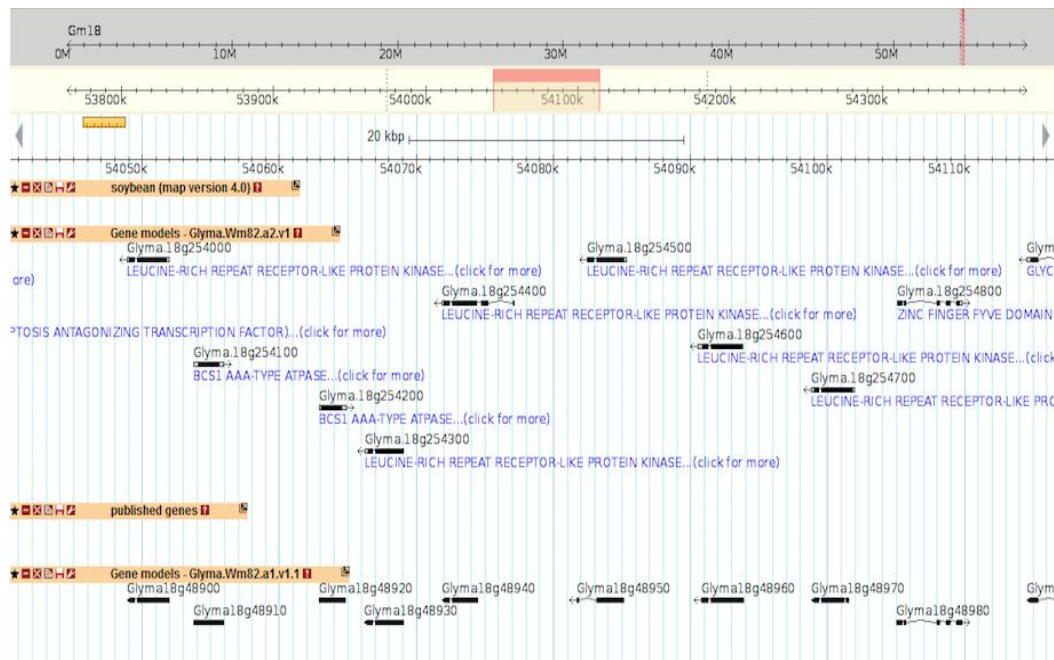
little red dots on the bottom are mappings of the 5' flanking region



The 5' flanking region maps to 135 loci across the entire genome, mostly with 100% identity. Given that the cultivar is sequence to $\sim 10\times$ depth, and during alignment we retain up to 25 secondary mappings, the huge pileup (2456) in this region (Gm13:3,092,592..3,147,949) is not surprising ($135 \times 10 = 1350$). It could be PI3 has 5-times tandem repeat, or more times in tandem.

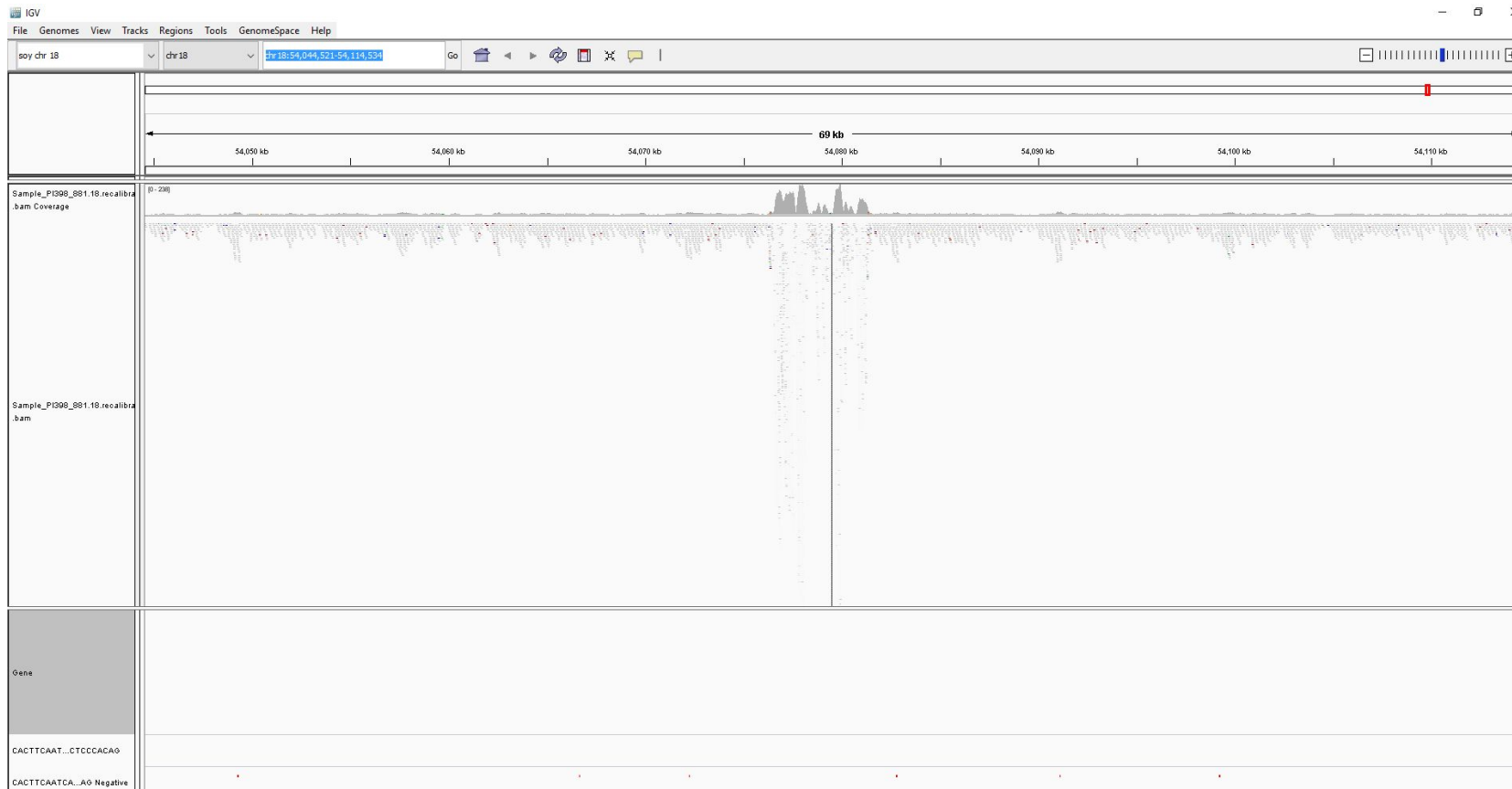
Bubble Caller on PI398881 relative to itself

- Second top hit
 - Var_2576 repeat 130.5193
- Mapped to a tandem repeat
- Leucine-rich receptor-like protein kinase



Soybase genome view

IGV view of the leucine-rich receptor-like kinase tandem repeat for PI398881 aligned to the newest reference

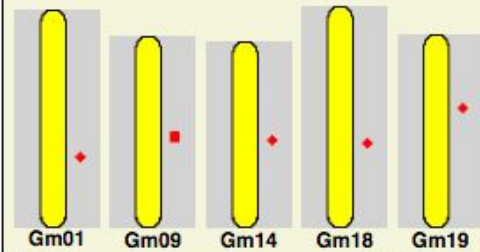


Path Divergence to call variants

- Top Hits from classified variant file for PI 398881
 - var_2557 length=17770 score = 37060.56
 - var_3467 length=14279 score=32133.22
 - var_2736 length=19967 score=24935.76
- The scores are significantly higher than the repeats found by Bubble Caller (with top hit score 171.79)

BLAT(similar to BLAST) result of the reference branch on Soybase:

The following 5 regions match your request.



Name	Type	Description	Position	Match Score
Alignment1	BLAT		Gm01:38302052..38316331	100.00
Alignment2	BLAT		Gm18:34940937..36739786	93.49
Alignment3	BLAT		Gm09:24784377..27295759	93.16
Alignment4	BLAT		Gm19:18067914..20202827	93.13
Alignment5	BLAT		Gm14:25041982..27222558	92.58

Path Divergence Repeats - PI398_881

From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
2	458	GmOgre_LTR	1214	1668	LTR/Gypsy	d	0.9628	1.4000	3971
459	10633	GmOgre_I	1	10194	LTR/Gypsy	d	0.9347	2.2613	83040
10637	11931	Gypsy-53_GM-I	9798	11075	LTR/Gypsy	d	0.7590	1.5376	5020
12277	12315	RETROFIT6_I	2869	2905	LTR/Copia	d	0.8158	1.5000	205
12648	14253	Gypsy-118_GM-I	11837	13440	LTR/Gypsy	d	0.8724	1.3261	10441
14455	15606	Gypsy-53_GM-I	13653	14793	LTR/Gypsy	d	0.7063	1.5353	2311
15640	15709	Gypsy-53_GM-LTR	1236	1308	LTR/Gypsy	d	0.9014	2.0000	517
15746	15774	Sat-37_GMa	3	31	Simple/Sat	d	0.9655	1.0000	229
15893	15959	Gypsy-53_GM-LTR	1236	1305	LTR/Gypsy	d	0.8971	1.5000	498
16068	16099	Sat-38_GMa	4	35	Simple/Sat	d	0.9375	2.0000	265
16102	16127	Sat-26_GMa	1	26	Simple/Sat	d	1.0000	99.0000	216
16128	16162	Sat-38_GMa	1	35	Simple/Sat	d	0.8286	1.5000	216
16193	16227	Sat-38_GMa	1	35	Simple/Sat	d	0.9143	3.0000	267
16256	16290	Sat-38_GMa	1	35	Simple/Sat	d	0.8286	1.5000	216
16324	16355	Sat-38_GMa	4	35	Simple/Sat	d	0.9062	1.5000	251

16358	16383	Sat-26_GMa	1	26	Simple/Sat	d	1.0000	99.0000	216
16452	16483	Sat-38_GMa	4	35	Simple/Sat	d	0.9375	2.0000	265
16486	16511	Sat-26_GMa	1	26	Simple/Sat	d	1.0000	99.0000	216
16580	16611	Sat-38_GMa	4	35	Simple/Sat	d	0.9062	3.0000	240
16614	16639	Sat-26_GMa	1	26	Simple/Sat	d	1.0000	99.0000	216
16853	16880	Sat-43_GMa	1	28	Simple/Sat	d	0.9643	1.0000	224
16881	16951	Gypsy-118_GM-I	15597	15667	LTR/Gypsy	d	0.8732	1.1250	508
16973	17770	GmOgre_LTR	401	1215	LTR/Gypsy	d	0.9487	1.3704	6736

Alignment of the highest scored repeat with transposon database (CENSOR)

var_2557 length=17770 score = 37060.56

var_3476 length=14279 score=32133.22

(2nd highest scored repeat)

From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
2	1069	GmOgre_LTR	1	1067	LTR/Gypsy	c	0.9551	1.5556	9180
1070	1140	Gypsy-118_GM-I	15597	15667	LTR/Gypsy	c	0.8873	1.1429	521
1199	1308	GmOgre_LTR	1	110	LTR/Gypsy	c	0.9727	1.5000	998
1309	1351	Sat-43_GMa	1	43	Simple/Sat	c	0.9535	2.0000	347
1593	1627	Sat-38_GMa	1	35	Simple/Sat	c	0.9143	3.0000	255
1721	1752	Sat-38_GMa	4	35	Simple/Sat	c	0.9062	3.0000	226
1821	1846	Sat-26_GMa	1	26	Simple/Sat	c	1.0000	99.0000	216
1849	1880	Sat-38_GMa	4	35	Simple/Sat	c	0.9062	3.0000	226
1949	1974	Sat-26_GMa	1	26	Simple/Sat	c	1.0000	99.0000	216
1977	2010	Sat-38_GMa	1	35	Simple/Sat	c	0.8857	3.0000	218
2041	2075	Sat-38_GMa	1	35	Simple/Sat	c	0.8286	1.5000	207
2076	2101	Sat-26_GMa	1	26	Simple/Sat	c	1.0000	99.0000	216
2104	2135	Sat-38_GMa	4	35	Simple/Sat	c	0.9062	3.0000	226
2169	2203	Sat-38_GMa	1	35	Simple/Sat	c	0.8286	1.5000	207
2204	2229	Sat-26_GMa	1	26	Simple/Sat	c	1.0000	99.0000	216
2232	2263	Sat-38_GMa	4	35	Simple/Sat	c	0.9062	3.0000	226
2332	2357	Sat-26_GMa	1	26	Simple/Sat	c	1.0000	99.0000	216

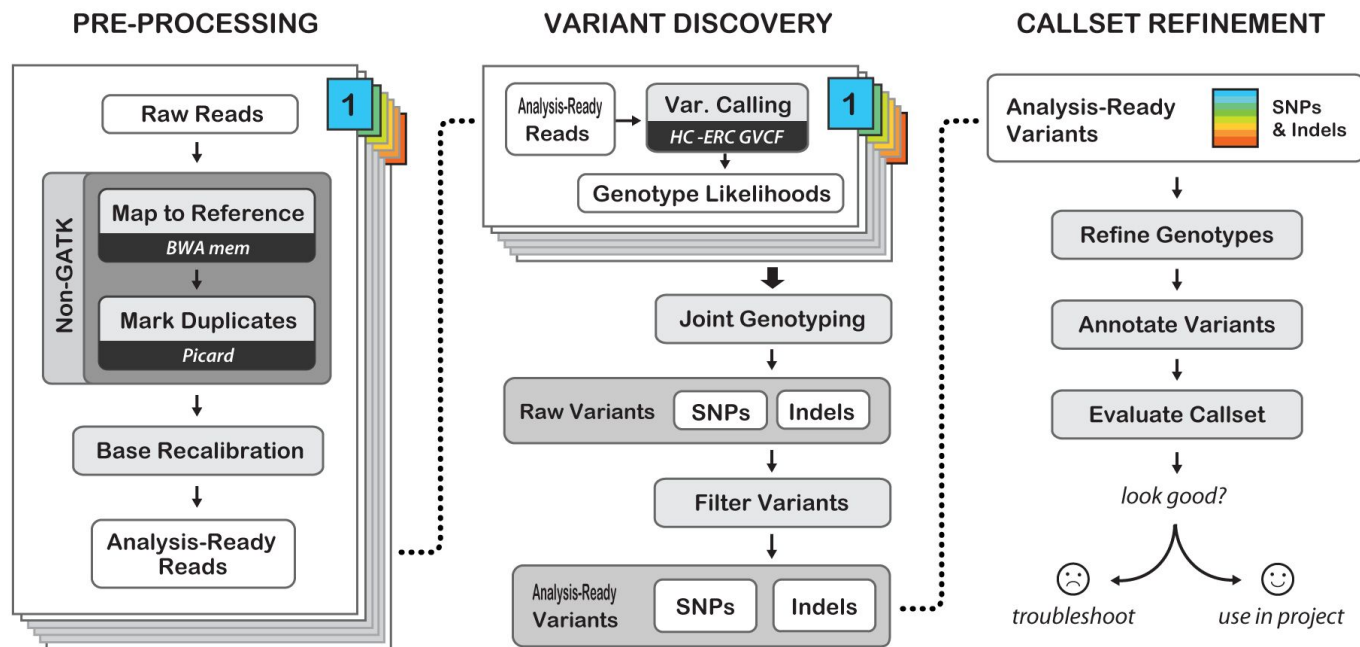
2500	2566	Gypsy-53_GM-LTR	1236	1305	LTR/Gypsy	c	0.8676	1.6000	456
2685	2713	Sat-37_GMa	3	31	Simple/Sat	c	0.8966	1.0000	203
2750	2819	Gypsy-53_GM-LTR	1236	1308	LTR/Gypsy	c	0.9014	2.0000	517
2852	4019	Gypsy-53_GM-I	13653	14793	LTR/Gypsy	c	0.7044	1.5434	2294
4184	4247	Gypsy-57_GR-LTR	453	509	LTR/Gypsy	d	0.7931	1.2222	242
4339	4468	Transib-2_HM	594	714	DNA/Transib	d	0.7120	1.6842	238
4640	4730	Gypsy-4_GR-I	5357	5436	LTR/Gypsy	c	0.7619	1.5000	243
4858	5169	Gypsy-53_GM-I	12329	12616	LTR/Gypsy	c	0.7010	1.5370	921
5173	5649	Gypsy-53_GM-I	11995	12513	LTR/Gypsy	c	0.7303	1.3626	1881
6918	7813	Copia-63_GM-LTR	23	1074	LTR/Copia	d	0.7864	1.4742	2933
7883	7977	Crack-2_HM	4269	4370	NonLTR/Daphne	d	0.7766	2.6667	211
8190	8347	GmOgre_I	4500	4658	LTR/Gypsy	c	0.6541	2.0769	445
8644	9254	GmOgre_I	3495	4190	LTR/Gypsy	c	0.7042	1.6893	1904
9538	9584	L1-52_DR	2062	2108	NonLTR/L1	c	0.7660	2.2000	212
9634	9878	GmOgre_I	2941	3176	LTR/Gypsy	c	0.6765	1.7805	670
10106	10138	L2-18_HRo	2472	2504	NonLTR/L2	d	0.8788	2.0000	216
10751	10869	GmOgre_I	662	789	LTR/Gypsy	c	0.6480	1.5833	343
10902	13699	GmOgre_I	1	2868	LTR/Gypsy	c	0.7667	1.5573	13040
13719	14279	GmOgre_LTR	1066	1627	LTR/Gypsy	c	0.9538	1.4000	4772

var_2735 length=19967 score=24935.76 (3rd highest scored repeat)

<u>Name</u>	<u>From</u>	<u>To</u>	<u>Name</u>	<u>From</u>	<u>To</u>	<u>Class</u>	<u>Dir</u>	<u>Sim</u>	<u>Pos/Mm:Ts</u>	<u>Score</u>
var_2735_branch_1	2	75	Gypsy-120_GM-I	6251	6327	LTR/Gypsy	d	0.7500	1.3077	327
var_2735_branch_1	78	141	Gypsy-120_GM-LTR	1	66	LTR/Gypsy	d	0.7692	1.4444	273
var_2735_branch_1	146	1089	SIRE1_LTR	100	999	LTR/Copia	c	0.8853	2.2973	5845
var_2735_branch_1	1090	1294	SIRE1_LTR	1	205	LTR/Copia	c	0.9463	1.2222	1725
var_2735_branch_1	1295	8756	SIRE1_INT	1	7258	LTR/Copia	c	0.9398	1.5543	57816
var_2735_branch_1	8757	9700	SIRE1_LTR	100	999	LTR/Copia	c	0.8853	2.2973	5845
var_2735_branch_1	9701	9905	SIRE1_LTR	1	205	LTR/Copia	c	0.9463	1.2222	1725

Genome Analysis Toolkit (GATK) Workflow used for comparison

Soybean reads with length 151. Tool used in alignment step: Novoalign (higher accuracy and versatility than BWA mem). Marking duplicate step: Novosort. Variant calling step: HaploTypeCaller.



Known Validated SNPs for Comparison

- Choi, IK., et al. (2007). A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176: 685–696. doi: 10.1534/genetics.107.070821
 - Designed PCR primers from known unigenes and expressed sequence tags
 - Isolated DNA from 6 Soybean Genotypes and performed PCR
 - Only the primer pairs that amplified a single discrete product on a gel were used for further analysis
 - PCR products were then digested and sequenced via Sanger sequencing
 - Sequence traces were then aligned using Phred and Phrap software
 - SNP discovery in the sequence alignments using machine learning algorithm based on PolyBayesSNP discovery software

Known validated SNPs for comparison

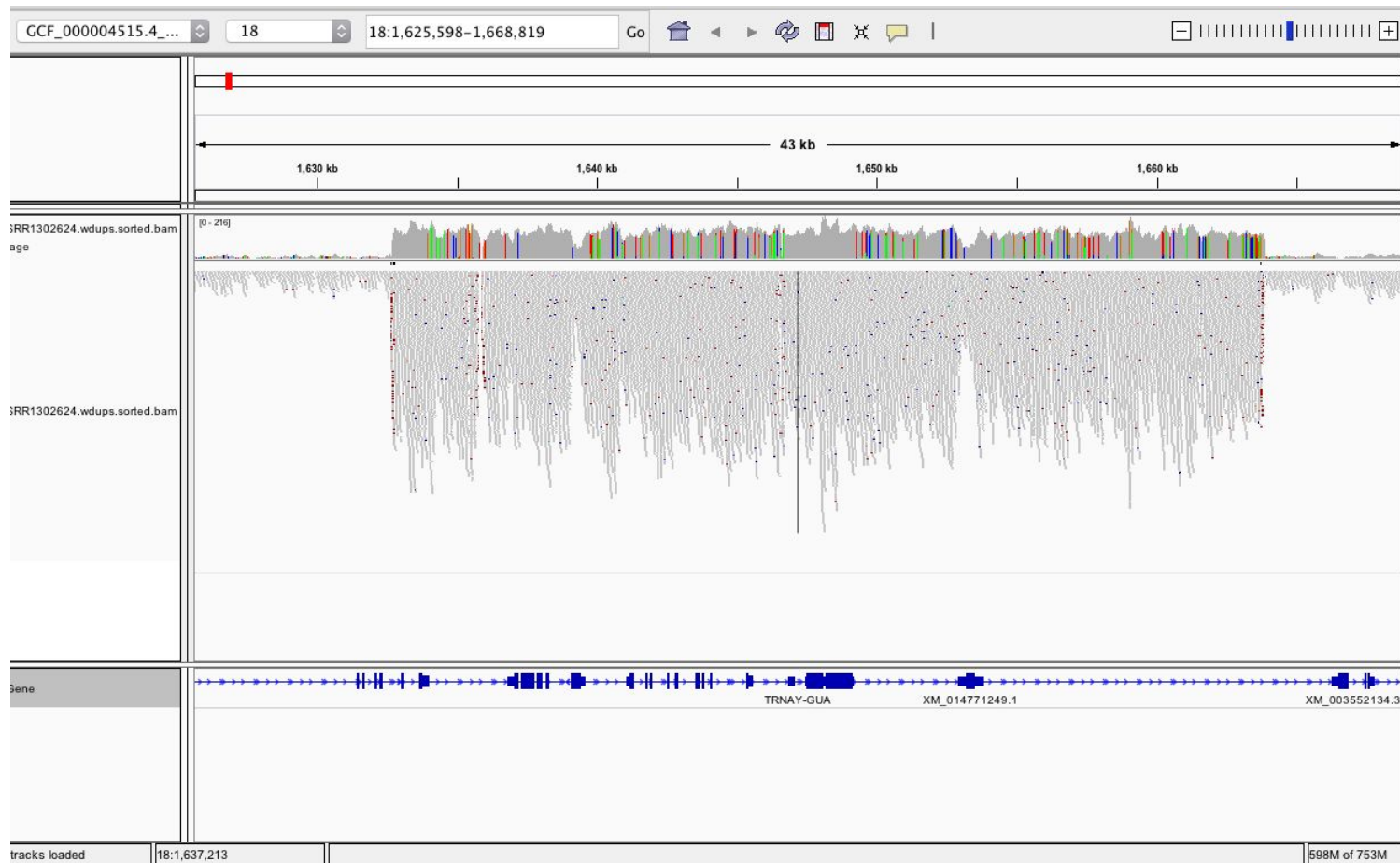
- PI574486: 85 Validated SNPs, Cortex found 42 (49%)
- PI398881: 44 Validated SNPs, Cortex found 32 (73%)
- Magellan: 71 Validated SNPs, Cortex found 14 (20%)
- Maverick: 49 Validated SNPs, Cortex found 1 (2%)

Cortex does not do well with SNP finding in low coverage samples. Validated SNPs are from Choi et al. (2007).

Next step: *Rhg1* Locus Repeats for PI209322

- Lee, TG., et al. (2015). Evolution and selection of *Rhg1*, a high copy-number variant nematode-resistance locus. *Molecular Ecology* 24: 1774-1791. doi: 10.1111/mec13138.
 - Describes that PI 209332 has ten copies of this locus
 - The Williams Reference was susceptible and did not have as many repeats as this sample
- We are currently running Cortex on 43 samples and looking to find this region.

IGV view of sample PI209332 at the Rhg1 locus



Acknowledgements

Kathleen Keating

Trained Junyu and I on genome assembly methods

Gloria Rendon

Wrote the GATK Bash workflow for use on Blue Waters and for training Junyu and I on variant calling workflows

Azza Ahmed

Contributed to converting GATK Bash workflow for use on iForge