NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Using SMRT Iso-Seq Sequencing to Dissect Polyploid Transcriptomes: Lessons Learned from Tetra- and Hexaploid Blueberries

Hamid Ashrafi, PhD

Hamed Bostan, PhD

North Carolina State University

PacBio East Coast UGM - June 2017

https://blueberry.cals.ncsu.edu/

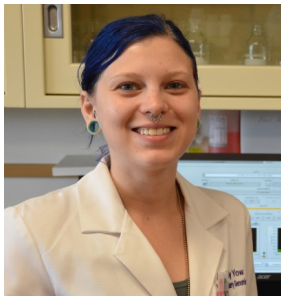
Find us on Facebook https://www.facebook.com/TeamVaccinium/

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Acknowledgements

Dr. George Yuan
PACIFIC BIOSCIENCES®



Ms. Ashley Yow
NCSU

Dr. Massimo Iorizzo
NCSU

Dr. Liz Tseng
PACIFIC BIOSCIENCES®

Dr. Rishi Aryal
NCSU

Dr. Hamed Bostan
NCSU

Funding Agencies

NC Blueberry Council NC Dept. Ag & CS

NC STATE UNIVERSITY Plants for Human Health INSTITUTE

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Blueberry Ploidy Level and Genome Size

- Naturally occurring 2X, 4X and 6X genomes
  - Most commercial varieties are 4X and 6X
  - Diploid blueberries are either ornamental or wild with no fresh or processed fruit commercial production use, but they have been used in breeding
- It is possible to cross a 2X with 4X and yet obtain a 4X progeny (unreduced gamete)
- It is possible to obtain a 5X genome by crossing a 4X and a 6X genome
  - But a 3X blueberry is rare or have not been successful
- Most breeding efforts are focused on 4X and 6X genomes
- It is believed that 4X blueberry is autotetraploid and 6X is a natural allohexaploid (2X + 4X ?)
- DNA content of a diploid genome is 2C ~1.37 pg, and the monoploid genome size is estimated ~670 Mb

NC STATE UNIVERSITY

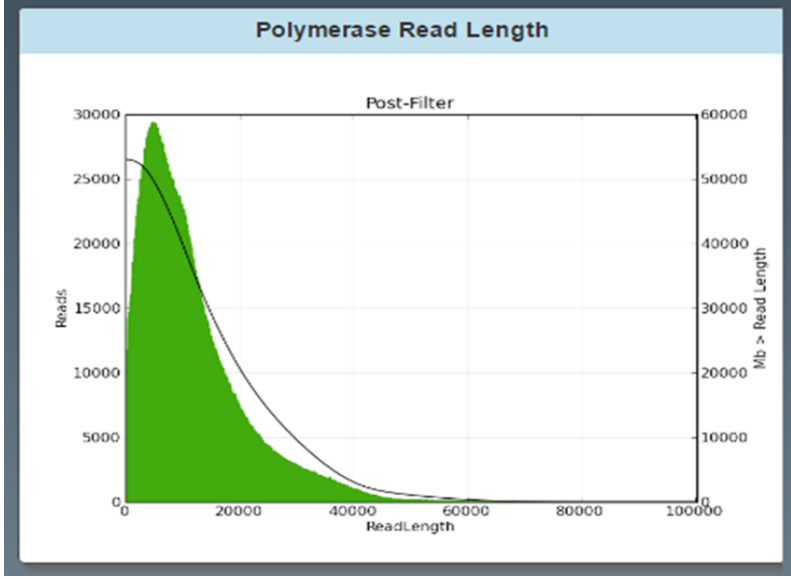
PACIFIC BIOSCIENCES®

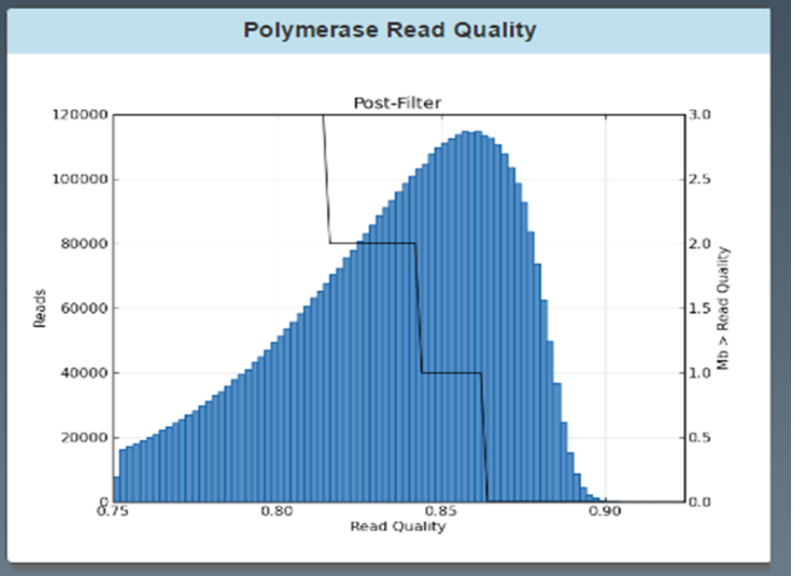
A Total of 54 RS II SMRT Cells Were Used for Blueberry Diploid Genome Assembly

SMRT Cells: 54 Movies: 54 Print Tech Support Files Email

Filtering

Metrics	Pre-Filter	Post-Filter
Polymerase Read Bases	55665028395	53234439147
Polymerase Reads	8115768	4540817
Polymerase Read N50	16319	16562
Polymerase Read Length	6858	11723
Polymerase Read Quality	0.49	0.834

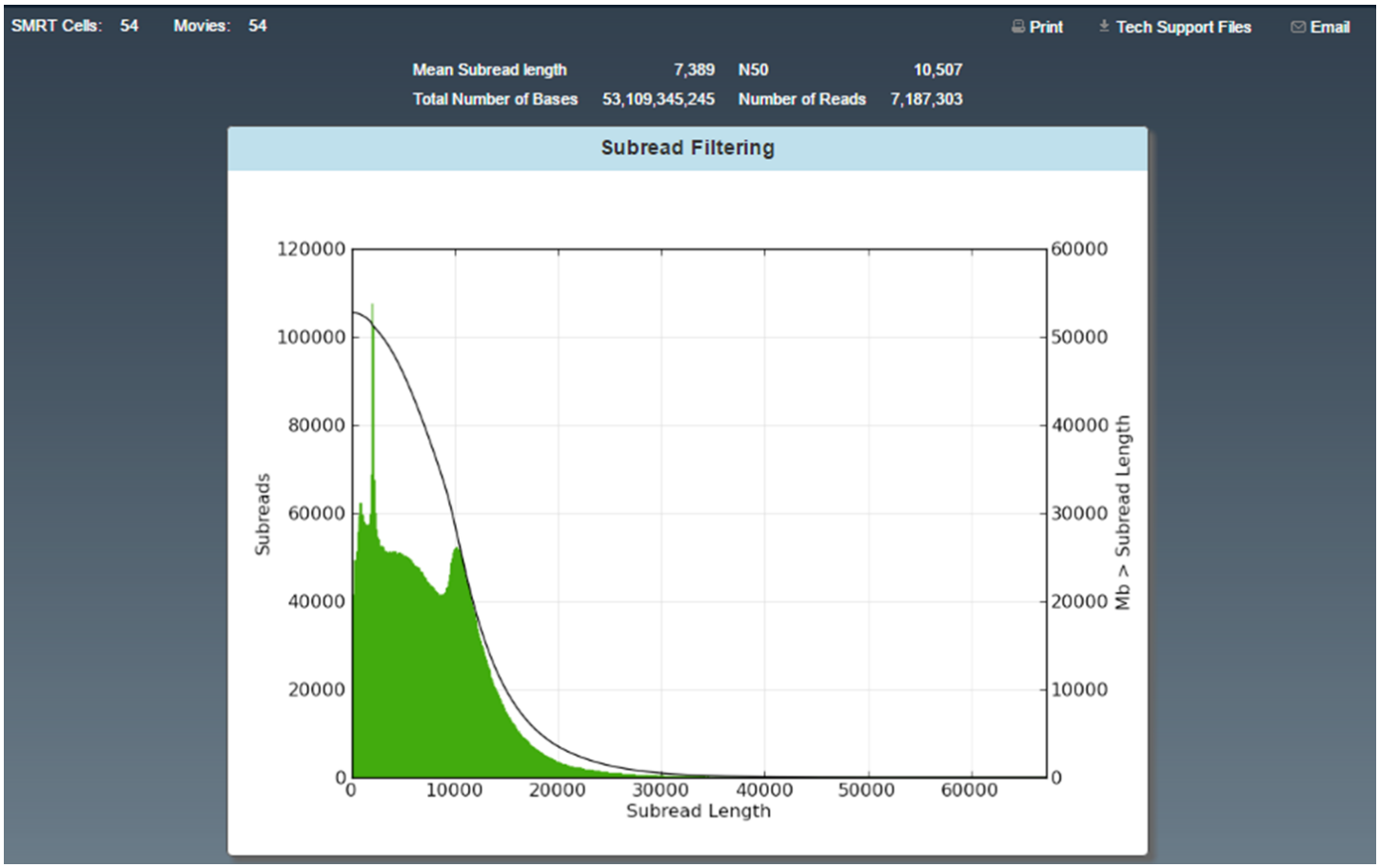




NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

A Total of 54 RS II SMRT Cells Were Used for Blueberry Diploid Genome Assembly



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Blueberry Genome Assembly by Falcon – Falcon Unzip

	Assembly	Total length (>= 0 bp)	# contigs	Largest contig	Total length	GC (%)	N50	L50
Run 1: Auto	a_ctg	361,971	36	39,382	361,893	39	15,275	9
	p_ctg	21,191,558	2,666	137,366	21,119,817	40	16,092	377
Run 2	a_ctg	106,730,494	4,105	152,359	106,730,020	38	31,169	1,147
	p_ctg	509,653,341	9,120	919,765	509,648,244	38	123,572	1,160
Run 3	a_ctg	84,496,567	3,930	106,558	84,496,093	38	25,130	1,127
	p_ctg	442,011,250	9,629	743,506	442,005,280	38	98,281	1,303
Run 4	a_ctg	23,833,333	1,467	63,821	23,833,086	38	17,711	491
	p_ctg	258,423,456	8,004	512,507	258,401,652	38	59,073	1,251
Run 5	a_ctg	127,543,692	4,328	246,408	127,542,942	38	35,511	1,193
	p_ctg	534,691,250	7,042	1,104,661	534,674,473	38	174,233	889
Run 6	a_ctg	37,663,019	2,454	70,925	37,661,940	38	17,074	772
	p_ctg	317,135,666	10,114	674,253	317,050,128	38	65,456	1,377
Run 7	a_ctg	5,920,512	473	74,955	5,919,651	38	13,824	150
	p_ctg	166,123,075	9,797	610,472	166,047,246	39	31,850	1,412
Run 8	a_ctg	150,735,294	3,437	247,829	150,735,131	38	58,249	798
	p_ctg	636,345,139	4,656	2,509,334	636,342,635	38	304,786	595
Run 9	a_ctg	536,745	40	28,422	536,363	38	16,315	14
	p_ctg	16,392,338	1,357	120,648	16,382,771	41	18,948	256
Run 10	a_ctg	597,910	41	29,506	597,910	38	17,838	14
	p_ctg	15,509,230	1,283	187,471	15,497,296	40	18,964	246
Run 11	a_ctg	33,120,782	1,943	84,803	33,120,618	38	18,653	629
	p_ctg	291,072,902	7,948	916,704	291,056,698	38	67,389	1,236
Run 12	a_ctg	33,523,822	1,968	84,803	33,523,658	38	18,552	638
	p_ctg	289,991,623	7,952	757,340	289,973,641	38	67,118	1,227
Run 13	a_ctg	150,816,695	3,431	247,829	150,816,532	38	58,249	797
	p_ctg	636,287,673	4,650	2,509,334	636,285,169	38	305,215	593
Run 14	a_ctg	99,228,819	4,179	132,812	99,228,502	38	27,976	1,199
	p_ctg	470,560,399	7,598	1,286,657	470,547,111	38	132,876	1,038
Run 15	a_ctg	33,120,455	1,944	84,803	33,120,291	38	18,614	631
	p_ctg	291,267,619	7,953	635,651	291,251,415	38	67,268	1,238
Run 16: Final	a_ctg	168,843,234	3,138	459,364	168,842,782	38	76,880	661
	p_ctg	658,738,700	4,536	2,585,193	658,729,482	38	426,430	430
CANU 54 Cells	blueberry.contigs	811,103,462	10,783	1,097,648	811,103,462	39	123,295	1,684
	blueberry.bubbles	776,178	32	46,866	776,178	37	24,165	14

	Falcon Unzip		Quiver	
Assembly	all_p_ctg	all_h_ctg	Consensus-P-quivered	Consensus-H-quivered
# contigs (>= 0 bp)	2,939	10,479	2,939	10,479
# contigs (>= 5000 bp)	2,939	9,964	2,939	9,965
# contigs (>= 10000 bp)	2,938	8,872	2,938	8,873
# contigs (>= 50000 bp)	2,398	2,209	2,398	2,221
Largest contig	2,500,485	680,366	2,508,523	682,940
Total length	633,240,445	415,765,557	635,573,105	417,065,786
GC (%)	38	39	38	39
N50	422,065	68,661	423,700	68,980
N75	185,127	29,958	185,727	30,071
L50	420	1,538	420	1,535
L75	987	3,891	987	3,886
# N's per 100 kbp	0	0	0	0

P fraction of GS(%)	A fraction of GS(%)	A fraction of P(%)
94.86	62.25	65.62

code 3 (false insert on the genome and gap on the transcript)

000160F 647243 647244 000160F 647243 647264 000160F_pilon 647342 GAGAGAGAGAGAGAGAGAGAGA . 000160F 647173 647533
000160F genome1 exon 647173 647533 93 + .
ID=c1138/f12p53/4042.mrna1.exon1;Name=c1138/f12p53/4042;Parent=c1138/f12p53/4042.mrna1;Target=c1138/f12p53/4042 1 339 + 1

Note: after removing of the sequence on the unzip genome, the gap disappeared so the sequence was repeated 2 times.

False INSERTION causing gap

000160F

Sequence ID: Query_186133 Length: 764783 Number of Matches: 9
Range 1: 647173 to 656596

Score	Expect	Identities	Gaps	Strand	Frame
17219 bits(9324)	0.0()	9402/9433(99%)	31/9433(0%)	Plus/Plus	

Features:

```
Query  1       CACTAGTCAAATGACAGCCGTTTTTTGATGCAACCATCGTTAGCCTCTCAGTATCTGAGT  60
Sbjct  647173  CACTAGTCAAATGACAGCCGTTTTTTGATGCAACCATCGTTAGCCTCTCAGTATCTGAGT  647232

Query  61      AATACTGTTT----------------------gagagagagagagagagagagagagaga  98
Sbjct  647233  AATACTGTTTGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA  647292

Query  99      gTAGTTTGGAAGGAAAGTGAAGGAAAAAGGGATGGGGATCCAAAAGAAGAGCTTTGGGTC  158
Sbjct  647293  GTAGTTTGGAAGGAAAGTGAAGGAAAAAGGGATGGGGATCCAAAAGAAGAGCTTTGGGTC  647352

Query  159     ATTTCGATCCATTTTCATGCATTCGGATGGCATTGATAAGCTGCTAATGGGACTCGGATT  218
Sbjct  647353  ATTTCGATCCATTTTCATGCATTCGGATGGCATTGATAAGCTGCTAATGGGACTCGGATT  647412

Query  219     CCTCGGGTCACTCGGCGACGGGCTTTCTATGCCCGTCATGTTGCTCTTCACTAGCAAGAT  278
Sbjct  647413  CCTCGGGTCACTCGGCGACGGGCTTTCTATGCCCGTCATGTTGCTCTTCACTAGCAAGAT  647472

Query  279     TATGAACAATATCGGGGGCGCTTCCGCCGGGGCCAAAAACTTCTCTCATACCATCAACAA  338
Sbjct  647473  TATGAACAATATCGGGGGCGCTTCCGCCGGGGCCAAAAACTTCTCTCATACCATCAACAA  647532

Query  339     GGTCCCTTCCCTCTCATAACtatatatatatatatatatatatatatatatatatatata  398
Sbjct  647533  GGTCCCTTCCCTCTCATAACTATATATATATATATATATATATATATATATATATATA  647592

Query  399     tatatatatatatatatatatatatatatatcacacatgcgtacatatgtatatatatCA  458
Sbjct  647593  TATATATATATATATATATATATATATATCACACATGCGTACATATGTATATATATCA  647652
```

000160F_pilon

Sequence ID: Query_179209 Length: 764876 Number of Matches: 9
Range 1: 647272 to 656682

Score	Expect	Identities	Gaps	Strand	Frame
17379 bits(9411)	0.0()	9411/9411(100%)	0/9411(0%)	Plus/Plus	

Features:

```
Query  1       CACTAGTCAAATGACAGCCGTTTTTTGATGCAACCATCGTTAGCCTCTCAGTATCTGAGT  60
Sbjct  647272  CACTAGTCAAATGACAGCCGTTTTTTGATGCAACCATCGTTAGCCTCTCAGTATCTGAGT  647331

Query  61      AATACTGTTTgagagagagagagagagagagagagagagTAGTTTGGAAGGAAAGTGAAG  120
Sbjct  647332  AATACTGTTTGAGAGAGAGAGAGAGAGAGAGAGAGAGAGTAGTTTGGAAGGAAAGTGAAG  647391

Query  121     GAAAAAGGGATGGGGATCCAAAAGAAGAGCTTTGGGTCATTTCGATCCATTTTCATGCAT  180
Sbjct  647392  GAAAAAGGGATGGGGATCCAAAAGAAGAGCTTTGGGTCATTTCGATCCATTTTCATGCAT  647451

Query  181     TCGGATGGCATTGATAAGCTGCTAATGGGACTCGGATTCCTCGGGTCACTCGGCGACGGG  240
Sbjct  647452  TCGGATGGCATTGATAAGCTGCTAATGGGACTCGGATTCCTCGGGTCACTCGGCGACGGG  647511

Query  241     CTTTCTATGCCCGTCATGTTGCTCTTCACTAGCAAGATTATGAACAATATCGGGGGCGCT  300
Sbjct  647512  CTTTCTATGCCCGTCATGTTGCTCTTCACTAGCAAGATTATGAACAATATCGGGGGCGCT  647571

Query  301     TCCGCCGGGGCCAAAAACTTCTCTCATACCATCAACAAGGTCCCTTCCCTCTCATAACta  360
Sbjct  647572  TCCGCCGGGGCCAAAAACTTCTCTCATACCATCAACAAGGTCCCTTCCCTCTCATAACTA  647631

Query  361     tatatatatatatatatatatatatatatatatatatatatatatatatatatatatata  420
Sbjct  647632  TATATATATATATATATATATATATATATATATATATATATATATATATATATATATA  647691

Query  421     tatatatatcacacatgcgtacatatgtatatatatCACCACGCAATGAACTTTTGAGTA  480
Sbjct  647692  TATATATATCACACATGCGTACATATGTATATATATCACCACGCAATGAACTTTTGAGTA  647751
```

code 5 (false insert on the genome and stopping the transcript mapping to continue- partial mapping)

000886F 22018 22019 000886F 22018 22053 000886F_pilon 22009 CAACAACAACAAAACATAACCATAGTCCCAAAGGGT . 000886F 21276
24514 000886F genome1 exon 21276 24514 98 - .
ID=c18093/f2p28/4207.mrna1.exon1;Name=c18093/f2p28/4207;Parent=c18093/f2p28/4207.mrna1;Target=c18093/f2p28/4207 1 3258 + 1

Note: in the unzip genome, the mapping stopped when reached to the sequence at base 22017 but in the pilon version the mapping continues since that sequence (interrupt) was removed so the gene is predicted complete.

False INSERTION causing stop

000886F

Sequence ID: Query_210919 Length: 200276 Number of Matches: 4
Range 1: 15008 to 22017

Score	Expect	Identities	Gaps	Strand	Frame
12936 bits(7005)	0.0()	7010/7012(99%)	2/7012(0%)	Plus/Minus	

Features:

```
Query  2462   TGTTCAGTCCACTATGATGTCTCCATGAATCTCTTATCAGCTCGAATCTTCAAAGAACTC  2521
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22017  TGTTCAGTCCACTATGATGTCTCCATGAATCTCTTATCAGCTCGAATCTTCAAAGAACTC  21958

Query  2522   AGGTTGTTCAAGAAATCTCTATGGGCATTTCGCACTTTGGGGTATAGCGGTTGTTGCTGC  2581
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21957  AGGTTGTTCAAGAAATCTCTATGGGCATTTCGCACTTTGGGGTATAGCGGTTGTTGCTGC  21898

Query  2582   TTGTTTCATTTCTTCTCTAAGACACTCACCCAGGCTCTTGTTTCTCATTATCAAATGTCT  2641
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21897  TTGTTTCATTTCTTCTCTAAGACACTCACCCAGGCTCTTGTTTCTCATTATCAAATGTCT  21838

Query  2642   TTTGATATGTGTTCTTTAGCCGATAAGTGATTACTTAACAATTACCTGTGTTAAGTTTCT  2701
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21837  TTTGATATGTGTTCTTTAGCCGATAAGTGATTACTTAACAATTACCTGTGTTAAGTTTCT  21778

Query  2702   ACCTTTGTCTGACGGATGTGCACAATTATCCCAACTAGTATTAGTTCTGTGGTCCAACAG  2761
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21777  ACCTTTGTCTGACGGATGTGCACAATTATCCCAACTAGTATTAGTTCTGTGGTCCAACAG  21718

Query  2762   AACTCATGATGAGGCTTTCTGTTTCCTATTCAAGTATTGTGAAGAACTGAAGTTTGACTT  2821
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21717  AACTCATGATGAGGCTTTCTGTTTCCTATTCAAGTATTGTGAAGAACTGAAGTTTGACTT  21658

Query  2822   CTTCAATAACTGCACTGATTTTGAGAAGACCATATCGGAGTGGTGCTTCTGTACCCTTCA  2881
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21657  CTTCAATAACTGCACTGATTTTGAGAAGACCATATCGGAGTGGTGCTTCTGTACCCTTCA  21598

Query  2882   ATTTGAGGATAAAGAGGACAAACATTCTTGAATTATTCATTTCATGAATGCAGCTTATAA  2941
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21597  ATTTGAGGATAAAGAGGACAAACATTCTTGAATTATTCATTTCATGAATGCAGCTTATAA  21538

Query  2942   CTTAAGTCCCTTTCAACATTGGTGGAAAGAGATGTTCATTCAAAGATTTGGAAAATGTGG  3001
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21537  CTTAAGTCCCTTTCAACATTGGTGGAAAGAGATGTTCATTCAAAGATTTGGAAAATGTGG  21478
```

000886F_pilon

Sequence ID: Query_245069 Length: 200282 Number of Matches: 3
Range 1: 14997 to 24469

Score	Expect	Identities	Gaps	Strand	Frame
17494 bits(9473)	0.0()	9473/9473(100%)	0/9473(0%)	Plus/Minus	

Features:

```
Query  1      CTTGTCAAAACGGGCGGTGTGCGATCTCCAATCGGTATAAACCGTCGCATACGAGACGAA  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24469  CTTGTCAAAACGGGCGGTGTGCGATCTCCAATCGGTATAAACCGTCGCATACGAGACGAA  24410

Query  61     TCTGTAAAACAGATAAAGGATCTGAGGTTGAAAATCTTTGACAATTTGAATCACGATCAG  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24409  TCTGTAAAACAGATAAAGGATCTGAGGTTGAAAATCTTTGACAATTTGAATCACGATCAG  24350

Query  121    GTAAAACGACAAACGTGAATTAGCATAATTGTGTCGATTTTATCACGCTTTAAGGTCATT  180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24349  GTAAAACGACAAACGTGAATTAGCATAATTGTGTCGATTTTATCACGCTTTAAGGTCATT  24290

Query  181    ATTTAGGGTTGCTGTAAATTAGATGATTGTCGTGTCTGACTAGTTCAATCGTGTGATTAA  240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24289  ATTTAGGGTTGCTGTAAATTAGATGATTGTCGTGTCTGACTAGTTCAATCGTGTGATTAA  24230

Query  241    GCGCATATGTTGTTACATTACAAGTTGTGAACTAGGATATTGAGGTGCTAGGGTTTGAGT  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24229  GCGCATATGTTGTTACATTACAAGTTGTGAACTAGGATATTGAGGTGCTAGGGTTTGAGT  24170

Query  301    TGATTCTGAAAGCATTTAGATTGGCGGCCTTTGATTTGATTGAGCCTAGGTGAGTTGGCG  360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24169  TGATTCTGAAAGCATTTAGATTGGCGGCCTTTGATTTGATTGAGCCTAGGTGAGTTGGCG  24110

Query  361    TTTTCTTAGTTTTGAGTTCAATCTTGAATGCCAAGTATTTTAGATTTGGGTTTTGGGGCT  420
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24109  TTTTCTTAGTTTTGAGTTCAATCTTGAATGCCAAGTATTTTAGATTTGGGTTTTGGGGCT  24050

Query  421    AATTACTTGCTGAAGGAACAATCGTGGTGATTTCTATGTTCCTATGTCTTTTGGTGGCAG  480
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  24049  AATTACTTGCTGAAGGAACAATCGTGGTGATTTCTATGTTCCTATGTCTTTTGGTGGCAG  23990

Query  481    GGGAGCATTGGGAATTGAGTTGGGTGTTGTTTCTGATGGGGCAACTGAAACTGAGTTGGG  540
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

code 6 (false deletion on the genome and gap on the genome when mapping the tanscript)

000104F 111391 111392 000104F 111391 000104F_pilon 111417 111427 . TCTCTCTCTCT 000104F 110997 111584 000104F
genome1 exon 110997 111584 98 - .
ID=c13750/f25p70/3400.mrna1.exon1;Name=c13750/f25p70/3400;Parent=c13750/f25p70/3400.mrna1;Target=c13750/f25p70/3400 1 591 + 1

Note: on the unizp genome, this sequence is missing which caused a gap on position 111391, when the sequence is inserted/corrected on the pilon genome, the gap is removed and replaced with the matches.

False DELETION causing gap

000104F

Sequence ID: Query_213451 Length: 924261 Number of Matches: 5
Range 1: 99245 to 111584

Score	Expect	Identities	Gaps	Strand	Frame
22718 bits(12302)	0.0()	12340/12355(99%)	15/12355(0%)	Plus/Minus	

Features:

```
Query  1       AAGTTGAAAAATACAGCTCATTGTGTTTTCAAGAAAAGAGGTTGCAAATCCAGCTATTAC  60
Sbjct  111584  AAGTTGAAAAATACAGCTCATTGTGTTTTCAAGAAAAGAGGTTGCAAATCCAGCTATTAC  111525

Query  61      AGGACACTGGAAAAAATTCCTCACTGTCGATCATTCCACAAATCCCAAATCCCAATTGCC  120
Sbjct  111524  AGGACACTGGAAAAAATTCCTCACTGTCGATCATTCCACAAATCC-AAATCCCAATTGCC  111466

Query  121     ATAAGATCTCTTCCCCCTCGGCGGCCTAATCTTTGACGTTTGCAATCTTATGTagagaga  180
Sbjct  111465  ATAAGATCTCTTCCCCCTCGGCGGCCTAATCTTTGACGTTTGCAATCTTATGTAGAGAG-  111407

Query  181     gagagagagagagagagagagagagagagagagGGTCCGTTTTAGTTGGGAGAGCCCACC  240
Sbjct  111406  ----------GAG-GAGAGAGAGAGAG-GAGAGGGTCCGTTTTAGTTGGGAGAGCCCACC  111359

Query  241     ACCAAGAGACGTGGTGATTGCTTCGATCTACCTTTTTACTTATCTTGATCAGATAGAGTT  300
Sbjct  111358  ACCAAGAGACGTGGTGATTGCTTCGATCTACCTTTTTACTTATCTTGATCAGATAGAGTT  111299

Query  301     CTTTTGGGTTCTCTGTAGCCATTGACTTCTTTGGACTGGGAATTGAGAAAATTTGGGATA  360
Sbjct  111298  CTTTTGGGTTCTCTGTAGCCATTGACTTCTTTGGACTGGGAATTGAGAAAATTTGGGATA  111239

Query  361     TTTGTTGTGGTGGTGGTGGTTTTGGTGGCTCAGTCTAGATGGGCTCGAAGCCGTTTTCTA  420
Sbjct  111238  TTTGTTGTGGTGGTGGTGGTTTTGGTGGCTCAGTCTAGATGGGCTCGAAGCCGTTTTCTA  111179

Query  421     TCCGGCTCCGACCTATCGTTCTTTAGAGACTTACTGGGACACCGACGACGACGCCCCTGG  480
Sbjct  111178  TCCGGCTCCGACCTATCGTTCTTTAGAGACTTACTGGGACACCGACGACGACGCCCCTGG  111119

Query  481     CCCACGCTGCGCCCACACTCTCACCGCCATCGCCGCTACTAAAACCCACGGCCCCCGCCT  540
Sbjct  111118  CCCACGCTGCGCCCACACTCTCACCGCCATCGCCGCTACTAAAACCCACGGCCCCCGCCT  111059
```

000104F_pilon

Sequence ID: Query_221427 Length: 924449 Number of Matches: 5
Range 1: 99270 to 111624

Score	Expect	Identities	Gaps	Strand	Frame
22816 bits(12355)	0.0()	12355/12355(100%)	0/12355(0%)	Plus/Minus	

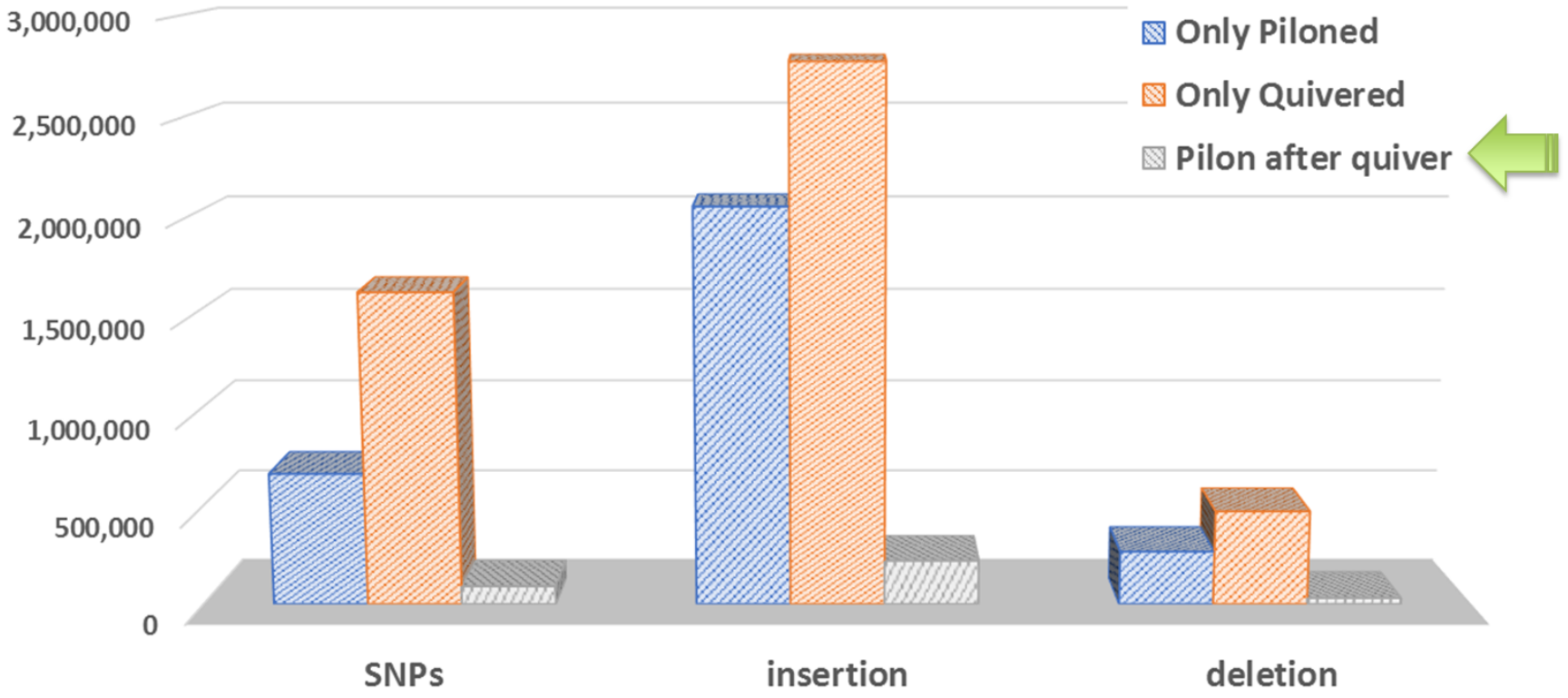
Features:

```
Query  1       AAGTTGAAAAATACAGCTCATTGTGTTTTCAAGAAAAGAGGTTGCAAATCCAGCTATTAC  60
Sbjct  111624  AAGTTGAAAAATACAGCTCATTGTGTTTTCAAGAAAAGAGGTTGCAAATCCAGCTATTAC  111565

Query  61      AGGACACTGGAAAAAATTCCTCACTGTCGATCATTCCACAAATCCCAAATCCCAATTGCC  120
Sbjct  111564  AGGACACTGGAAAAAATTCCTCACTGTCGATCATTCCACAAATCCCAAATCCCAATTGCC  111505

Query  121     ATAAGATCTCTTCCCCCTCGGCGGCCTAATCTTTGACGTTTGCAATCTTATGTagagaga  180
Sbjct  111504  ATAAGATCTCTTCCCCCTCGGCGGCCTAATCTTTGACGTTTGCAATCTTATGTAGAGAGA  111445

Query  181     gagagagagagagagagagagagagagagagagGGTCCGTTTTAGTTGGGAGAGCCCACC  240
Sbjct  111444  GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGGGTCCGTTTTAGTTGGGAGAGCCCACC  111385

Query  241     ACCAAGAGACGTGGTGATTGCTTCGATCTACCTTTTTACTTATCTTGATCAGATAGAGTT  300
Sbjct  111384  ACCAAGAGACGTGGTGATTGCTTCGATCTACCTTTTTACTTATCTTGATCAGATAGAGTT  111325

Query  301     CTTTTGGGTTCTCTGTAGCCATTGACTTCTTTGGACTGGGAATTGAGAAAATTTGGGATA  360
Sbjct  111324  CTTTTGGGTTCTCTGTAGCCATTGACTTCTTTGGACTGGGAATTGAGAAAATTTGGGATA  111265

Query  361     TTTGTTGTGGTGGTGGTGGTTTTGGTGGCTCAGTCTAGATGGGCTCGAAGCCGTTTTCTA  420
Sbjct  111264  TTTGTTGTGGTGGTGGTGGTTTTGGTGGCTCAGTCTAGATGGGCTCGAAGCCGTTTTCTA  111205

Query  421     TCCGGCTCCGACCTATCGTTCTTTAGAGACTTACTGGGACACCGACGACGACGCCCCTGG  480
Sbjct  111204  TCCGGCTCCGACCTATCGTTCTTTAGAGACTTACTGGGACACCGACGACGACGCCCCTGG  111145

Query  481     CCCACGCTGCGCCCACACTCTCACCGCCATCGCCGCTACTAAAACCCACGGCCCCCGCCT  540
Sbjct  111144  CCCACGCTGCGCCCACACTCTCACCGCCATCGCCGCTACTAAAACCCACGGCCCCCGCCT  111085
```

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Comparison of Base Correction Using Different Methods



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Worth to mention that:

The total number of bases Pilon corrected was < 0.07% of the total genome size assembled

So if believe in Pilon, after Falcon assembly and Quivering, we already have a genome assembled of over 99.93% accuracy.

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Iso-Seq Project

Things to consider

1. Choice germplasm source for RNA extraction
   - Diploid (for genome annotation as well as comparison)
   - Tetraploid
   - Hexaploid
2. Variety or cultivar
   - W85-20 *V. caesariense* (N.J. blueberry)
   - O'Neal *V. corymbosum* (native to north east of the U.S.)
   - Premier *V. virgatum* or *ashei* (native to southern eastern U.S.)

3. Tissues

NC STATE UNIVERSITY

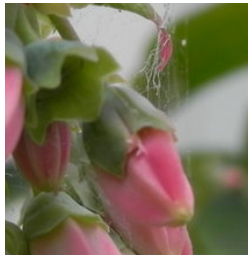
PACIFIC BIOSCIENCES®

Iso-Seq Project

Tissue Samples Were Collected from Field Grown Blueberries; Leaf, Flower, Fruit, Root

Leaf

Flower Stage 1



Flower Stage 3

Flower Stage 5

Fruit Stage 1

Fruit Stage 2

Fruit Stage 3

Fruit Stage 4

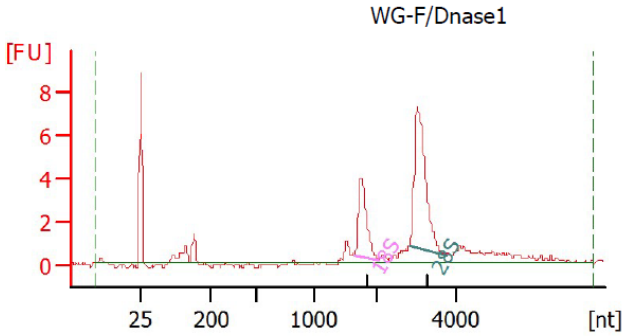
Root

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Total RNA Extraction Was Attempted by Different Kits

- Sigma Plant RNA extraction Kit
- Bioanalyzer was used to check the quality of RNA
- The same RNA for both Illumina and Iso-Seq libraries
- Iso-Seq libraries were made with size selection option
- KAPA stranded RNA-Seq Kit to make Illumina libraries



Overall Results for sample 3 : WG-F/Dnase1

RNA Area:	46.8
RNA Concentration:	36 ng/µl
rRNA Ratio [28s / 18s]:	2.3
RNA Integrity Number (RIN):	9.7 (B.02.08)
Result Flagging Color:	
Result Flagging Label:	RIN: 9.70

Fragment table for sample 3 : WG-F/Dnase1

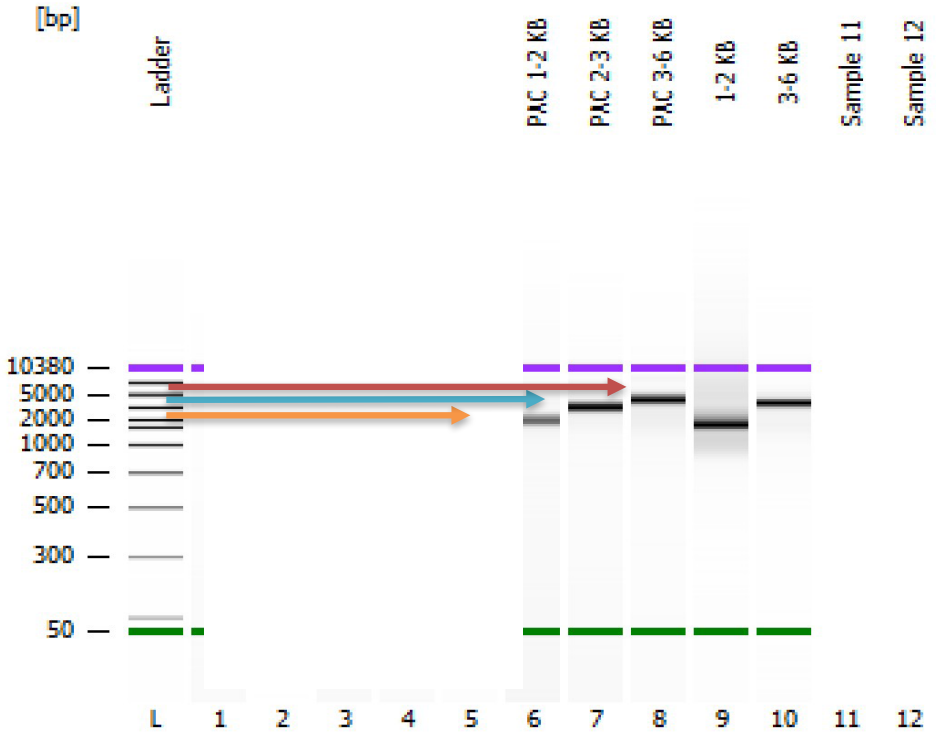
Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1,651	2,091	7.2	15.4
28S	2,836	3,733	16.2	34.7

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Barcoded Library Construction for PacBio Iso-seq Sequencing

BluePippin Size Selection



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Making Barcoded Iso-Seq Libraries

- There were (are) six barcoded adapters available to make pooled libraries

	Primer Sequence	16-mer barcode	oligo dT
dT_BC1	AAGCAGTGGTATCAACGCAGAGTAC	tcagacgatgcgtcat	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
dT_BC2	AAGCAGTGGTATCAACGCAGAGTAC	ctatacatgactctgc	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
dT_BC3	AAGCAGTGGTATCAACGCAGAGTAC	tactagagtagcactc	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
dT_BC4	AAGCAGTGGTATCAACGCAGAGTAC	tgtgtatcagtacatg	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
dT_BC5	AAGCAGTGGTATCAACGCAGAGTAC	gatctctactatatgc	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
dT_BC1	AAGCAGTGGTATCAACGCAGAGTAC	acagtctatactgctg	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

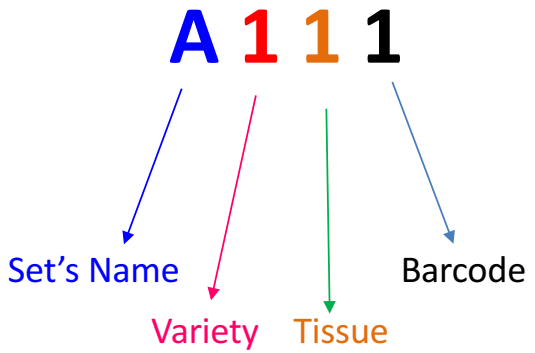
- Nine tissues for each variety (4X and 6X)
  - Only 4X and 6X libraries were barcoded (RSII Sequencing)
  - 2X library was not barcoded (2 Sequel SMRT cells to date)

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

We defined sets (A, B and C)

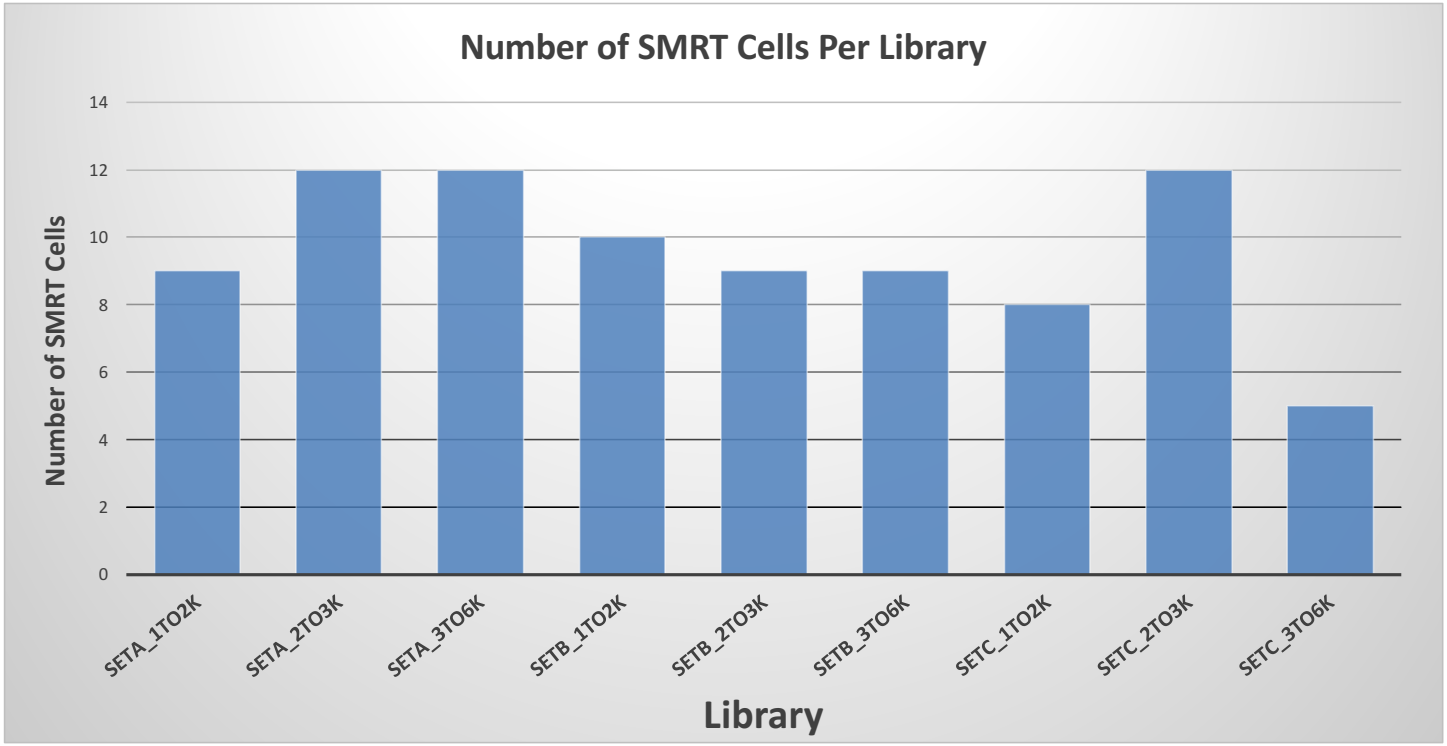
- Our approach
  - Make groups of six and having an indexing convention
    - A111, A122, A133, A144, A155, A166 (1-2, 2-3, 3-6 kb)
    - B171, B182, B193, B214, B225, B236 (1-2, 2-3, 3-6 kb)
    - C241, C252, C263, C274, C285, C296 (1-2, 2-3, 3-6 kb
  - SMRT analysis 2.3.0 to analyze the data



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

On Average We Ran 9.5 SMRT Cells Per Library (87 RSII total)

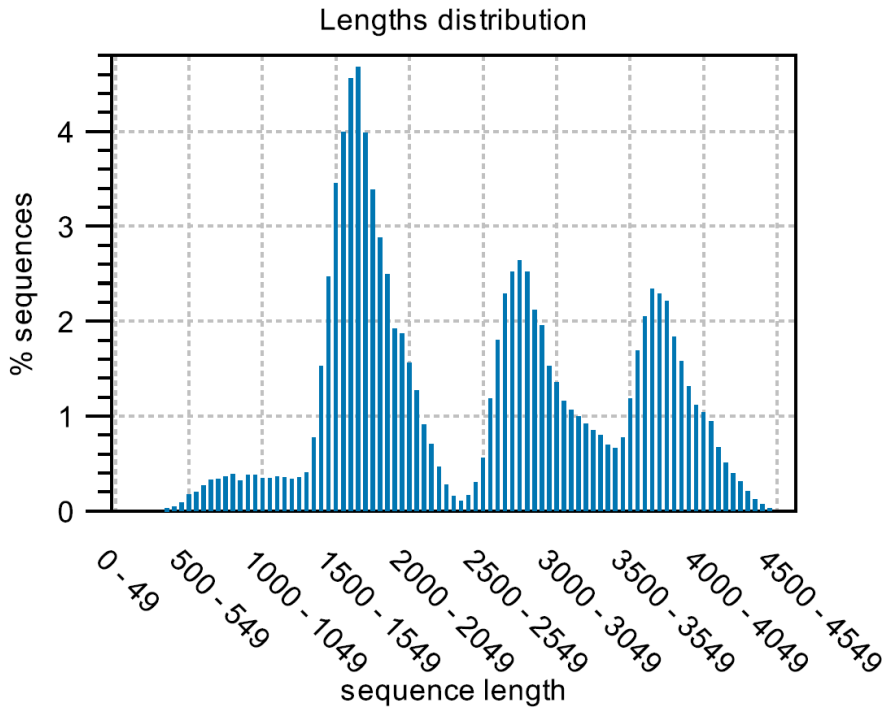


NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Quiver Was Used to Obtain HQ and LQ Data

- Full Length Non-Chimeric (FLNC) = 1,624,690
- Tetraploid Genome LQ = 773,571
- Tetraploid Genome HQ = 141,399 (351,343,616 nt)

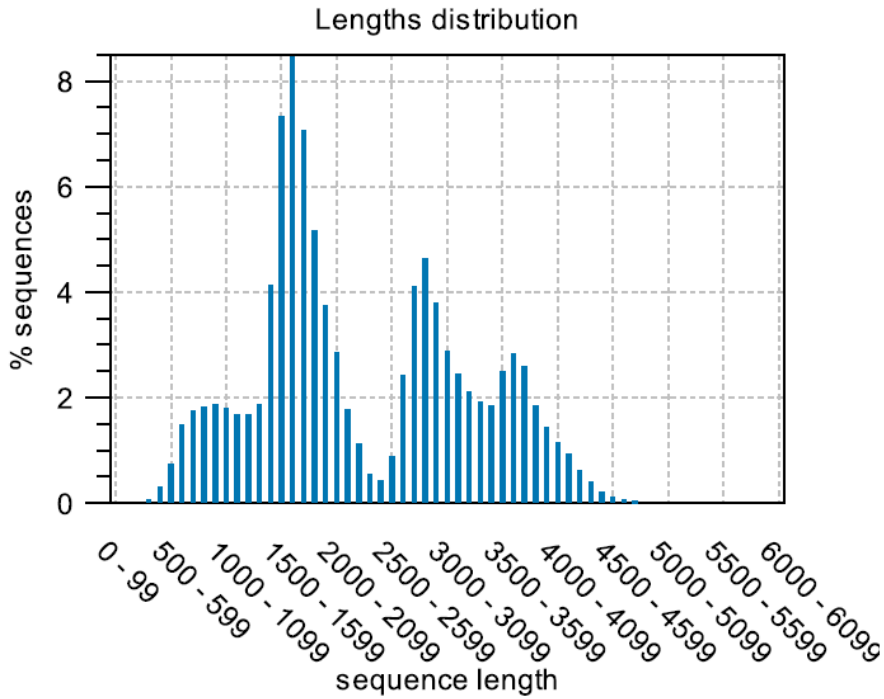


NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Quiver Was Used to Obtain HQ and LQ Data

- Full Length Non-Chimeric (FLNC) = 1,302,432
- Hexaploid Genome LQ = 614,378
- Hexaploid genome HQ = 110,050 and (250,964,213 nt)

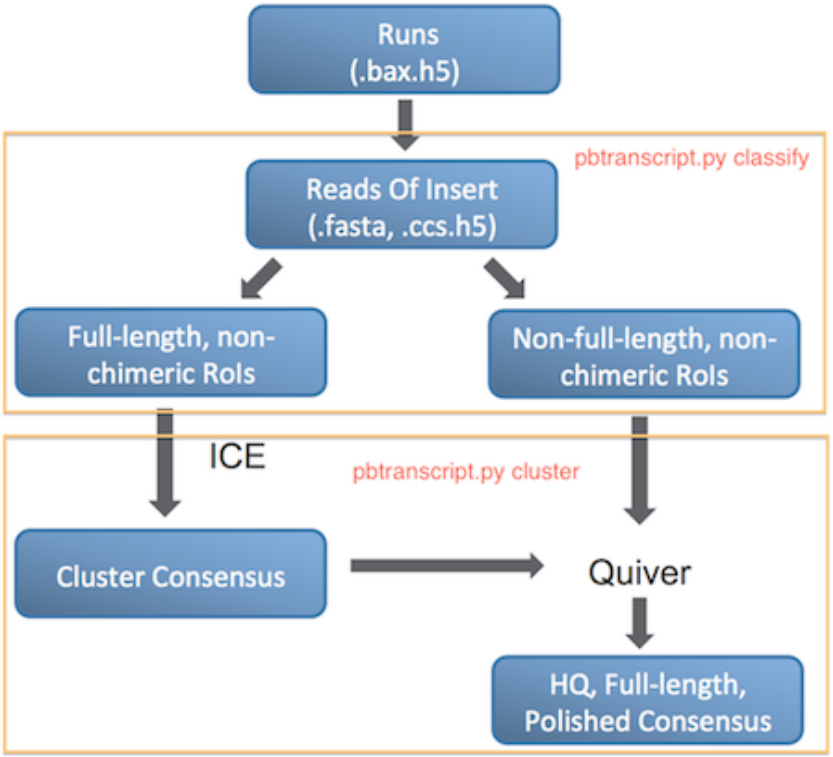


NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Data analysis pipeline (RSII Data)

All SMRT cells of each cultivar were used to output FLNC, LQ and HQ reads

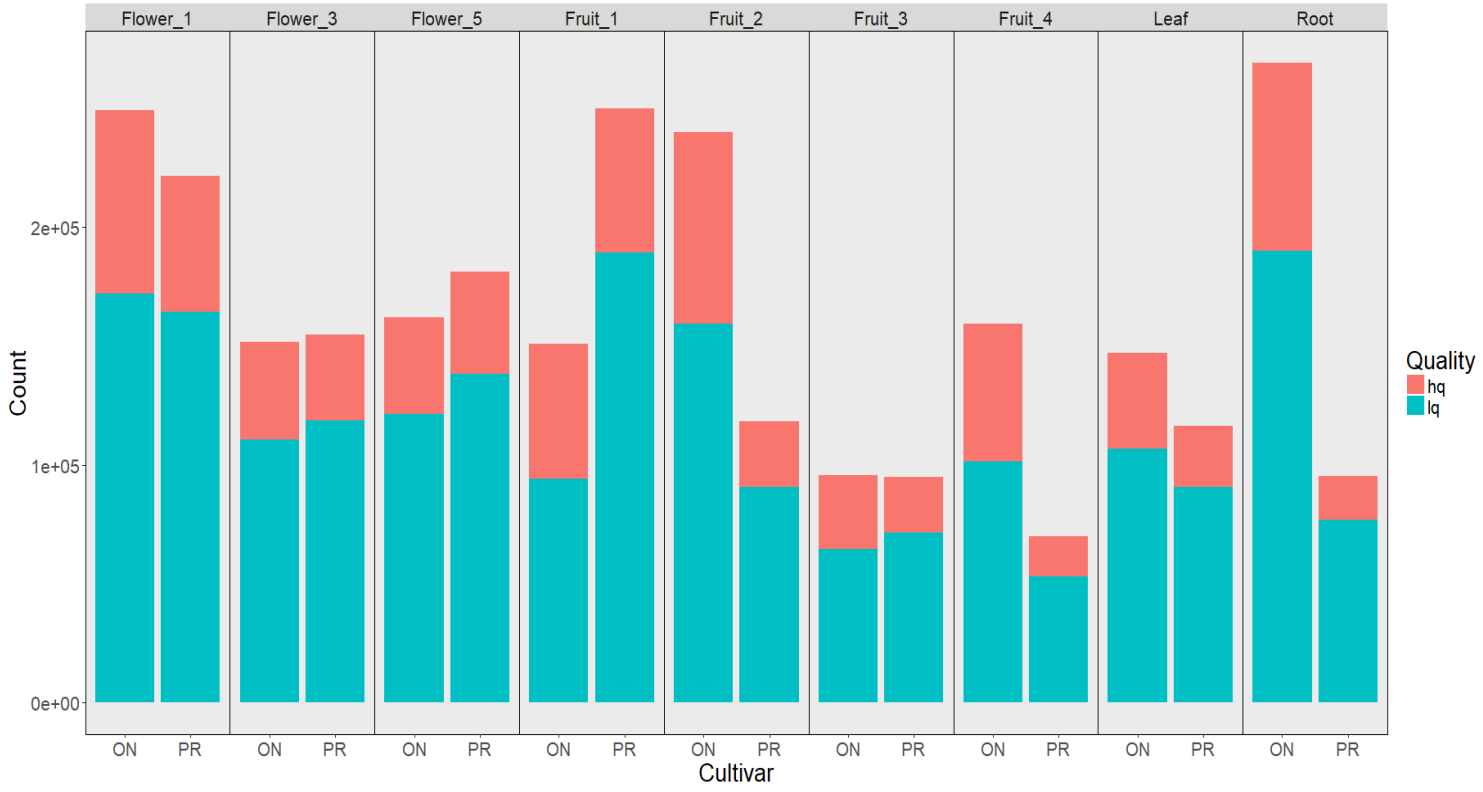


https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-%28v2.3%29-Tutorial-%231.-Getting-full-length-reads

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

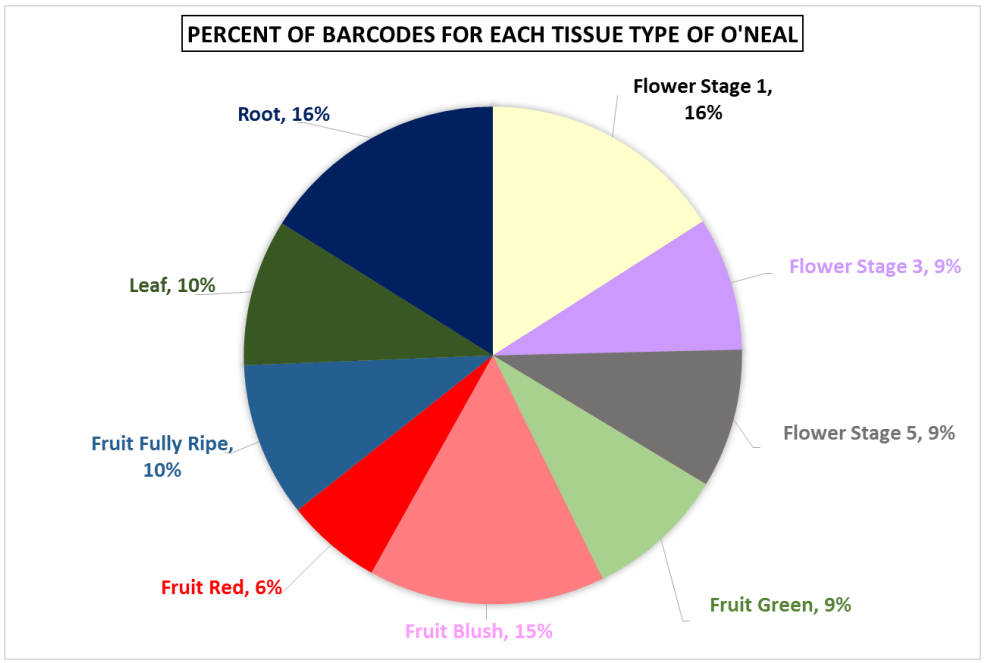
Proportion of LQ and HQ Reads in 4X and 6X Blueberry Genotypes Separated by Different Tissues

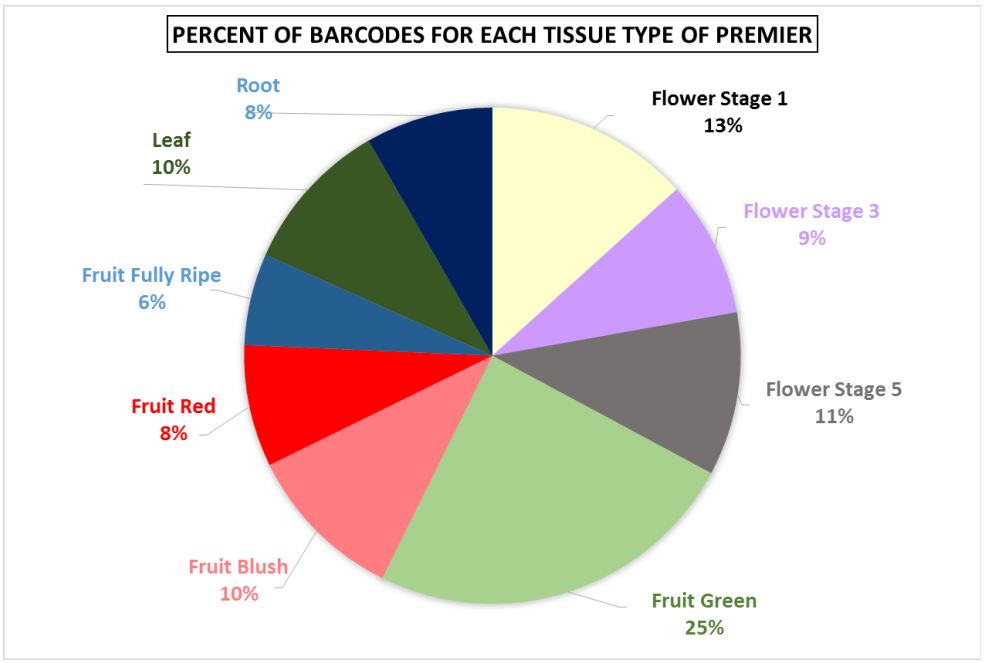


NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Demultiplexing Full Length Non-chimeric Reads (FLNC)

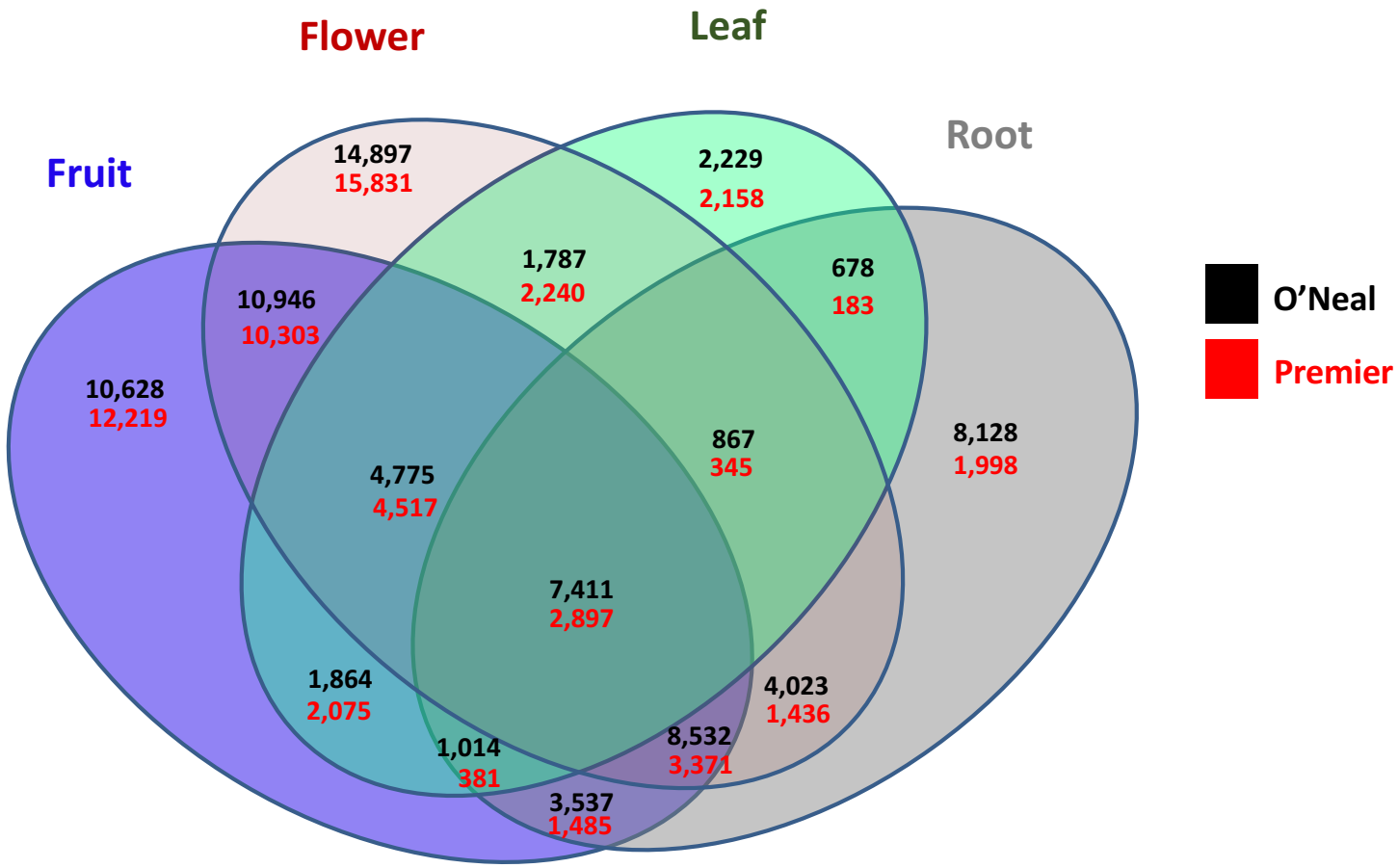




NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

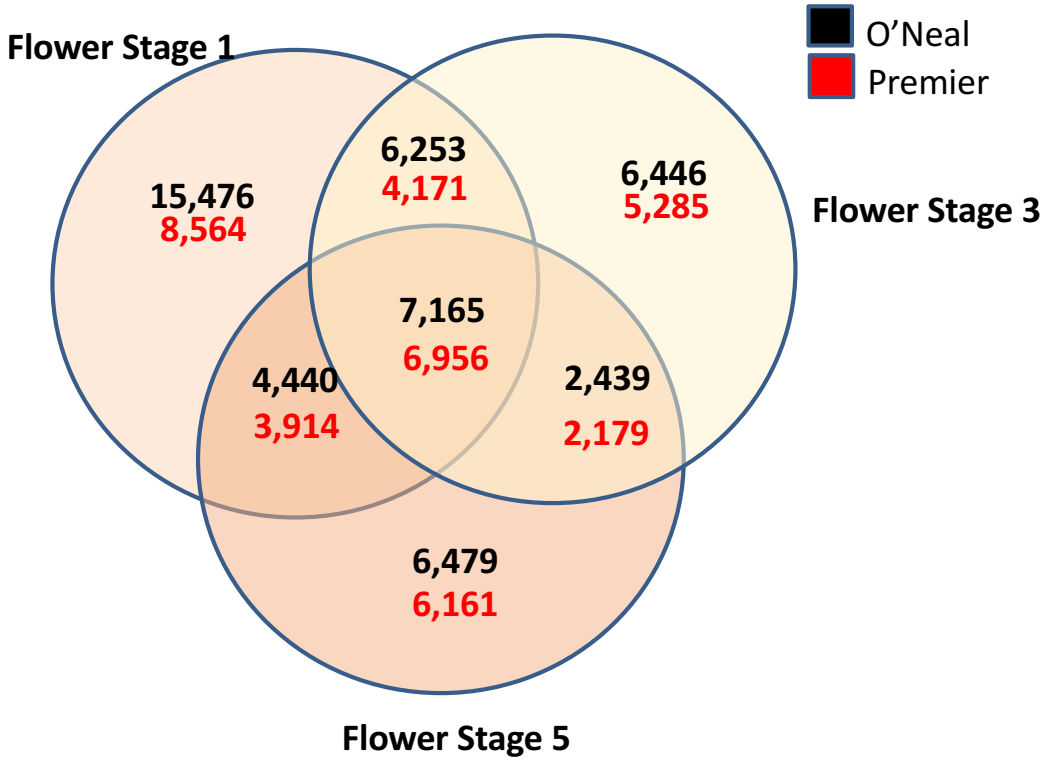
Overlap of all PacBio Isoforms in Four Tissues of 4X and 6X Genotypes



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

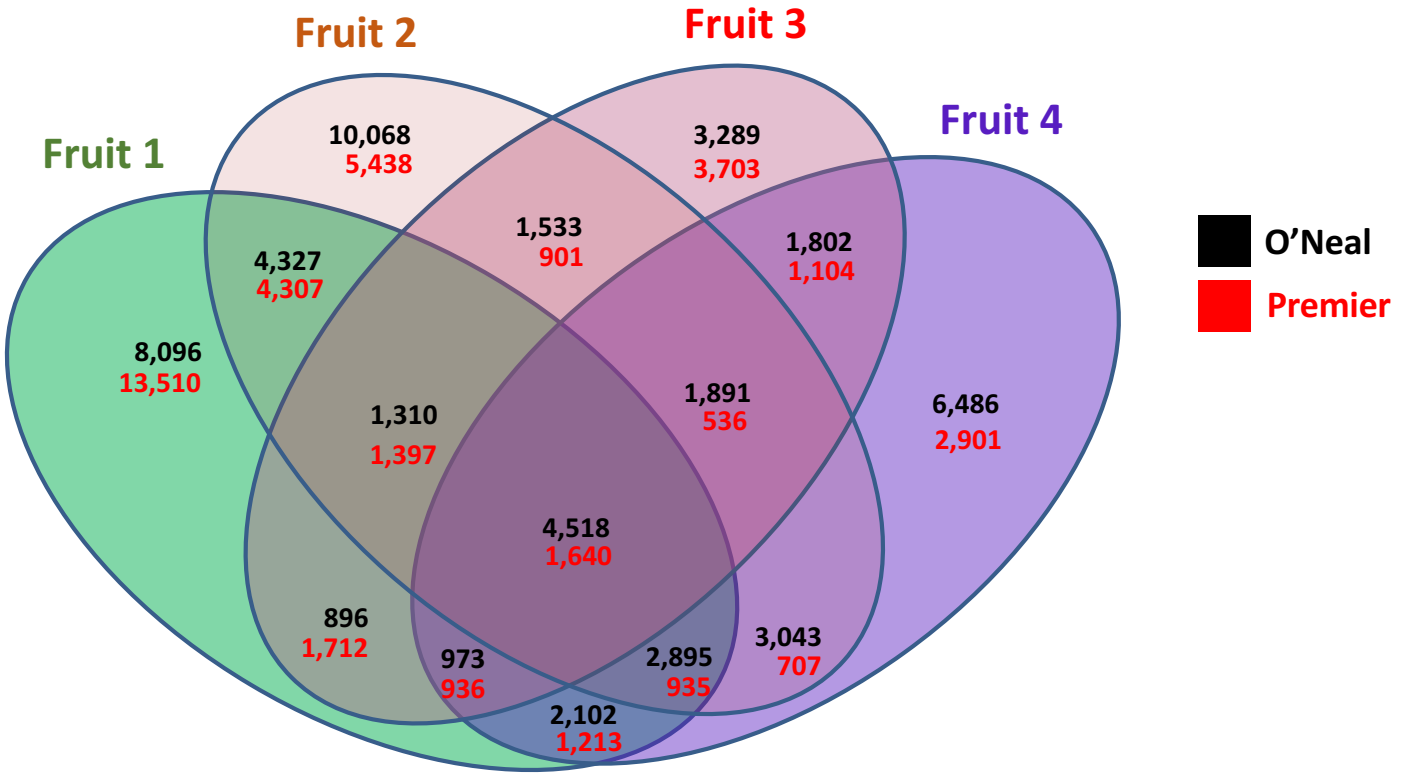
Overlap of all PacBio Isoforms in Three Flower Development Stages of 4X and 6X Genotypes



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Number of Iso-Seq Sequences Separated by Tissue Type in 4X and 6X Genomes



GMAP Was Used to Map HQ Reads to the Reference Sequence

- Two versions of reference sequence was used
  - Quivered genome
  - Pilon corrected Quivered genome

  (https://github.com/broadinstitute/pilon/wiki)
- Mapping was done to primary contigs
- It ran once with default parameters only for 4X and 6X iso-seq data
- It ran for the second time with the following parameters for 2x, 4X and 6X
  - *--min-trimmed-coverage = 0.95 & --min-identity=0.95*

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

GMAP Default Values					
No. of Iso-Seqs Mapped to Quivered Genome			**No. of Iso-Seqs Mapped to Quivered – Piloned Genome**		
2X	**4X**	**6X**	2X	**4X**	**6X**
-	140,664 (99.48%)	109,672 (99.63%)	-	140,096 (99.07%) 568	109,099 (99.13%) 573
No. of Iso-Seqs Uniquely Mapped to Quivered Genome			**No. of Iso-Seqs Mapped to Quivered – Piloned Genome**		
-	100,441	71,146	-	103,106 2,665	77,613 6,467

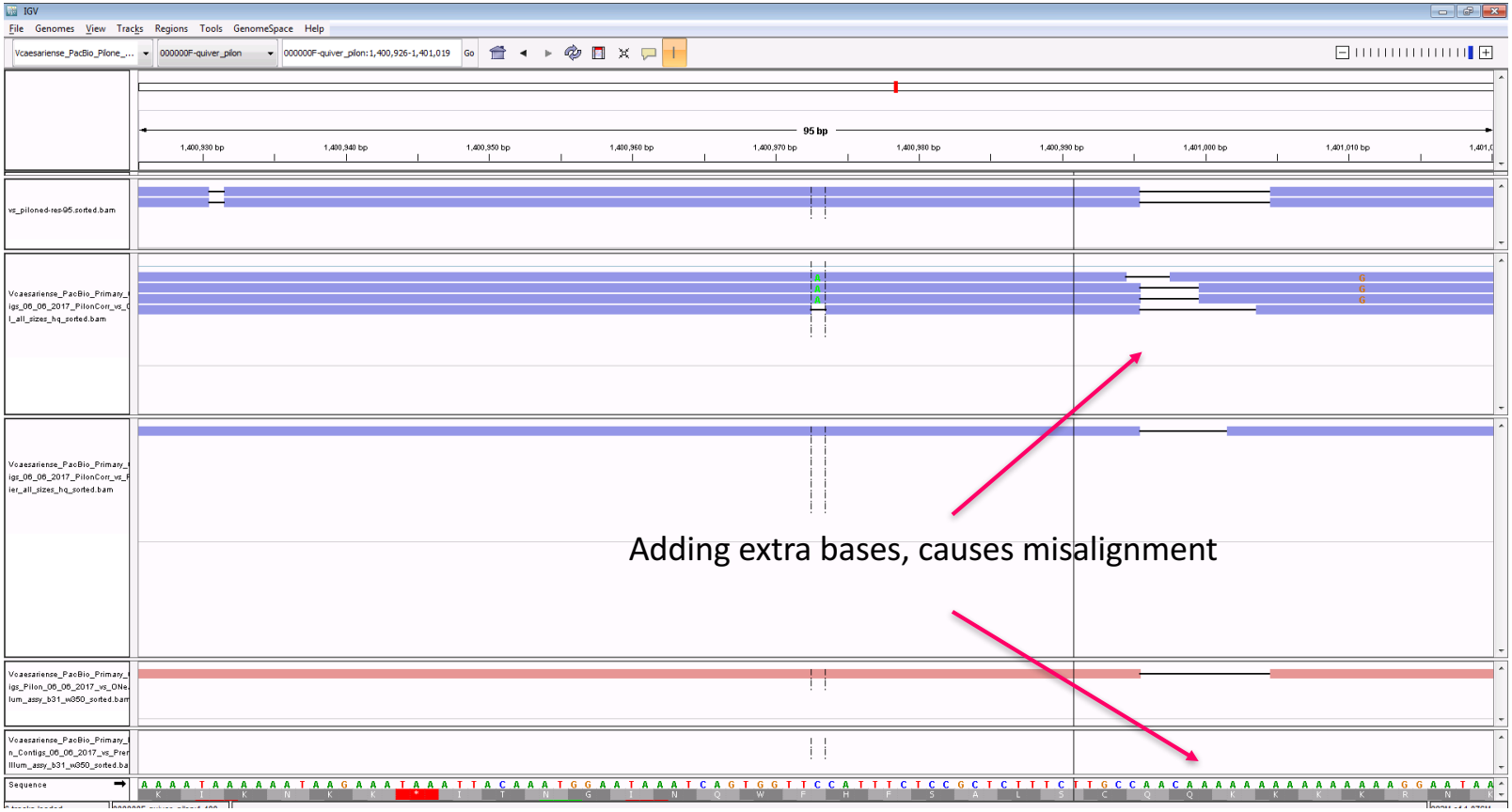
The increase in the number of uniquely mapped sequences to the Pilon corrected genome, may indicate that it is better to make this correction.

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

GMAP Run With min cov 0.95 and min identity 0.95					
No. of Iso-Seqs Mapped to Quivered Genome			No. of Iso-Seqs Mapped to Quivered – Piloned Genome		
2X	4X	6X	2X	4X	6X
-	127,312 (90.00%)	96,681 (87.85%)	-	127,307 (90.00%) ↓5	96,709 (87.87%) ↑28
No. of Iso-Seqs Uniquely Mapped to Quivered Genome			No. of Iso-Seqs Mapped to Quivered – Piloned Genome		
29,512	100,428	74,097	29,509 ↓3	100,329 ↓99	74,074 ↓23

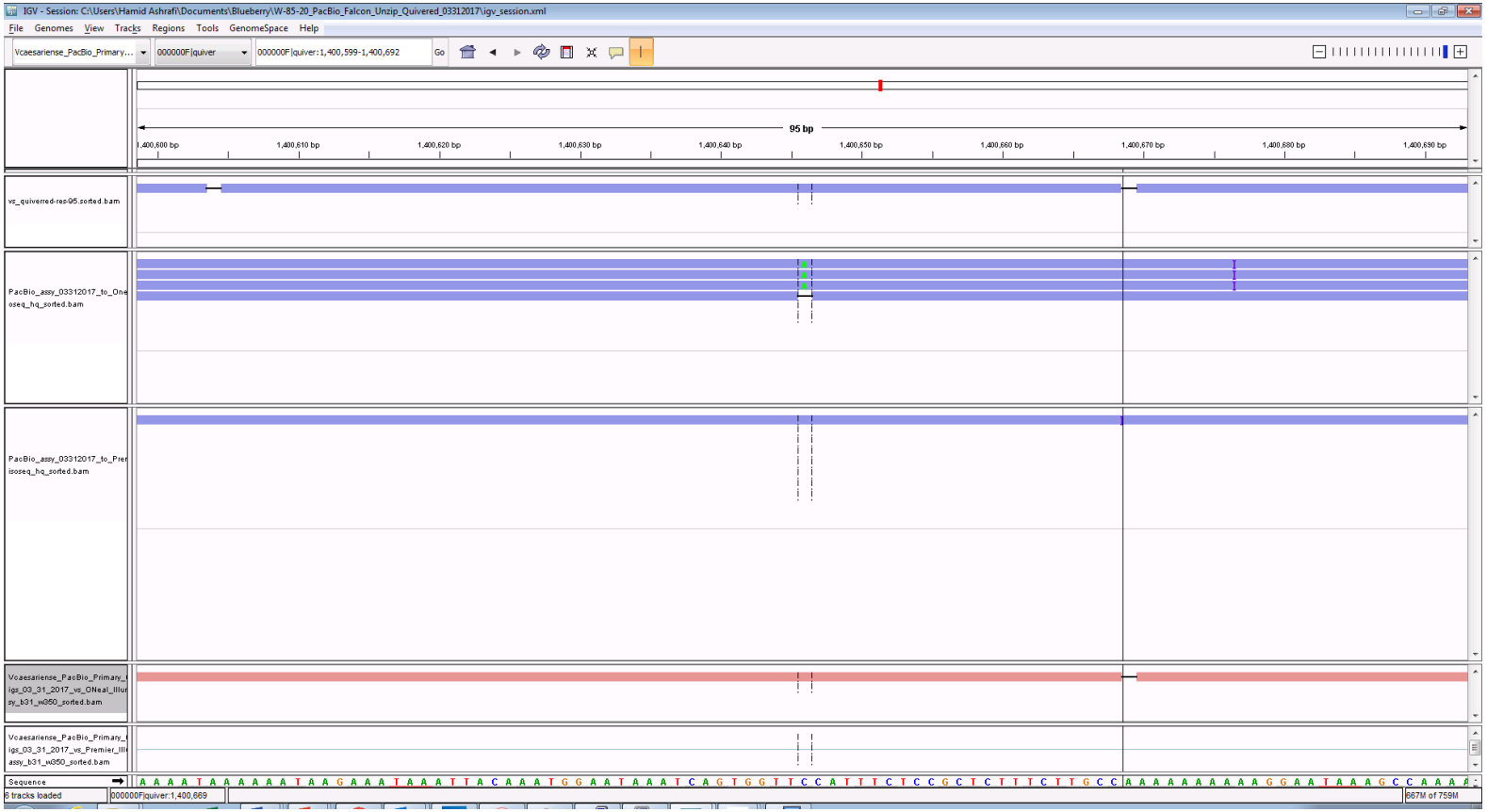
Should We Correct The Genome with Pilon?



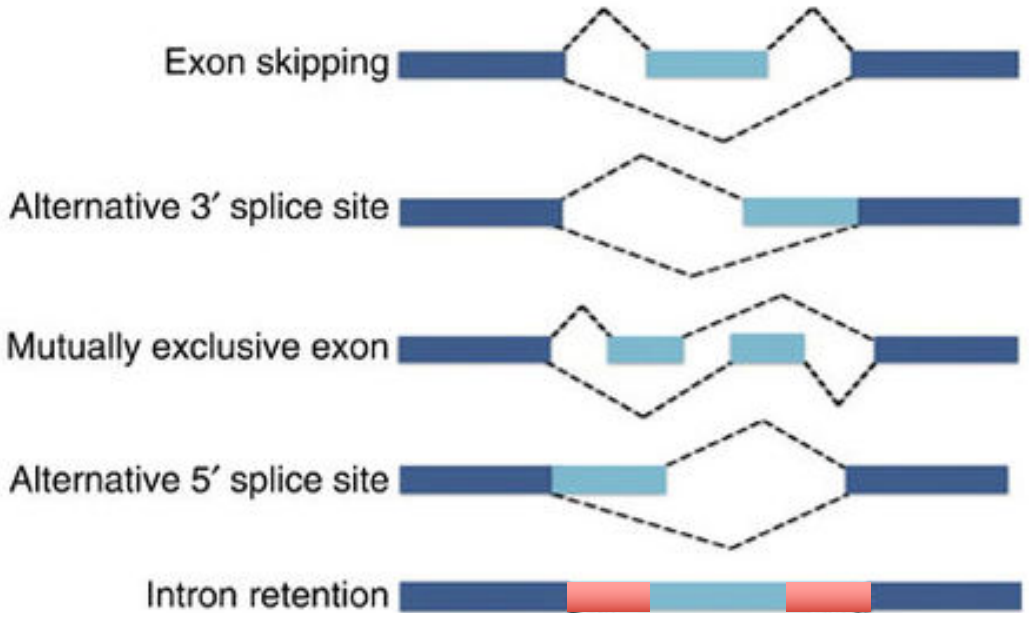
NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Aligning Iso-Seq to Only Quivered Genome



Five Alternative Splicing Modes



Wang et al. (2016) https://www.nature.com/articles/ncomms11708

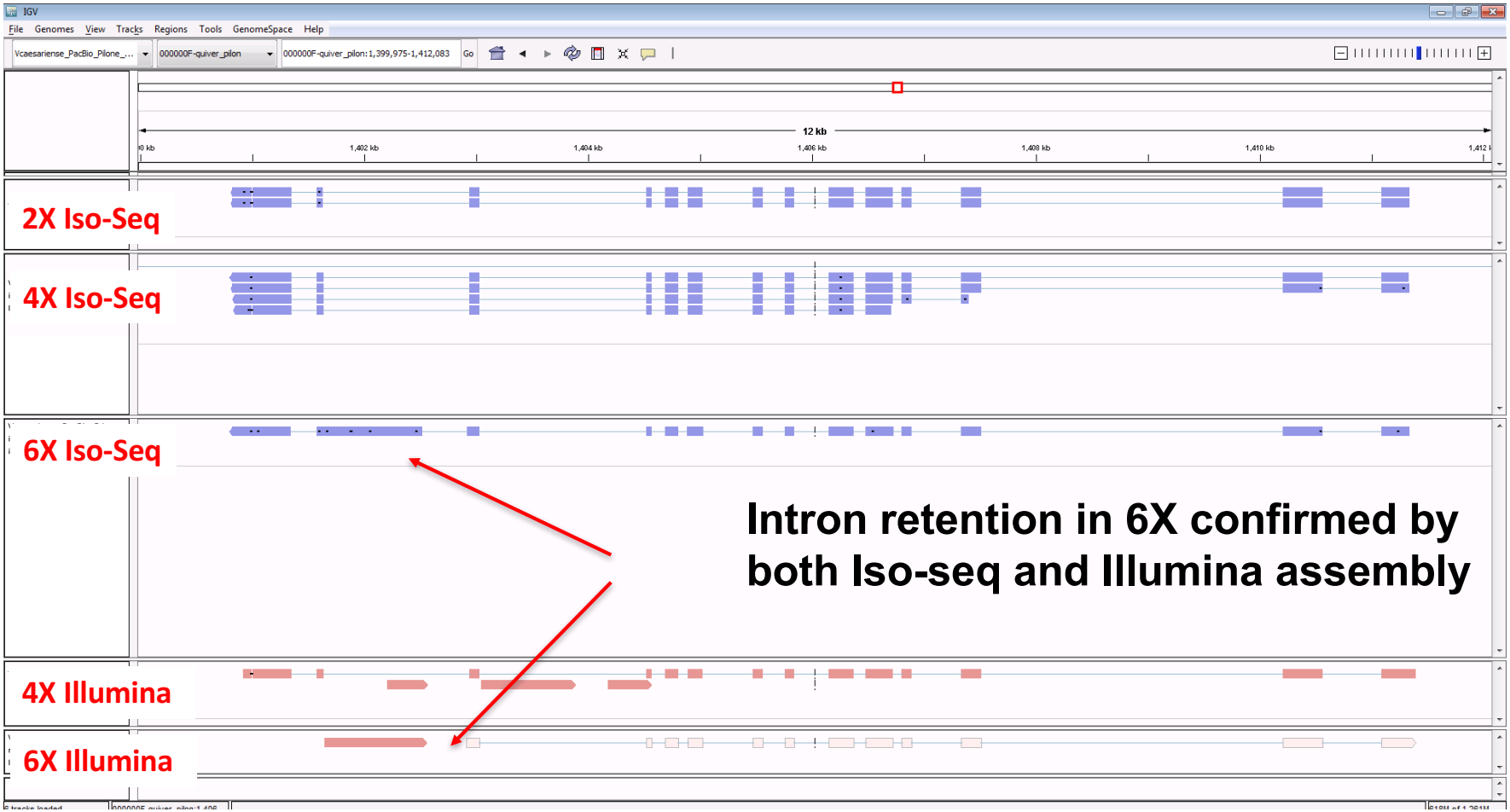
NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Polymorphisms in hexaploids not in 2X
Introns in Illumina assembly

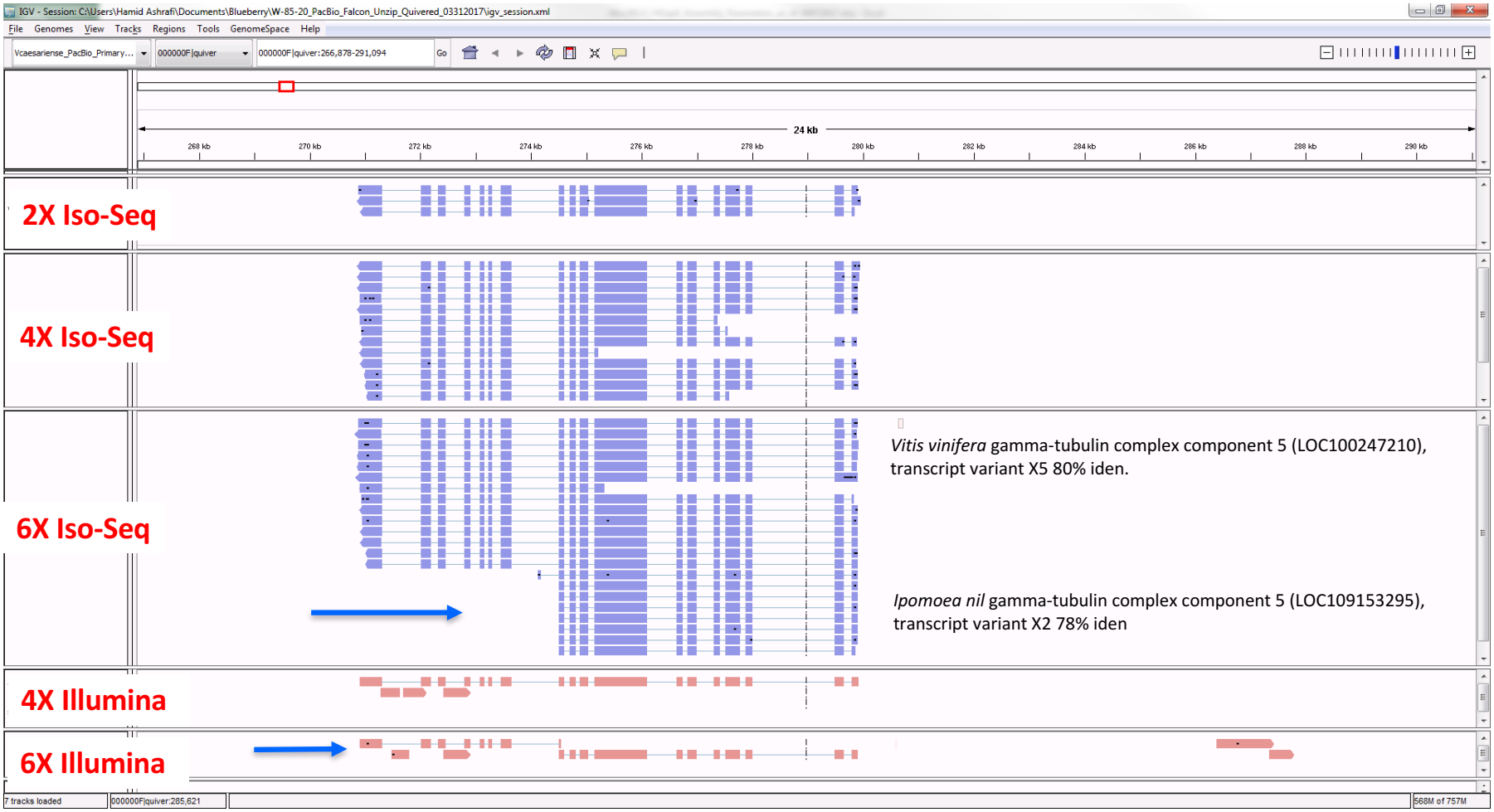
NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®



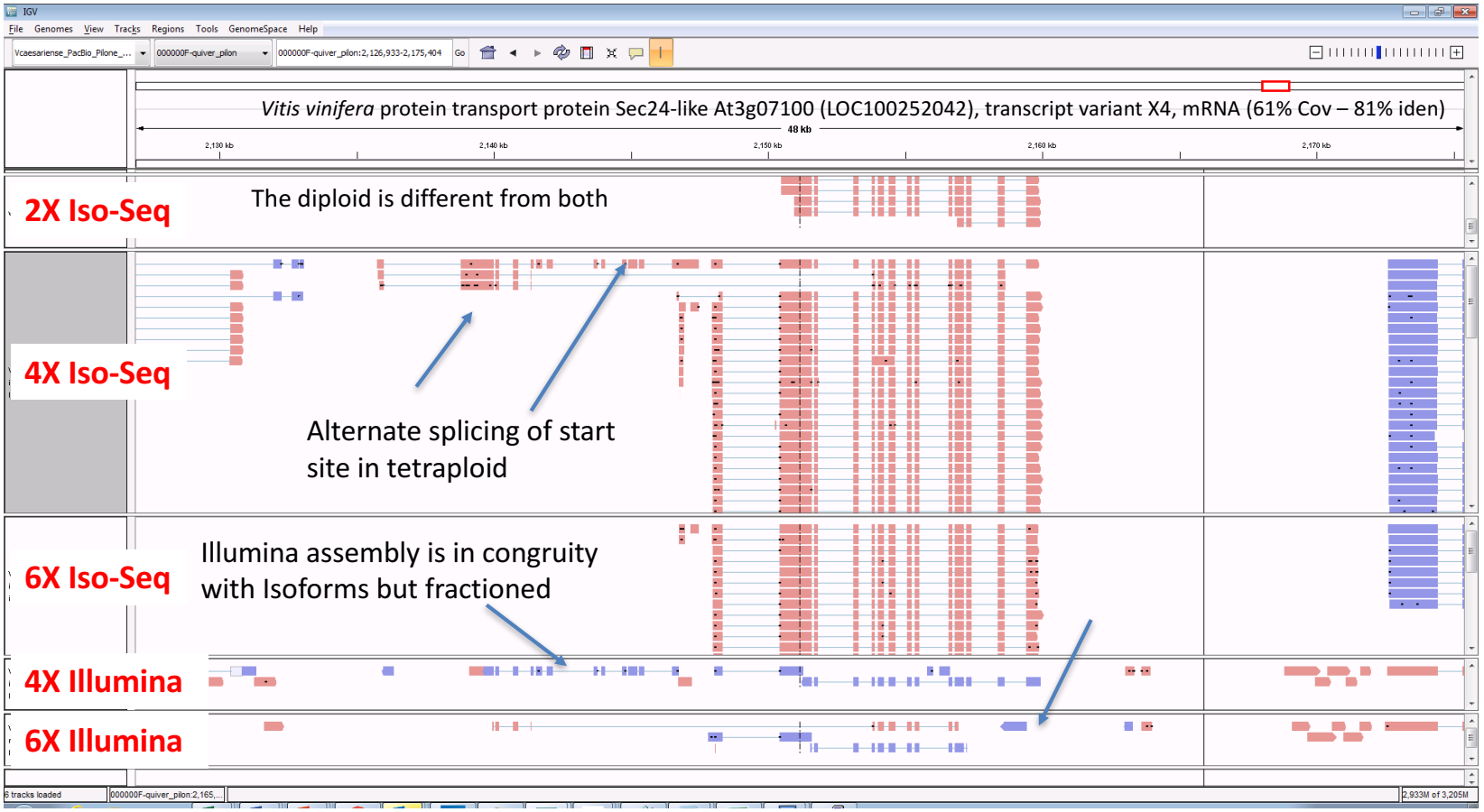
Vitis vinifera gamma-tubulin complex component 5 (LOC100247210), transcript variant X5 80% iden.

Ipomoea nil gamma-tubulin complex component 5 (LOC109153295), transcript variant X2 78% iden

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

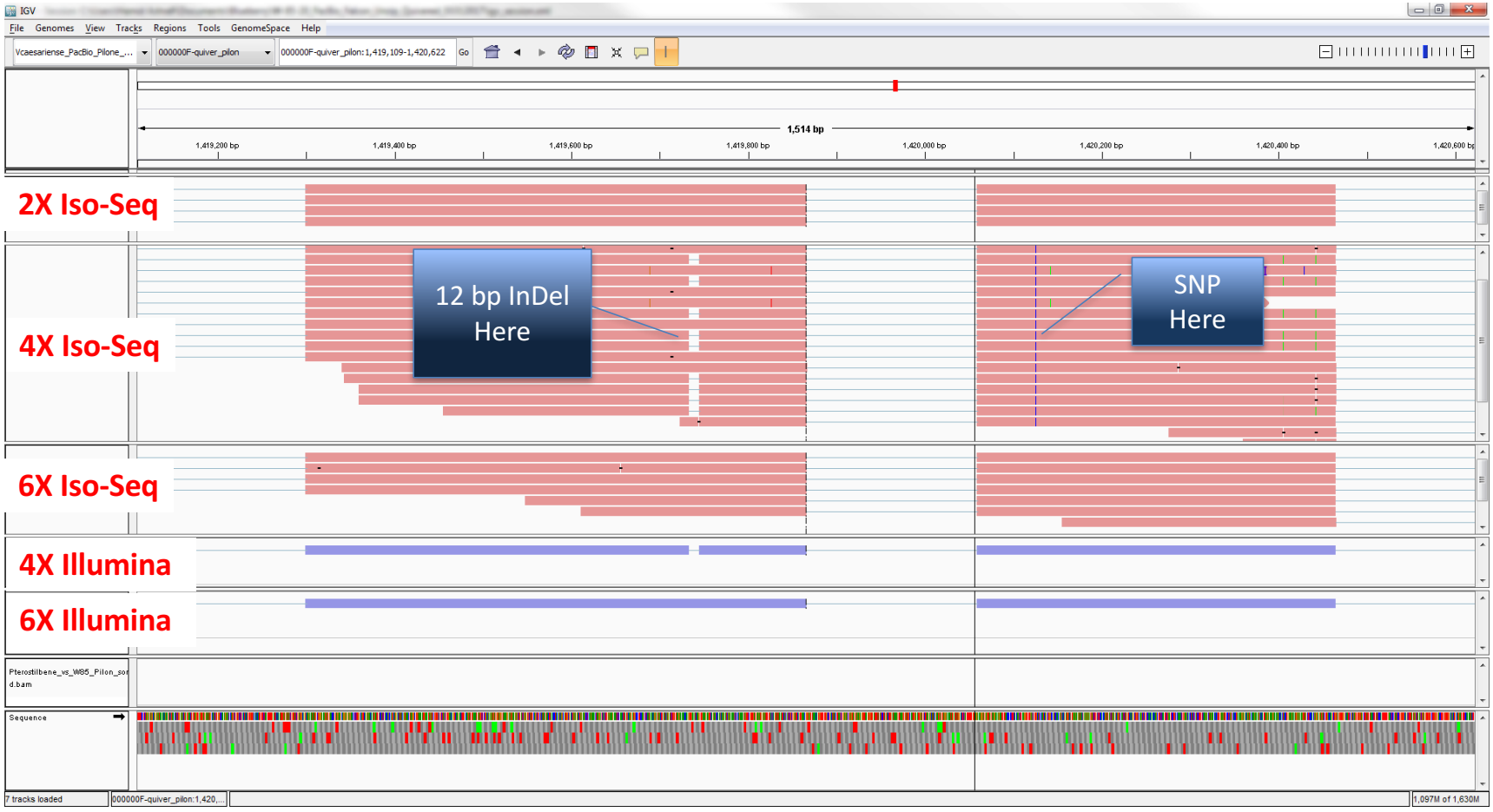
Alternate Splicing of Start Site of a Transport Protein in 4X



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

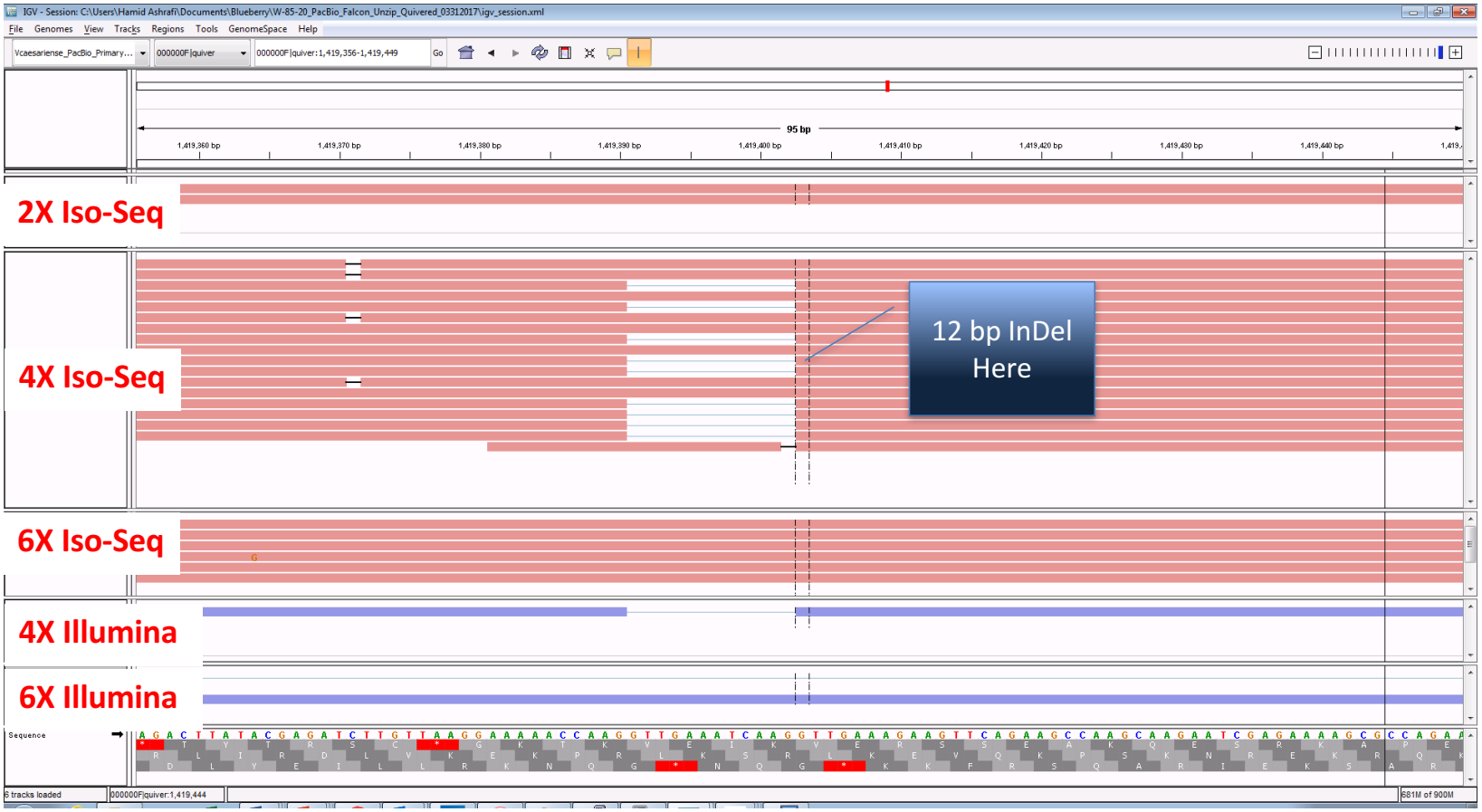
Isoform Analysis Unrevealed Other Forms of Variation (1/3)



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

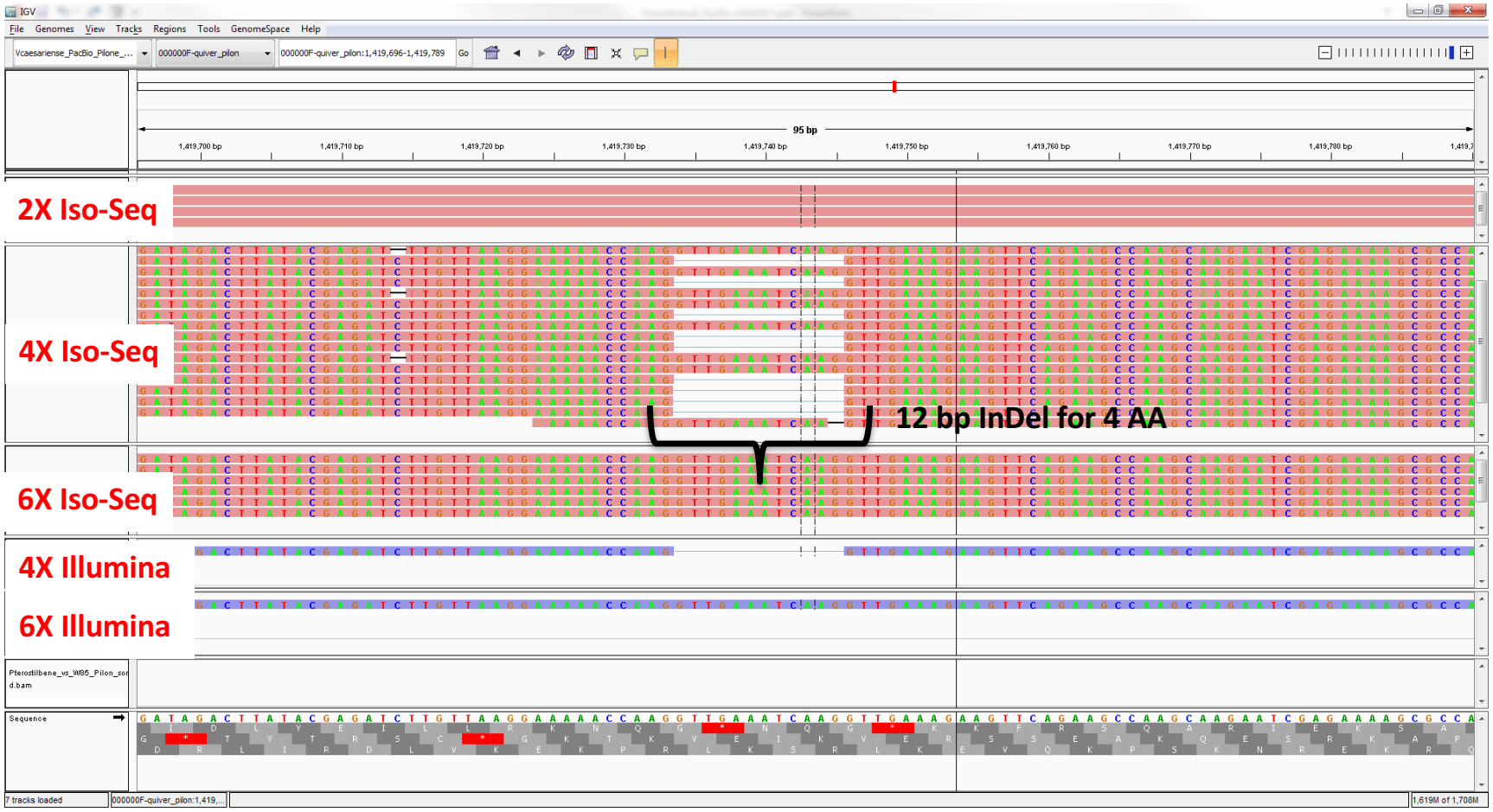
Isoform Analysis Unrevealed Other Forms of Variation (2/3)



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Isoform Analysis Unrevealed Other Forms of Variation (3/3)

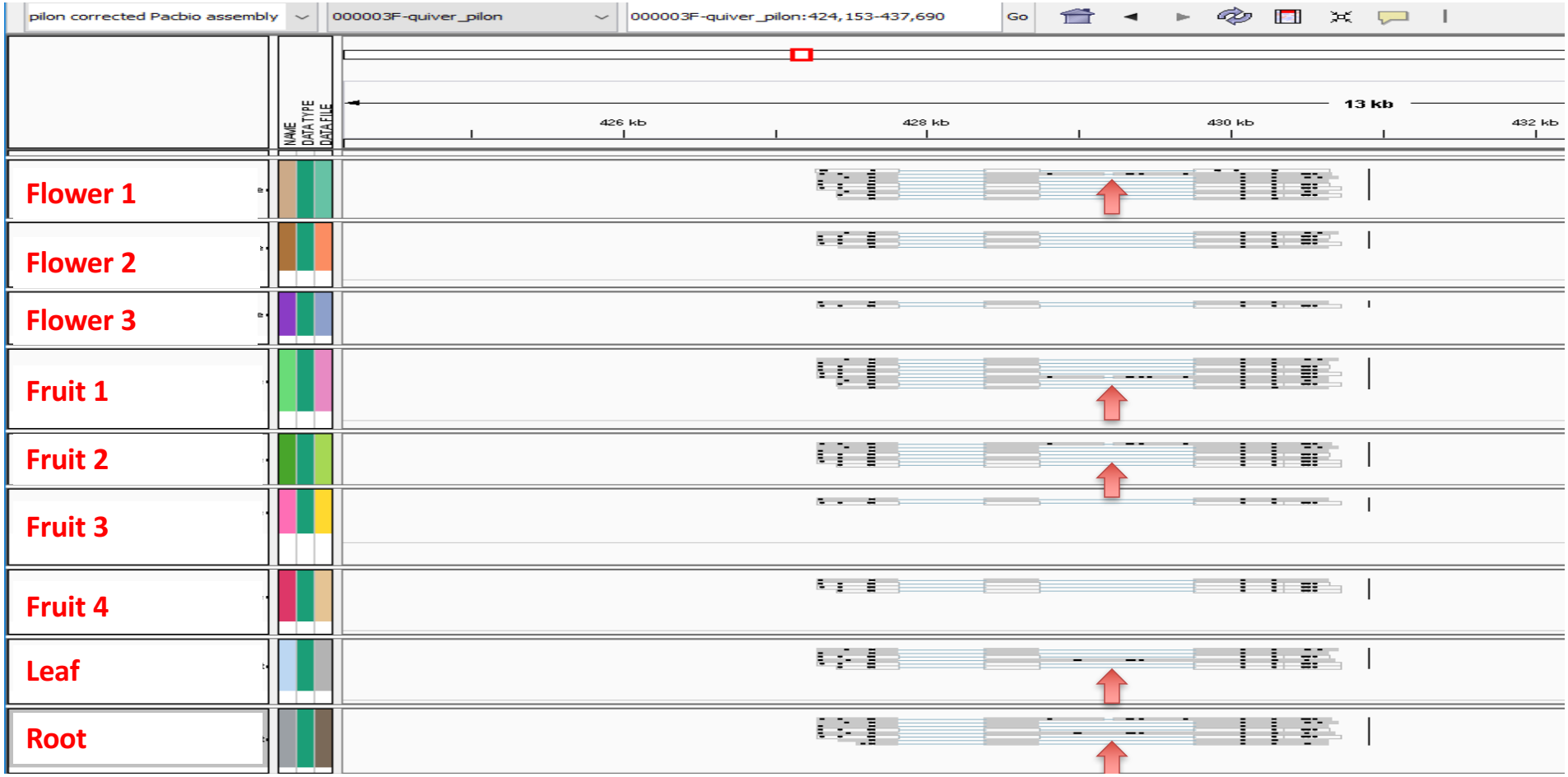


Tissue Specific Iso-Forms

NC STATE UNIVERSITY

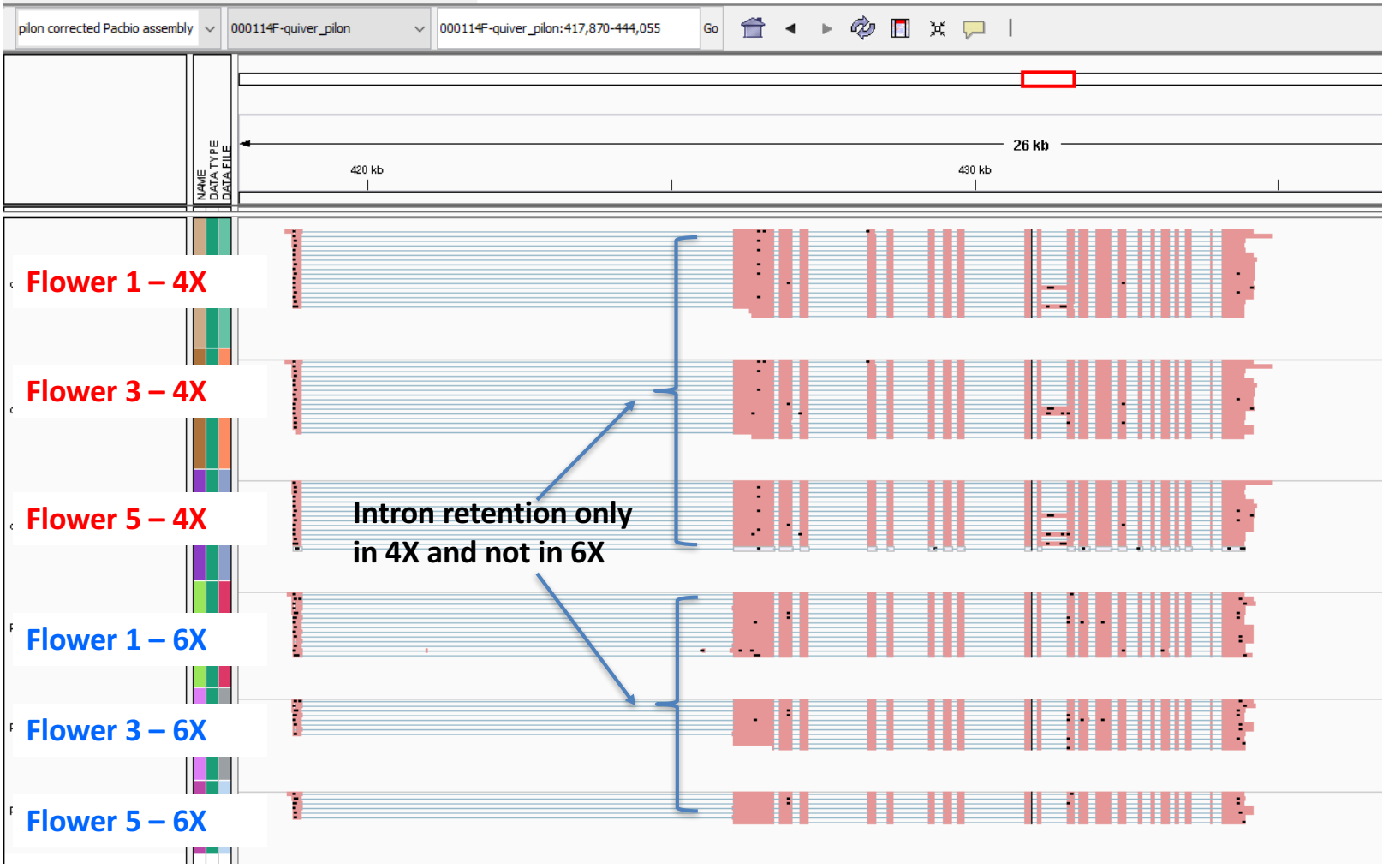
PACIFIC BIOSCIENCES®

Tissue Specific Intron Retention



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®



NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Conclusions

- Using 54 SMRT RSII cells we were able to achieve a decent assembly of blueberry genome.
- Barcoding the RNA-Seq libraries helped us to save for the cost of the project.
- What makes different blueberries is more than simple SNP in the genes.
- We believe alternative splicing, Iso-forms and structural variants are responsible for a large proportion of variations and evolution of blueberry.
- We need to use the term “autoployploidy” with caution for blueberry and maybe other plants species.
- Although the chromosomes can pair during meiosis, this does not mean that sub-genomes are identical at the gene and iso-forms levels.
- Defining the iso-forms to have a biological meaning and connecting them to the actual function of the genes remains a challenging task.

NC STATE UNIVERSITY

PACIFIC BIOSCIENCES®

Thank You
Questions?