Assembly and Analysis of DNA Sequencing Data

Xiaoqiu Huang
Department of Computer Science
Iowa State University

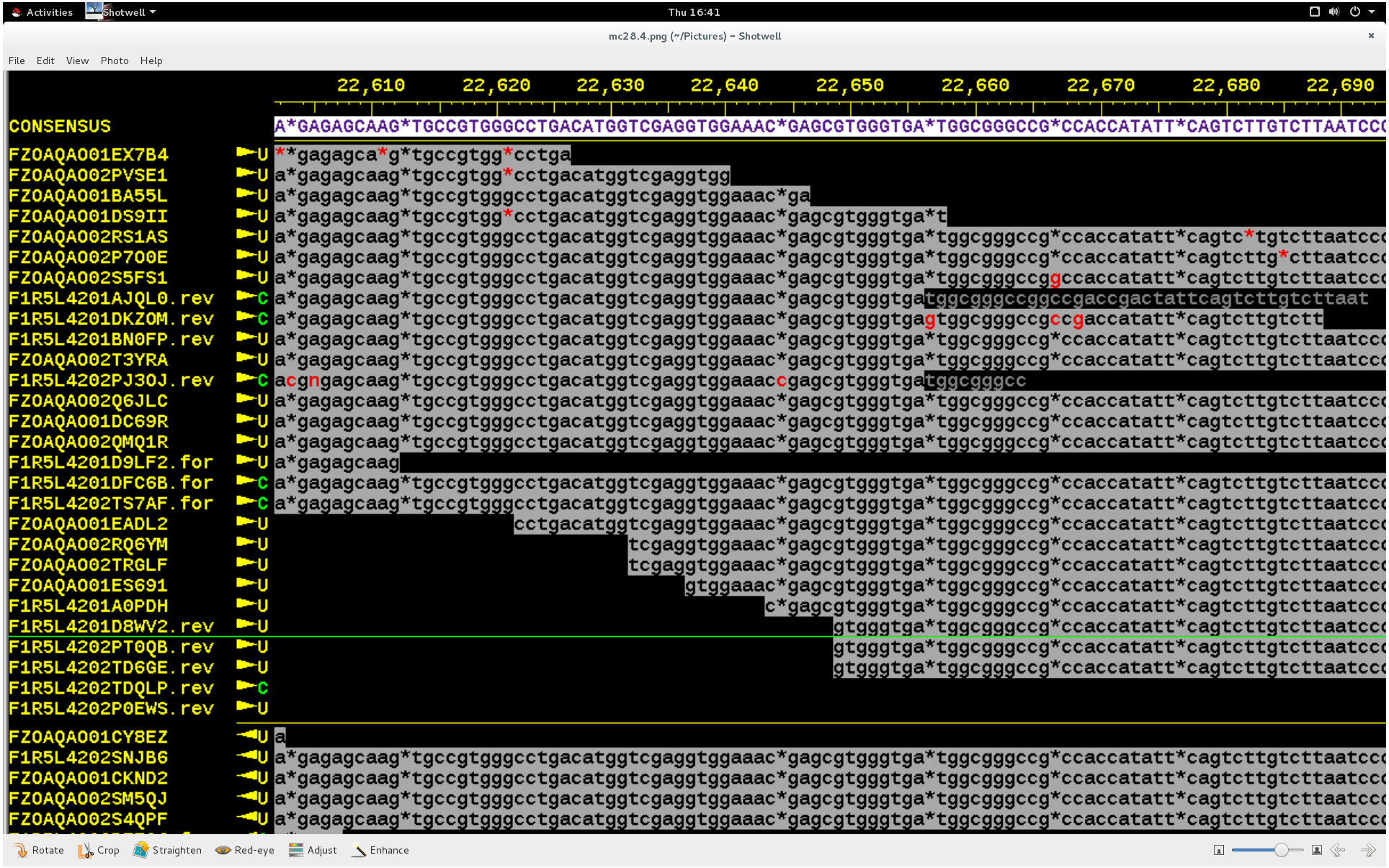
Raw DNA Sequencing Data are Huge and Short

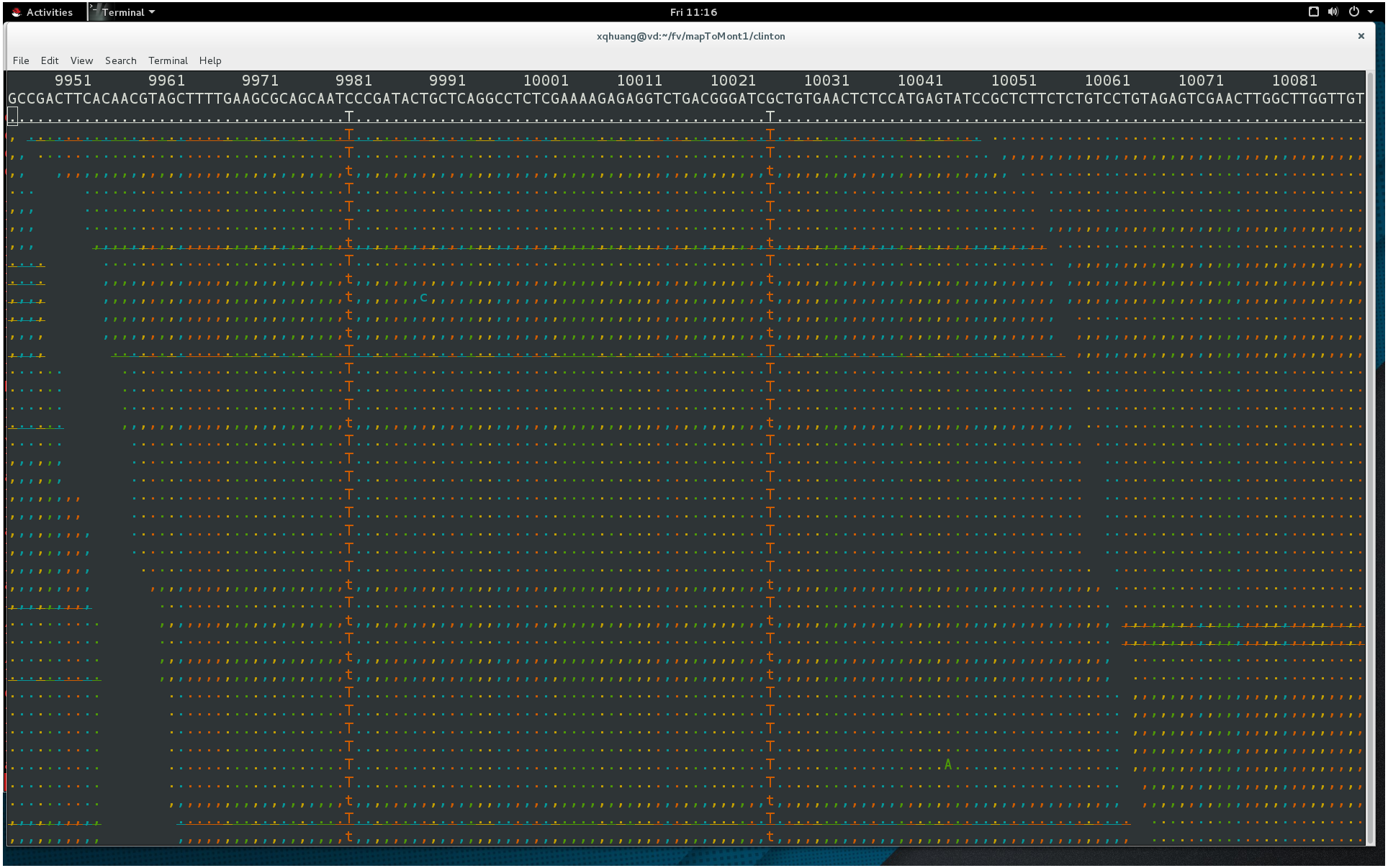
- A huge number (up to a billion) of DNA sequences can be produced by a sequencing instrument in a day.
- DNA sequences are short (e.g. 100 to 200 base pairs in length).
- Their error rates are often below 1%.


Applications Places System
Tue Jul 22, 4:50 PM Xiaoqiu Huang
xqhuang@vd:~/fv/34551
File Edit View Search Terminal Help
@HWI-ST412_0141:3:1:15599:2216#0/1
CTCGCAGAAGTGCGGCCCATACCAACTCAACAGGGTCAACTCGGCCCACAGCACTGGCAGCCTGGGCACCGACAGCATGCTCAAGAGCGCTCGGGAACCTCC
+HWI-ST412_0141:3:1:15599:2216#0/1
gccccccccgdgccccccccccccdcceeceeggdcagcdgefggggdfgggggfgegdfd^ef_f\fddaddbbce_cffcfedecdfedcdbdWdddc
@HWI-ST412_0141:3:1:15653:2238#0/1
AGGGGCGGCAAACATGGAAGATGACTTCATGAGCTCTGTCAAGGCCAGCGAAGATGTCGAGTTGCGGGACTTGTGGTAGTTGGAGAGGCCACTCGATGGGGG
+HWI-ST412_0141:3:1:15653:2238#0/1
gccccc_YUccccagccdcccecgcececddgggdgccdgdgggggggdefdffdegcgggfWfffdddK_BBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST412_0141:3:1:15678:2237#0/1
GTGAGAAGGTGAATGATATGAGGGCAATGAGATACTTGATGGGCTCCCACAGCTGCTCCGCGATGCAGCGGTTCCATATTCCCACCATGAGGTTTGCAAAGT
+HWI-ST412_0141:3:1:15678:2237#0/1
fd^a^a^^^fceccdcccceccccccccgccgccdggeggggeggggggggggggggfeggcfggggfgfefdfdccbdgggggddeeeaecbff][_]_N
@HWI-ST412_0141:3:1:15724:2239#0/1
TCAGCCCCGGAGGTGGTTCCAGCTGCTCCGGTCAGCAGACCCCTGATTGGTTATGGGGCCAGGGTGACGCGGTGCCGAATGGGAGACCTCTAATAAGCCAGC
+HWI-ST412_0141:3:1:15724:2239#0/1
gccccccccgccgccecccccceecdccccggeggcecegdcedggccegggggggggegggggggeggggg`[^\eefedfZdddd_eeeeefeffcc
@HWI-ST412_0141:3:1:15772:2167#0/1
TCATCCGCGCAAGCTTGAATCCTGGAGTGTCCATTGTCTGTTTCGGGCAACACNGATTGCNCANNAGAACNNGTNNNNTCNNNGCTNNNNNNNNNNNNNNN
+HWI-ST412_0141:3:1:15772:2167#0/1
gccccccccccdccccc_ccccdcdgcggggeggdggdcdgggfgggfggcc^BcBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST412_0141:3:1:15964:2199#0/1
AATGCGACGAGCAGACTCAGCTGCTGGATGATGCTCGACCGCGACAAGTCAGAACTCAAAGGGTGAGTGCCATAAAAGGTTGGACGGAGAAGCTCCTGACAG
+HWI-ST412_0141:3:1:15964:2199#0/1
gcdccccccgdddcecddeccecddccgggggggeggggeggeggffe_gegfgfggaggfffdfffdf_eed^ecddbdfeefddbddedaaddeW_d`X_
@HWI-ST412_0141:3:1:15753:2227#0/1
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAGACGAACCACGGAGCTCAGACGTGCGGA
+HWI-ST412_0141:3:1:15753:2227#0/1
gccccccccgcdcccccgcccgcddccdcggedgeeggbgeNdcdcffffgggggfcdaddddVc_aBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST412_0141:3:1:16022:2151#0/1
TCGGAGCGAGAGGACACGCTTGTCTANCTGTNNTTGNNTANNACNNNNNGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+HWI-ST412_0141:3:1:16022:2151#0/1
gceggcccgcgecgcdcedgddg`gaBaabBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST412_0141:3:1:16212:2152#0/1
:
xqhuang@vd:~/fv/...
xqhuang@vd:~/fv/...
xqhuang@lhb-ngs:...
xqhuang@map:~/h...
xqhuang@vd:~/fv/...
xqhuang@map:~
xqhuang
huang
talk
2014

Assembly and Mapping

- Short redundant DNA sequences are put together to construct longer non-redundant genomic or RNA sequences (assembly).
- DNA sequences are aligned to a reference sequence (mapping) to find single nucleotide polymorphisms (SNPs).


Activities
Shotwell
Thu 16:41
mc28.4.png (~/Pictures) – Shotwell
File Edit View Photo Help
22,610 22,620 22,630 22,640 22,650 22,660 22,670 22,680 22,690
CONSENSUS
FZOAQA001EX7B4
FZOAQA002PVSE1
FZOAQA001BA55L
FZOAQA001DS9II
FZOAQA002RS1AS
FZOAQA002P7O0E
FZOAQA002S5FS1
F1R5L4201AJQL0.rev
F1R5L4201DKZOM.rev
F1R5L4201BN0FP.rev
FZOAQA002T3YRA
F1R5L4202PJ3OJ.rev
FZOAQA002Q6JLC
FZOAQA001DC69R
FZOAQA002QMQ1R
F1R5L4201D9LF2.for
F1R5L4201DFC6B.for
F1R5L4202TS7AF.for
FZOAQA001EADL2
FZOAQA002RQ6YM
FZOAQA002TRGLF
FZOAQA001ES691
F1R5L4201A0PDH
F1R5L4201D8WV2.rev
F1R5L4202PT0QB.rev
F1R5L4202TD6GE.rev
F1R5L4202TDQLP.rev
F1R5L4202P0EWS.rev
FZOAQA001CY8EZ
F1R5L4202SNJB6
FZOAQA001CKND2
FZOAQA002SM5QJ
FZOAQA002S4QPF
Rotate Crop Straighten Red-eye Adjust Enhance


Activities
Terminal
Fri 11:16
xqhuang@vd:~/fv/mapToMont1/clinton
File Edit View Search Terminal Help
9951 9961 9971 9981 9991 10001 10011 10021 10031 10041 10051 10061 10071 10081
GCCGACTTCACAACGTAGCTTTTGAAGCGCAGCAATCCCGATACTGCTCAGGCCTCTCGAAAAGAGAGGTCTGACGGGATCGCTGTGAACTCTCCATGAGTATCCGCTCTTCTCTGTCCTGTAGAGTCGAACTTGGCTTGGTTGT

Methods for Addressing the Challenges

- Dynamic programming
- Hashing
- Graph algorithms
- Parallel processing

Impact of Finding Genetic Variation

- Molecular evolution
- Association of genetic variation with traits
- Personalized lifestyle
- Precision agriculture

Genetic Variation and Reproduction

- Genetic variation is the ‘raw material’ for natural selection.
- Sexual reproduction generates genetic variation by recombination.
- Asexual reproduction lacks mechanisms to generate genetic variation.

Asexual reproduction in Fungi

- Many important plant pathogens have a long period of asexual reproduction.
- Two asexual plant pathogens:

  the fungus *Verticillium dahliae*

  the fungus *Fusarium virguliforme*

Fungus *Fusarium virguliforme*

- Sudden death syndrome (SDS) of soybean first appeared in Arkansas in 1971 and then spread in 30 years to all major soybean-producing regions in the U.S.
- SDS in North America is mostly caused by the fungus *Fusarium virguliforme (Fv)*.
- SDS is similar to Bean Root Rot (BRR).

Low Genetic Diversity but Variable Aggressiveness

- Low genetic diversity is detected in *Fv* populations based on small sequence data.
- But *Fv* isolates show variable aggressiveness on soybean.
- Paradox: *Fv* is subject to Muller's Ratchet but is highly adaptive.

Genome Sequence Data

- 454 sequences of up to 400 bp:
  3 Gb of data for *Fv* isolate Mont-1, single reads and paired reads with 3-kb and 20-kb inserts.
- Illumina sequences of 102 bp:
  390 Gb of data for 10 SDS/BRR isolates, 39 Gb of paired-end reads with a 300-bp insert per isolate.

Experimental Results

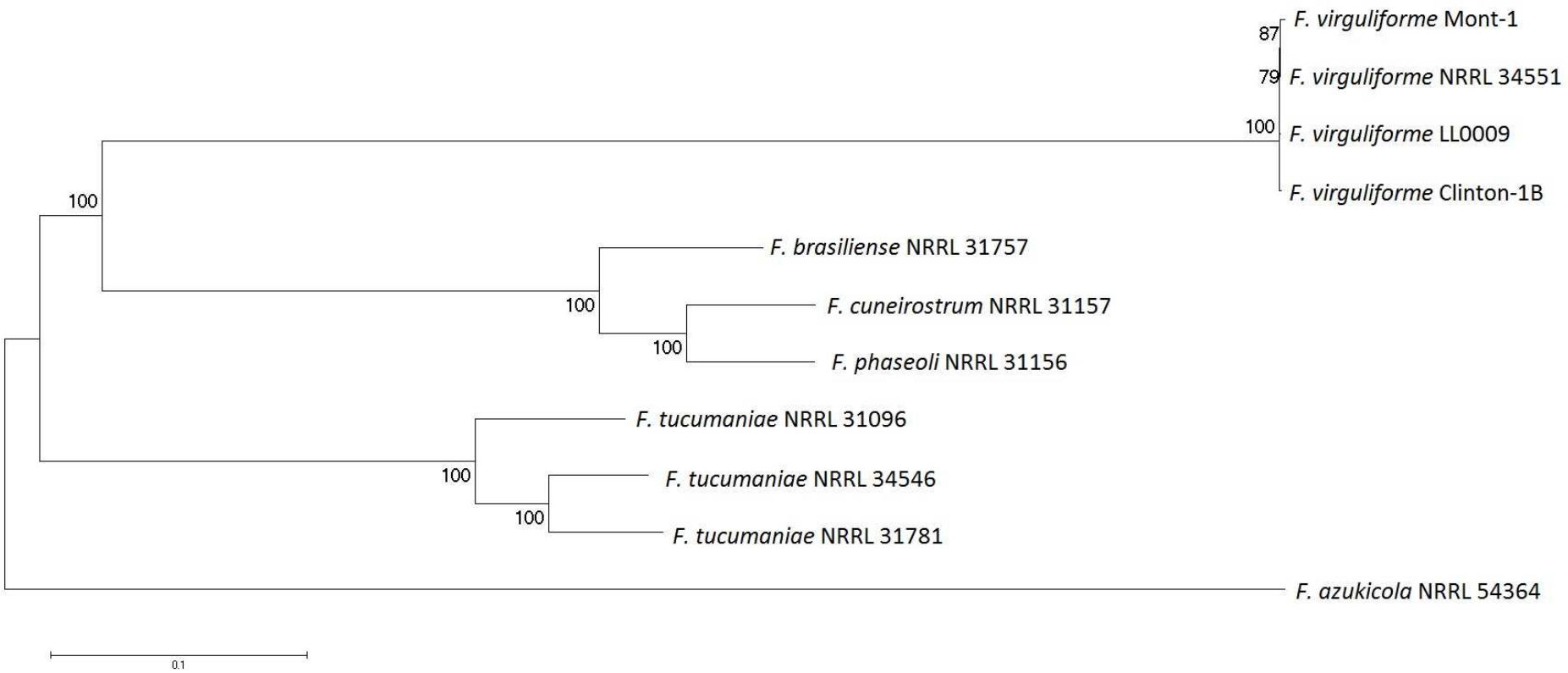
- PCAP was used to produce a chicken genome assembly on ABI 3730 reads (Hillier et al., Nature 432, 695-716).
- PCAP.454 was used to produce a genome assembly of *Fv* Mont-1 (Srivastava et al. PLoS One 9(1), e81832).
- PCAP.Illumina produced an assembly on each of the 10 SDS/BRR isolates.

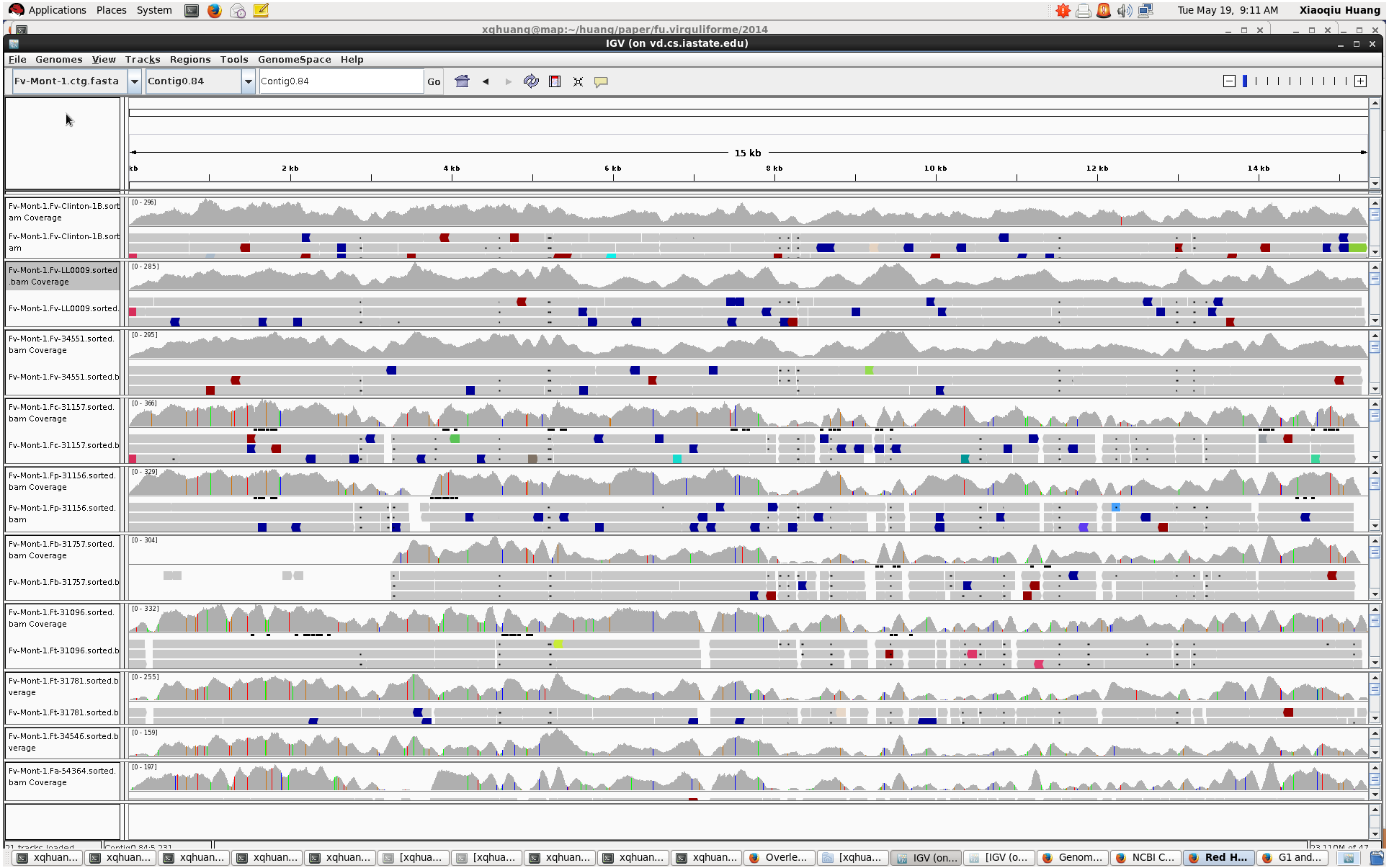
Data Comparison

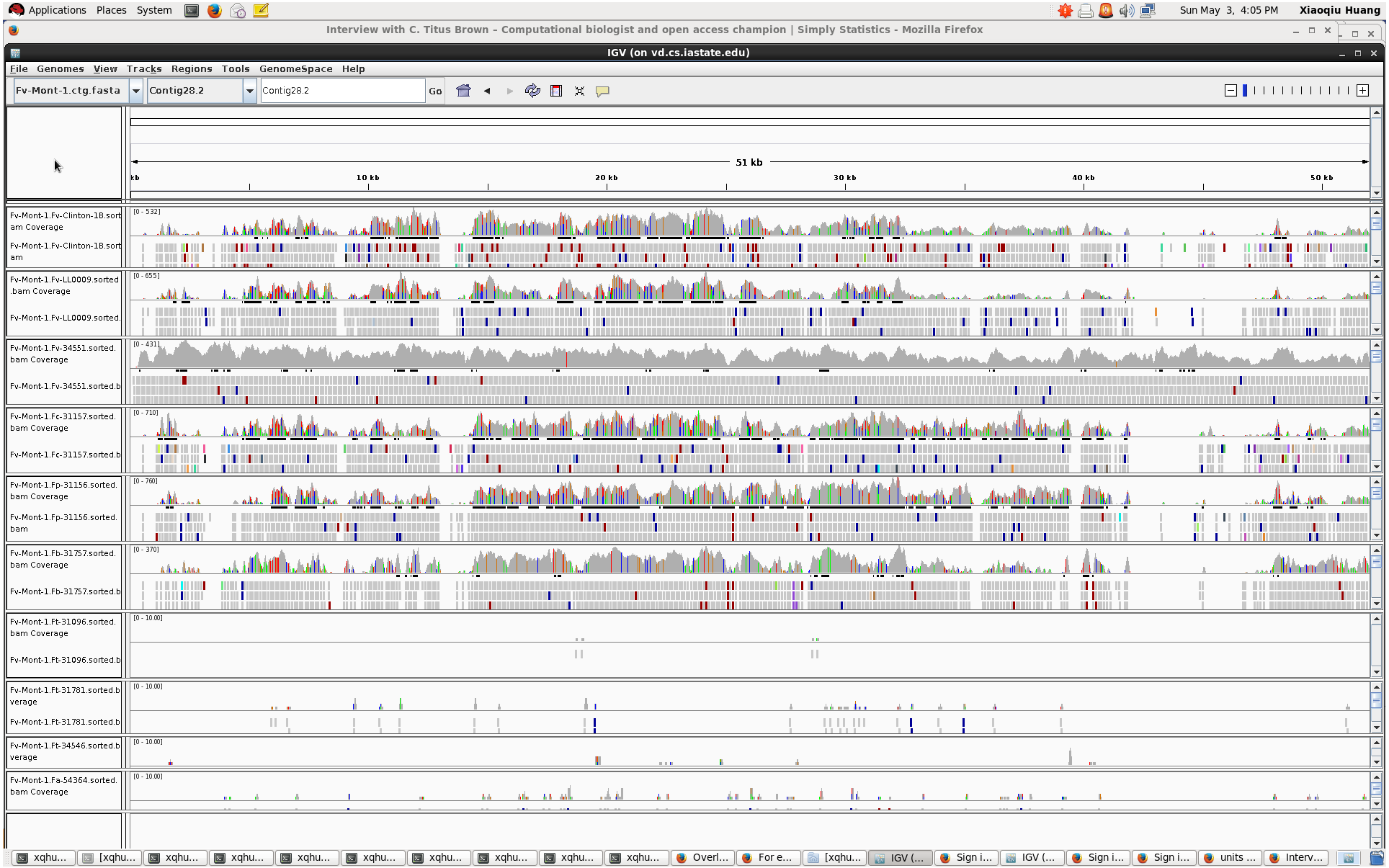
Type	No. of reads (in millions)	Size (Gb)
3730 Chicken	9.8	20
454 Fungus *Fv*	5	3
Illumina F. *Ft*	151	44

Computation Time Comparison

Type	#Processors	Time
3730 Chicken	100	1 week
454 Fungus *Fv*	8	2 days
Illumina F. *Ft*	1	2 days


F. virguliforme Mont-1
F. virguliforme NRRL 34551
F. virguliforme LL0009
F. virguliforme Clinton-1B
F. brasiliense NRRL 31757
F. cuneirostrum NRRL 31157
F. phaseoli NRRL 31156
F. tucumaniae NRRL 31096
F. tucumaniae NRRL 34546
F. tucumaniae NRRL 31781
F. azukicola NRRL 54364
87
79
100
100
100
100
100
100
0.1


Applications Places System
Tue May 19, 9:11 AM Xiaoqiu Huang
xqhuang@map:~/huang/paper/fu.virguliforme/2014
IGV (on vd.cs.iastate.edu)
File Genomes View Tracks Regions Tools GenomeSpace Help
Fv-Mont-1.ctg.fasta
Contig0.84
Contig0.84
Go
15 kb
kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
Fv-Mont-1.Fv-Clinton-1B.sorted.bam Coverage
Fv-Mont-1.Fv-Clinton-1B.sorted.bam
[0 - 296]
Fv-Mont-1.Fv-LL0009.sorted.bam Coverage
Fv-Mont-1.Fv-LL0009.sorted.
[0 - 285]
Fv-Mont-1.Fv-34551.sorted.bam Coverage
Fv-Mont-1.Fv-34551.sorted.b
[0 - 295]
Fv-Mont-1.Fc-31157.sorted.bam Coverage
Fv-Mont-1.Fc-31157.sorted.b
[0 - 366]
Fv-Mont-1.Fp-31156.sorted.bam Coverage
Fv-Mont-1.Fp-31156.sorted.bam
[0 - 329]
Fv-Mont-1.Fb-31757.sorted.bam Coverage
Fv-Mont-1.Fb-31757.sorted.b
[0 - 304]
Fv-Mont-1.Ft-31096.sorted.bam Coverage
Fv-Mont-1.Ft-31096.sorted.b
[0 - 332]
Fv-Mont-1.Ft-31781.sorted.bverage
Fv-Mont-1.Ft-31781.sorted.b
[0 - 255]
Fv-Mont-1.Ft-34546.sorted.bverage
[0 - 159]
Fv-Mont-1.Fa-54364.sorted.bam Coverage
[0 - 197]
IGV (on...
Red H...


Applications Places System
Sun May 3, 4:05 PM Xiaoqiu Huang
Interview with C. Titus Brown – Computational biologist and open access champion | Simply Statistics - Mozilla Firefox
IGV (on vd.cs.iastate.edu)
File Genomes View Tracks Regions Tools GenomeSpace Help
Fv-Mont-1.ctg.fasta
Contig28.2
Contig28.2
Go
51 kb
kb
10 kb
20 kb
30 kb
40 kb
50 kb
Fv-Mont-1.Fv-Clinton-1B.sorted.bam Coverage
Fv-Mont-1.Fv-Clinton-1B.sorted.bam
[0 - 532]
Fv-Mont-1.Fv-LL0009.sorted.bam Coverage
Fv-Mont-1.Fv-LL0009.sorted.
[0 - 655]
Fv-Mont-1.Fv-34551.sorted.bam Coverage
Fv-Mont-1.Fv-34551.sorted.b
[0 - 431]
Fv-Mont-1.Fc-31157.sorted.bam Coverage
Fv-Mont-1.Fc-31157.sorted.b
[0 - 710]
Fv-Mont-1.Fp-31156.sorted.bam Coverage
Fv-Mont-1.Fp-31156.sorted.bam
[0 - 760]
Fv-Mont-1.Fb-31757.sorted.bam Coverage
Fv-Mont-1.Fb-31757.sorted.b
[0 - 370]
Fv-Mont-1.Ft-31096.sorted.bam Coverage
Fv-Mont-1.Ft-31096.sorted.b
[0 - 10.00]
Fv-Mont-1.Ft-31781.sorted.bverage
Fv-Mont-1.Ft-31781.sorted.b
[0 - 10.00]
Fv-Mont-1.Ft-34546.sorted.bverage
[0 - 10.00]
Fv-Mont-1.Fa-54364.sorted.bam Coverage
[0 - 10.00]
xqhu...
[xqhu...
xqhu...
xqhu...
xqhu...
xqhu...
xqhu...
xqhu...
xqhu...
xqhu...
xqhu...
Overl...
For e...
[xqhu...
IGV (...
Sign i...
IGV (...
Sign i...
Sign i...
units ...
Interv...

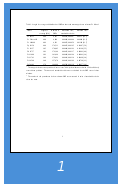


Table 1: Length of coverage and distribution of SNPs when reads were mapped onto reference *Fv* Mont-1

Isolate	Length of coverage (Mb)	Number of SNPs	Mean SNP rate/ standard deviation[a]	Max SNP rate[b]
Fv 34551	49.9	4,955	0.00003/0.00007	0.00177 (23.7)
Fv Clinton-1B	49.6	8,269	0.00006/0.00044	0.00960 (21.5)
Fv LL0009	49.2	8,541	0.00007/0.00052	0.01129 (21.7)
Fp 31156	40.0	178,511	0.00447/0.00125	0.01097 (5.2)
Fc 31157	39.5	176,065	0.00446/0.00123	0.01126 (5.5)
Fb 31757	39.3	172,100	0.00435/0.00117	0.00903 (4.0)
Ft 31096	39.3	181,420	0.00462/0.00128	0.00943 (3.8)
Ft 31781	39.2	172,823	0.00441/0.00114	0.00829 (3.4)
Ft 34546	38.9	157,076	0.00412/0.00102	0.00726 (3.1)
Fa 54363	37.9	188,209	0.00506/0.00119	0.00957 (3.8)

[a] The mapped reference was partitioned into at least 700 disjoint windows each with 35-kb sufficiently covered base positions. The mean and standard deviation were calculated for the SNP rates of these windows.

[b] The number in the parentheses is the maximum SNP rate measured in units of standard deviation above the mean.

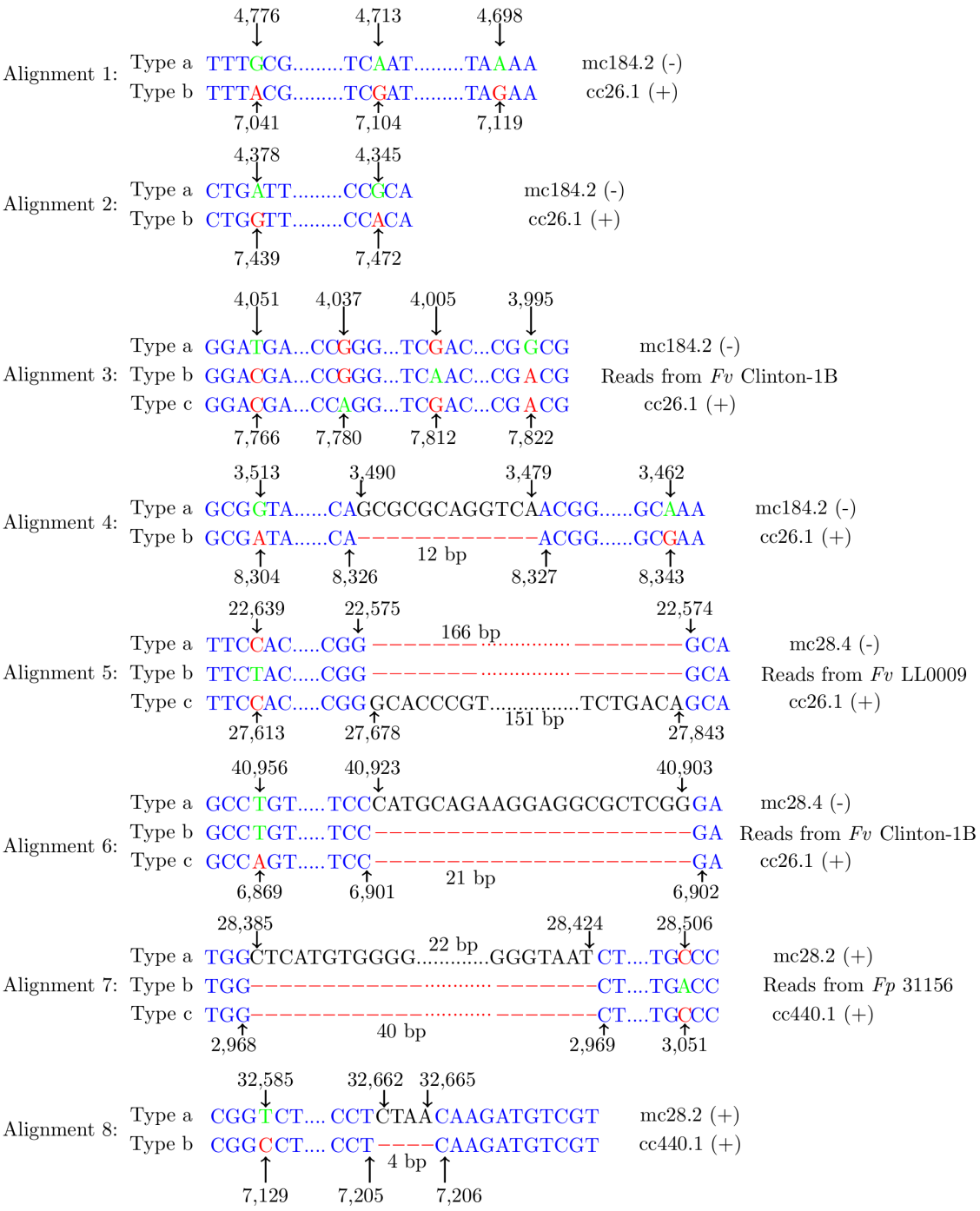

Alignment 1:
4,776
4,713
4,698
Type a TTTGCG.....TCAAT.....TAAAA mc184.2 (-)
Type b TTTACG.....TCGAT.....TAGAA cc26.1 (+)
7,041
7,104
7,119
Alignment 2:
4,378
4,345
Type a CTGATT.....CCGCA mc184.2 (-)
Type b CTGGTT.....CCACA cc26.1 (+)
7,439
7,472
Alignment 3:
4,051
4,037
4,005
3,995
Type a GGATGA...CCGGG...TCGAC...CGGCG mc184.2 (-)
Type b GGACGA...CCGGG...TCAAC...CGACG Reads from Fv Clinton-1B
Type c GGACGA...CCAGG...TCGAC...CGACG cc26.1 (+)
7,766
7,780
7,812
7,822
Alignment 4:
3,513
3,490
3,479
3,462
Type a GCGGTA.....CAGCGCGCAGGTCAACGG.....GCAAA mc184.2 (-)
Type b GCGATA.....CA----------ACGG.....GCGAA cc26.1 (+)
12 bp
8,304
8,326
8,327
8,343
Alignment 5:
22,639
22,575
22,574
166 bp
Type a TTCCAC.....CGG -------.....------GCA mc28.4 (-)
Type b TTCTAC.....CGG -------.....------GCA Reads from Fv LL0009
Type c TTCCAC.....CGGGCACCCGT.....TCTGACAGCA cc26.1 (+)
151 bp
27,613
27,678
27,843
Alignment 6:
40,956
40,923
40,903
Type a GCCTGT.....TCCCATGCAGAAGGAGGCGCTCGGGA mc28.4 (-)
Type b GCCTGT.....TCC----------GA Reads from Fv Clinton-1B
Type c GCCAGT.....TCC----------GA cc26.1 (+)
21 bp
6,869
6,901
6,902
Alignment 7:
28,385
28,424
28,506
22 bp
Type a TGGCTCATGTGGGG.....GGGTAATCT....TGCCC mc28.2 (+)
Type b TGG----------.....-------CT....TGACC Reads from Fp 31156
Type c TGG----------.....-------CT....TGCCC cc440.1 (+)
40 bp
2,968
2,969
3,051
Alignment 8:
32,585
32,662
32,665
Type a CGGTCT.... CCTCTAACAAGATGTCGT mc28.2 (+)
Type b CGGCCT.... CCT----CAAGATGTCGT cc440.1 (+)
4 bp
7,129
7,205
7,206

Table 1: The number of reads from the isolate that link all alleles in the sequence

Sequence[a]	Number of reads from the isolate that cover the sequence[b]					
	Fv Clinton-1B	*Fv* LL0009	*Fv* 34551	*Fc* 31157	*Fp* 31156	*Fb* 31757
A1.Ta	16	16	16	0	0	0
A1.Tb	32	18	0	48	46	8
A2.Ta	88	149	147	0	0	9
A2.Tb	85	114	0	115	0	0
A3.Ta	52	41	61	0	0	0
A3.Tb	78	52	0	152	0	0
A3.Tc	33	0	0	34	46	0
A4.Ta	162	134	77	121	0	97
A4.Tb	54	0	0	127	242	0
A5.Ta	39	27	57	0	0	18
A5.Tb	0	8	0	39	0	0
A5.Tc	85	0	0	69	65	0
A6.Ta	0	0	46	0	0	0
A6.Tb	72	0	0	0	209	0
A6.Tc	116	121	35	554	244	74
A7.Ta	0	0	98	0	0	0
A7.Tb	0	31	0	0	54	0
A7.Tc	35	0	0	42	50	39
A8.Ta	0	35	34	40	42	0
A8.Tb	70	0	0	51	59	42

[a] Each sequence is denoted by its alignment number and type letter (Figure X): e.g., Types a and b in Alignment 1 are denoted by A1.Ta and A1.Tb, respectively.

[b] A read covers a sequence in a set of polymorphic sequences if the read and the sequence have the same allele at every occurrence of polymorphism.

Conclusion

- Low SNP rates show that the assembly consensus sequence is accurate.
- They confirm that *Fv* is asexual in reproduction mode.
- *Fv* genome has mutational hot regions.
- *Fv* has novel mechanisms for generating mutations.

Acknowledgements

- Madan Bhattacharyya at ISU
- Leonor Leandro at ISU
- Kerry O'Donnell at USDA ARS
- Subodh Srivastava at U. of Arkansa
- Shiaw-Pyng Yang at Monsanto
- Colleagues at WashU