

Overview and Implementation of the GBS Pipeline

**Qi Sun
Computational Biology Service Unit
Cornell University**

Genetic Marker Discovery and Genotyping

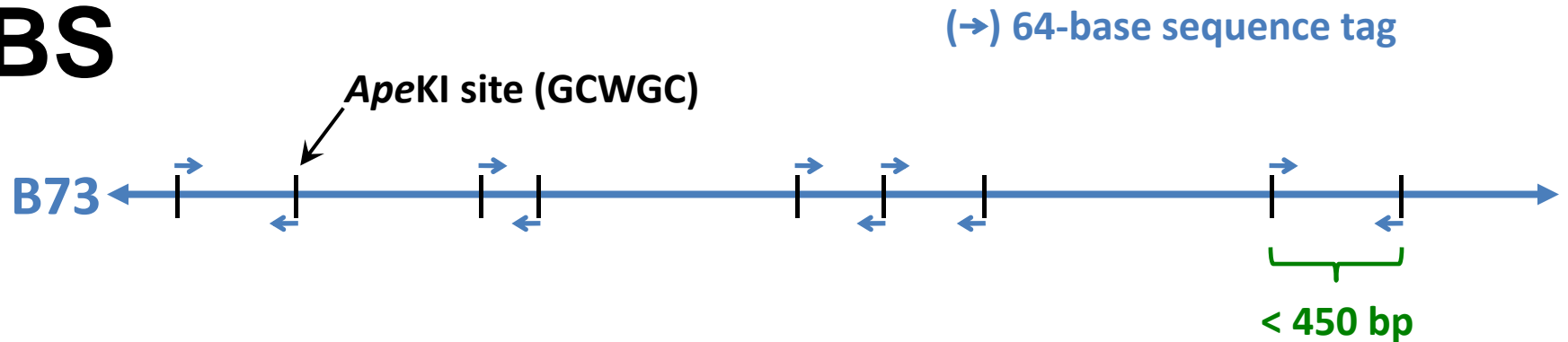
GBS

vs

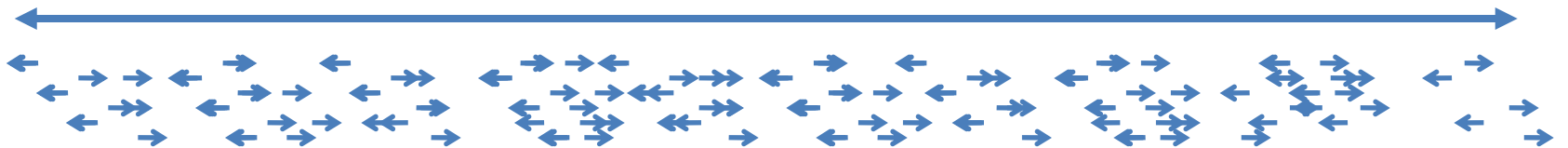
Whole Genome Shotgun

Reduced Genome Representation through GBS

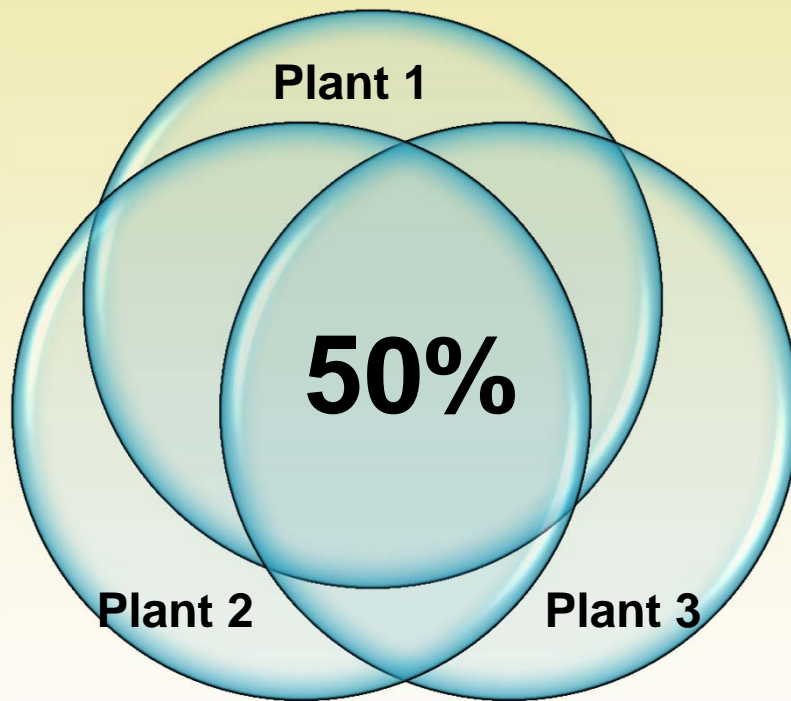
GBS



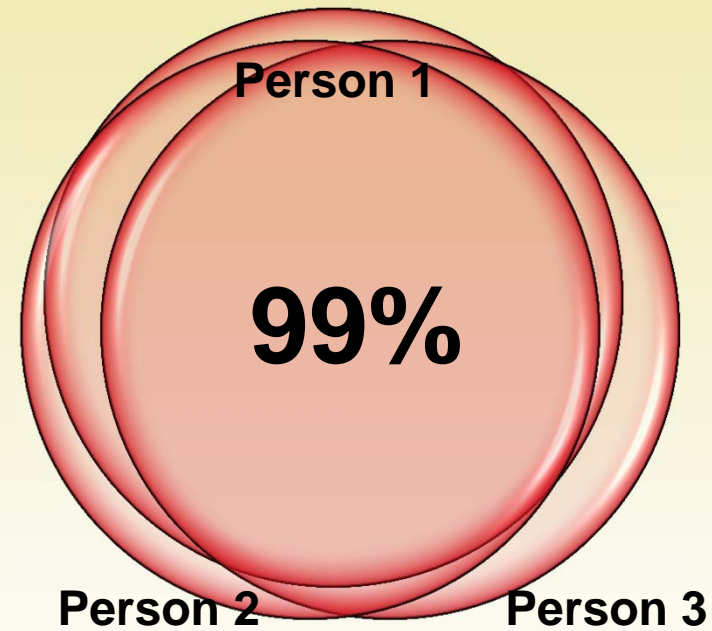
WGS



Only 50% of the maize genome is shared between two varieties



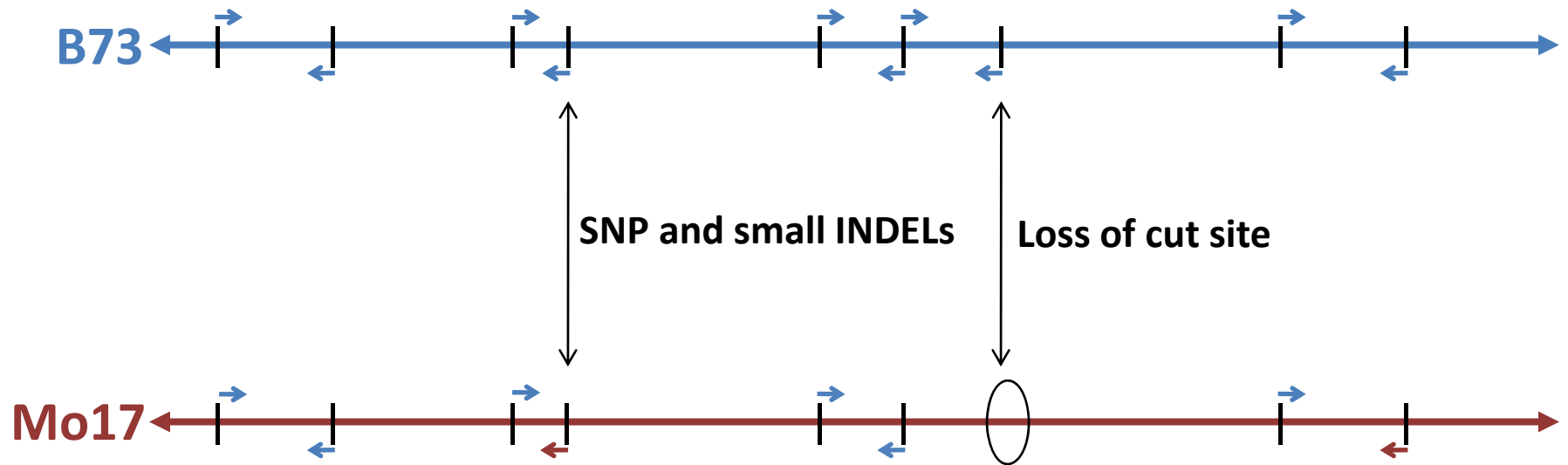
Maize



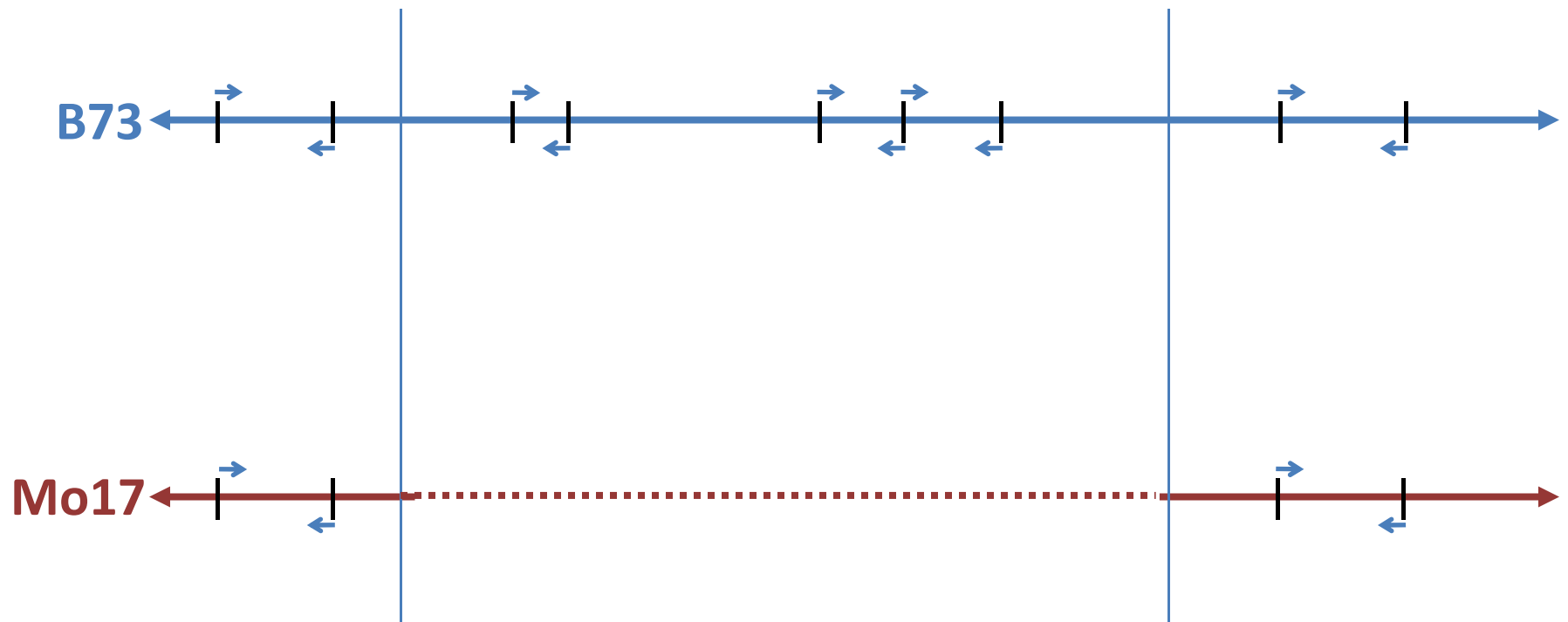
Humans

Fu & Dooner 2002, Morgante et al. 2005, Brunner et al 2005
Numerous PAVs and CNVs - Springer, Lai, Schnable in 2010

Identification of markers with/without the reference genome



Identification of Presence/Absence Variations (PAV)



GBS Data Analysis Pipeline

GBS Experimental Design

What is appropriate library coverage?

- Coverage can be controlled by enzyme choices:
 - ApeKI (5bp) – Maize, sorghum, grape, switchgrass, cacao
 - PstI (6bp) – wheat, barley, maize
 - PstI (7bp – using ApeKI overhangs) – scrub jay (Chen)
 - HindIII - cacao

What is appropriate library coverage?

- Coverage can be controlled by degree of multiplexing in sequencing:
 - 48 plexing
 - 96 plexing
 - 384 plexing
- Desired coverage depends on how many markers are needed and tolerance of missing data.

GBS Data Analysis Pipeline

Terms Used in the Pipeline

Reads -> Tags -> Aligned Tags -> SNPs/INDELS

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGC**A**TGTAGACGGGC

.....

Reads -> Tags -> Aligned Tags -> SNPs/INDELS

Tag 1 { CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC
CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC
CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

Tag 2 { CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC
CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC
CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

Tag 3 { CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGC**A**TGTAGACGGGC

.....

Reads -> Tags -> Aligned Tags -> SNPs/INDELS

┌ CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCGTGCATGGGACACAAGCGTGTAGACGGGC

Tag
Maize NAM population (5000 lines)
2.6 billion reads
6 million tags

Tag
└ CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCGTGCATGGGACACAAGCGTGTAGACGGGC

Tag 3 ┌ CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCGTGCATGGGACACAAGC**A**TGTAGACGGGC

.....

Reads -> Tags -> Aligned Tags ->

SNPs/INDELS

ApeK I site

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

.....

Reads -> Tags -> Aligned Tags ->

SNPs/INDELS

ApeK I site

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**C**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGCGTGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGC**A**TGTAGACGGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGGACACAAGC**A**TGTAGACGGGC

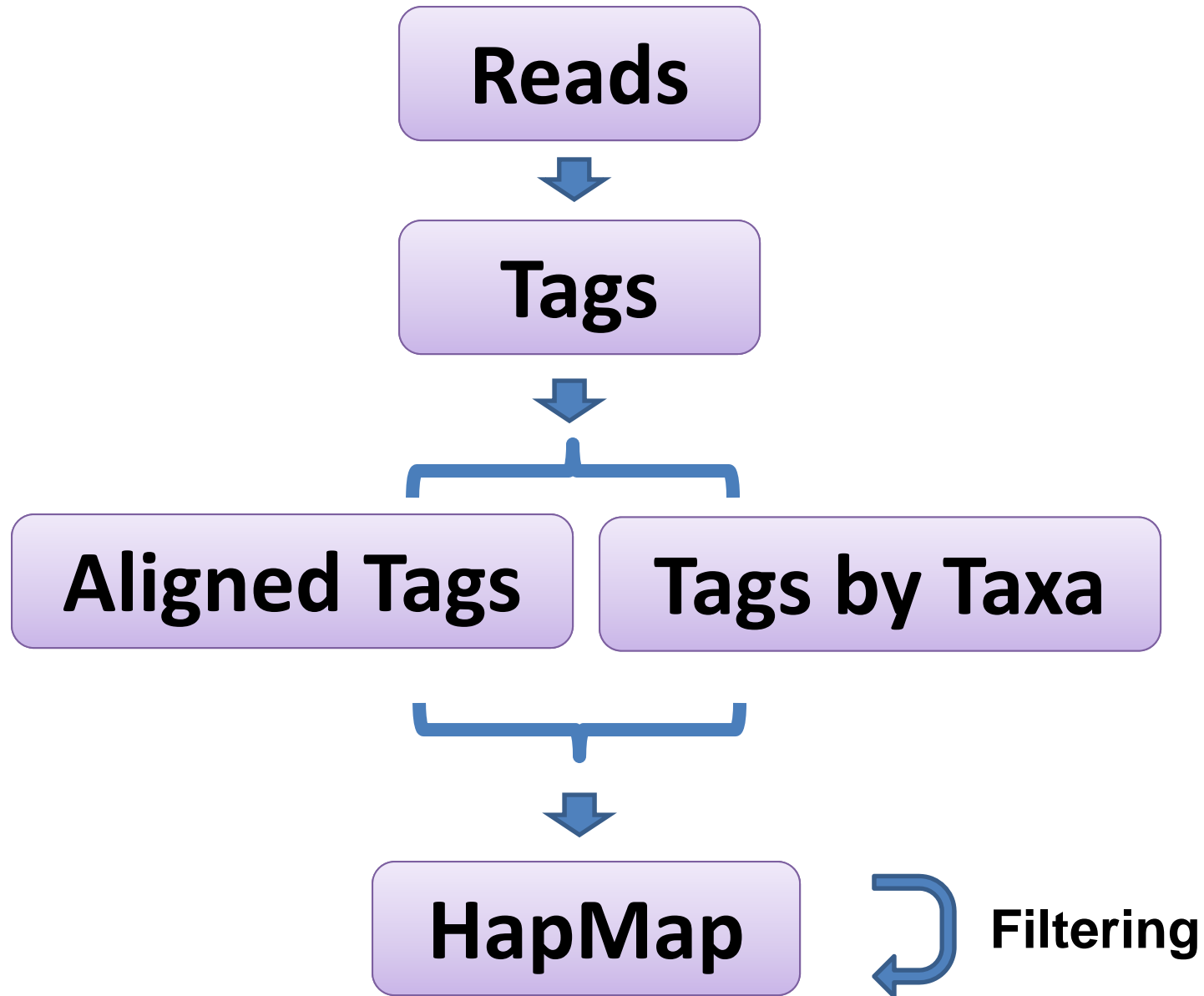
CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGC

CAGCAAAAAAAAAAAGAGGGATG**G**GGCGGCTTGCCTGCATGGC

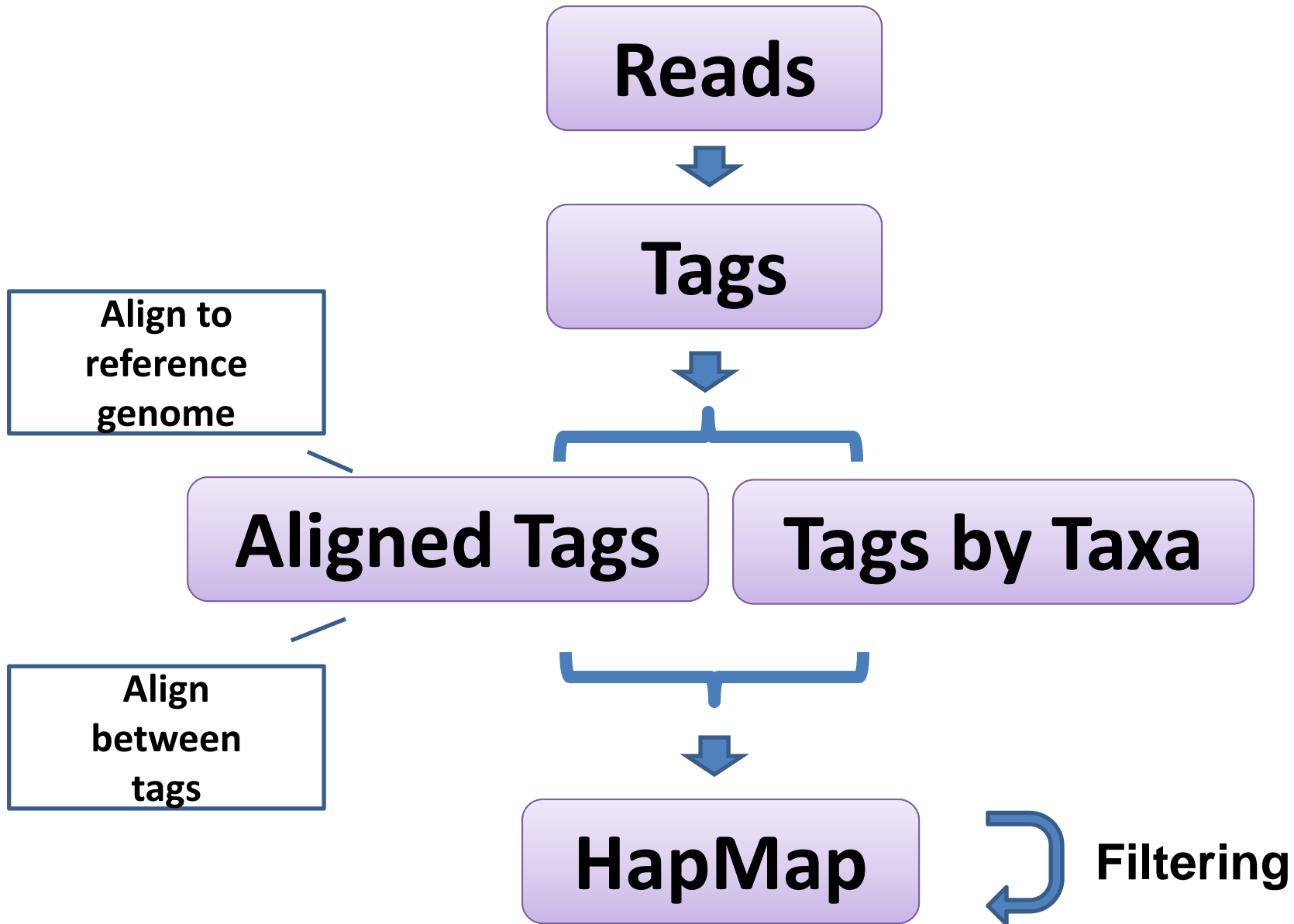
.....

Reproducible sequencing errors

Summary of the GBS pipeline



Summary of the GBS pipeline



GBS Data Analysis Pipeline

File Formats / Data Structure

QSEQ and FASTQ file formats

QSEQ file (pre-QC filter):

HWUSI-EAS690	1	6	54	282	1708	0	1	GAATCAGCTTGCTC CAATCATGCTTTGT TTATTGAATGAAGA GTTCACTCGTAGAA ATGTCCGTTCTTCC TTTGAGCATATCCG T	aaaaa`aaaa`aaa_`^ a_`a_``a``]a``a``^[a __^_`^Y_]`^_]_X]_ Z[^X[W[XUTTTY VVWXVRVUTTRVV VWXX	1
--------------	---	---	----	-----	------	---	---	---	--	---

FASTQ file (post-QC filter):

```
@HWUSI-EAS690:2:1:0:1179#0/1
GAATCAGCTTGCTCCAATCATGCTTTGTTTATTGAATGAAGAGTTCACTCGTAGAAATGTCCGTTCTTCCTTTGAG
CATATCCGT
+HWUSI-EAS690:2:1:0:1179#0/1
aaaaa`aaaa`aaa_`^a_`a_``a``]a``a``^[a__^_`^Y_]`^_]_X]_Z[^X[W[XUTTYVVWXVRVUTTRVVVWXX
```

*** New Illumina pipeline (v1.8) produces FASTQ files only**

TagCounts File (.TC)

26442466	2	
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCTACACCTCATACCACAGGAC	64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAGAATTTTATGTTTCCTACCTCCAACCCAGGACTTT	64	36
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCCTATACCTCATCCCACAGGACTT	64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTTATTTCTACACCTCATACCACAGGACTT	64	25
CAGCAAAAAAAAAAAAAAAAAAACCAAGAATTTGATGTCTCAAACCCCAACACACAGGCTT	64	37
CAGCAAAAAAAAAAAAAAAAAAACCAAGAATTTTTTGTCTCAAACCCCAACCCCAGGCCT	64	1
CAGCAAAAAAAAAAAAAAAAAAAGGGGTTTTGAATAAAAAAACTGAAGGAGCAGAAAAAAAA	58	17
CAGCAAAAAAAAAAAAAAAAAAACCAAGAAATTTGATGTTTCATACCTCATACCACAGGACT	64	23
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCTACACCTCATACCACAGGACT	64	2
CAGCAAAAAAAAAAAAAAAAAAACCAAAAATTTTATGTCTCAAACCCCAACCCCAGGGCTTC	64	1
CAGCAAAAAAAAAAAAAAAAAAACCAATAATTTGATGTCTACACCTCATACCACAGGGCTTC	64	35

TagCounts File (.TC)

26442466	Total #Tags	2	
CAGCAAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGAC		64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAGAATTTTATGTTTCCTACCTCCAACCCAGGACTTT		64	36
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCCTATACCTCATCCCACAGGACTT		64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTTATTTCTCATACCTCATACCACAGGACTT		64	25
CAGCAAAAAAAAAAAAAAAAAAACCAAGAAATTTGATGTCTCAAACCCCAACACACAGGCTT		64	37
CAGCAAAAAAAAAAAAAAAAAAACCAAGAAATTTTTGTCTCAAACCCCAACCCCAGGCCT		64	1
CAGCAAAAAAAAAAAAAAAAAAAGGGGTTTTGAATAAAAAAACTGAAGGAGCAGAAAAAAAAA		58	17
CAGCAAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTTTCATACCTCATACCACAGGACT		64	23
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCT			

Tags

Tag size

#Reads

Tags-By-Taxa File (.TBT)



6040401	2	88																
Tag	Length	08.073	08.073	08.073	08.073	08.073												
	h	1-19	1-29	1-6	1-24	1-37												
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCAA AGGACTT	64	0	0	1	0	0												
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCAC AGGACTT	64	0	1	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCCC AGGACTT	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCAC AGGACTT	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAAACCAAA AGGACTT	64	1	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAACAGG ACTTCCC	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAGGCC AAGAAAGG	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAGGGATGGGGCGGCTTGCGGGCATGGGACACAAGCGTG TAGACGGGC	64	0	1	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAGGGATGGGGCGGCTTGCGTGCATGGGACACAACCGTG TAGACGGGC	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCCTG TAGACGGGC	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGGG TAGACGGGC	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTG GAGACGGGC	64	0	0	0	0	0												
CAGCAAAAAAAAAAAAAAAAAAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTG GAGACGGGC	64	0	1	0	0	0												

Tags

Taxa

0. absence
1. presence

Tags-On-Physical-Map File (.TOPM)

	1020631	2	4					
CAGAGATAAACGCAAGCAGAAGCAGGAGGAGGGAGGCAAAGATGACAAGGCCATGGGCA GTGAC	64	1	11	-1	28230985	28230922	0	
CAGAGATAAACTAAAGCATCGATCCATCAAATGAATTATGTAATTGCAAAAGTCAATTATGC C	64	1	11	-1	3832211	3832148	0	
CAGAGATAAAGAGAGAGAAAACTCAAAAAGAAAAGCATATCATGCAGGCGACGCAGAGAG ATA	64	1	6	-1	27993074	27993011	0	
CAGAGATAAATCCAAACGTTTTCTGTTCCGGTAGGGTCTGTCAGCGAGGCAGAAAAA AA	51	1	7	1	3250207	3250257	0	
CAGAGATAAATTTGAACGGTAGTTAACTAGCAAATTCGTCTATATTAATTCATAGTATGAC A	64	1	2	1	34487829	34487892	0	
CAGAGATAACTAGTTAATTAGTTGCATGCAAGTGTGTGCTCTTGTGCTATTGATTAAGCTGG TA	64	1	2	1	1249312	1249375	0	
CAGAGATAACTTATATATTTGCATCAAAGATAATCCAGCCGACAGCCGACATATGTAAGA CG	64	1	5	-1	24463120	24463057	0	
CAGAGATAAGCTTTAGGGCACGAATCATATCTTCAACAGCATCTCTATAGCACGCAATGGC AG	64	1	3	-1	27908515	27908452	0	
CAGAGATAAGGATACCTACTGTGGATAAATACACAGAATATACACGTGACGTGACCGGAC ATC	64	1	1	1	41245835	41245898	0	
CAGAGATAATAGTAGGATTCTGCGCAATGCAATCGAGTTTCACGCCTCTGTTCTTCAGTTCT TG	64	1	6	-1	28001431	28001368	0	
CAGAGATAATATCATCAGTACCTTCTTTATTTCCAGAACGCCATCACCATCAGCTGAAAAAA A	57	1	6	-1	8065543	8065487	2	
CAGAGATAATGCTCTTACACGGTACAAGGAGAAGAAGAAGAGAAGAAAGTAAGTGCATCTTA CT	64	2*	*	*		*	0	
CAGAGATAATTAATATGTTTAGTTTTGACAGGTGAAGGAACAATAGTTAGCTTATCTGTCAT G	64	1	1	1	24319222	24319285	0	
CAGAGATAATTGACAATGACGTAGGCATATTGACAACCTCAAAAATTATGCACAAAAATGTCC A	64	1	5	1	8796128	8796191	0	
CAGAGATAACAACCATCAAGAGAATCAAAGCAACGATACTCGGCTGAGCTGAAAAAAAAAAAA AA	50	1	8	1	23956554	23956603	0	
CAGAGATAACAACGACGTGGCAAGGACAAGTCAGCCACGCGAGTTCACTGGACGGCGGACA CGGG	64	1	4	1	2274574	2274637	0	
CAGAGATACAGTGTGGGTTTCGATCATCCGGGATACTTCCATTTCTTATTTCTTTACTCAGAA A	64	1	7	1	20512714	20512777	0	
CAGAGATACATAGGCTAACCGGCAGTTGCCAACTAAGACTAATAAAAACTTCATCAAGAAAA AA	64	1	8	1	23636734	23636797	0	
CAGAGATACATAGGCTAATCGGCAGTTGCCAACTAAGACTAATAAAAACTTCATCAAGAAAA AA	64	1	8	1	23636734	23636797	1	
CAGAGATACATCATAACTGTGATACCTCAGTAGGACCACCAGGGTTGTGACTAGCTGTAAAA AT	64	1	3	-1	29627153	29627090	0	

Tags-On-Physical-Map File (.TOPM)

1020631	2	4						
CAGAGATAAACGCAAGCAGAAGCAGGAGGAGGGAGGCAAAGATGACAAGGCCATGGGCA GTGAC	64	1	11	-1	28230985	28230922	0	
CAGAGATAAACTAAAGCATCGATCCATCAAAATGAATTATGTAATTGCAAAAGTCAATTATGC C	64	1	11	-1	3832211	3832148	0	
CAGAGATAAAGAGAGAGAAAACTCAAAAAGAAAAGCATATCATGCAGGCGACGCAGAGAG ATA	64	1	6	-1	27993074	27993011	0	
CAGAGATAAATCCAAACGTTTTCTGTTCCGGTAGGGTCTGTCAGCGAGGCAGAAAAA AA	51	1	7	1	3250207	3250257	0	
CAGAGATAAATTTGAACGGTAGTTAACTAGCAAATTCGTCTATATTAATTCATAGTATGAC A	64	1	2	1	34487829	34487892	0	
CAGAGATAACTAGTTAATTAGTTCATGCAAGTGTGTGCTCTTGTGCTATTGATTAAGCTGG TA	64	1	2	1	1249312	1249375	0	
CAGAGATAACTTATATTTGCATCCAAAGATAATCCAGCCGACAGCCGACATATGTAAGA CG	64	1	5	-1	24463120	24463057	0	
CAGAGATAAGCTTTAGGGCACGAATCATATCTTCAACAGCATCTCTATAGCACGCAATGGC AG	64	1	3	-1	27908515	27908452	0	
CAGAGATAAGGATACCTACTGTGGATAAATACACAGAATATACACGTGACGTGACCGGAC ATC	64	1	1	1	41245835	41245898	0	
CAGAGATAATAGTAGGATTCTGCGCAATGCAATCGAGTTTCACGCCTCTGTTCTTCAGTTCT TG	64	1	6	-1	28001431	28001368	0	
CAGAGATAATATCATCAGTACCTTCTTTATTTCCAGAACGCCATCACCATCAGCTGAAAAA A	57	1	6	-1	8065543	8065487	2	
CAGAGATAATGCTCTTACACGGTACAAGGAGAAGAAGAAGAGAAGAAAGTAAGTGCATCTTA CT	64	2*	*	*		*	0	
CAGAGATAATTAATATGTTTAGTTTGACAGGTGAAGGAACAATAGTTAGCTTATCTGTCAT G	64	1	1	1	24319222	24319285	0	
CAGAGATAATTGACAATGACGTAGGCATATTGACAACCTCAAAAATTATGCACAAAAATGTCC A	64	1	5	1	8706122	8706101	0	
CAGAGATAACAACCATCAAGAGAATCAAAGCAACGATACTCGGCTGAGCTGAAAAA AA	50							

In the current pipeline, the TOPM file is derived from the SAM output of the BWA alignment tool.

Tags-On-Physical-Map File (.TOPM)

	1020631	2	4					
CAGAGATAAACGCAAGCAGAAGCAGGAGGAGGGAGGCAAAGATGACAAGGCCATGGGCA GTGAC	64	1	11			35	28230922	0
CAGAGATAAACTAAAGCATCGATCCATCAAAATGAATTATGTAATTGCAAAAGTCAATTATGC C	64	1	11			11	3832148	0
CAGAGATAAAGAGAGAGAAAAACTCAAAAAGAAAAGCATATCATGCAGGCGACGCAGAGAG ATA	64	1	6	-1			27993074	27993011
CAGAGATAAATCCAAACGTTTTCTGTTCCGGTAGGGTCTGTCAGCGAGGCAGAAAAA AA	51	1	7	1			3250207	3250257
CAGAGATAAATTTGAACGGTAGTTAACTAGCAAATTCGTCTATATTAATTCATAGTATGAC A	64	1	2	1			34487829	34487892
CAGAGATAACTAGTTAATTAGTTCATGCAAGTGTGTGCTCTTGTGCTATTGATTAAGCTGG TA	64	1	2	1			1249312	1249375
CAGAGATAACTTATATATTTCATCCAAAGATAATCCAGCCGACAGCCGACATATGTAAGA CG	64	1	5	-1			24463120	24463057
CAGAGATAAGCTTTAGGGCACGAATCATATCTTCAACAGCATCCTCTATAGCACGCAATGGC AG	64	1	3	-1			27908515	27908452
CAGAGATAAGGATACCTACTGTGGATAAATACACAGAATATACACGTGACGTGACCGGAC ATC	64	1	1	1			41245835	41245898
CAGAGATAATAGTAGGATTCTGCGCAATGCAATCGAGTTTCACGCCTCTGTTCTTCAGTTCT TG	64	1	6	-1			28001431	28001431
CAGAGATAATATCATCAGTACCTTCTTTATTTCCAGAACGCCATCACCATCAGCTGAAAAA A	57	1	6	-1		80		2
CAGAGATAATGCTCTTACACGGTACAAGGAGAAGAAGAAGAGAAGAAAGTAAGTGCATCTTA CT	64	2*	*	*				0
CAGAGATAATTAATATGTTTAGTTTTGACAGGTGAAGGAACAATAGTTAGCTTATCTGTCAT G	64	1	1					0
CAGAGATAATTGACAATGCGTAGGCATATTGACAACCTCAAAAATTATGCACAAAAATGTCC A	64							

Max #mismatches

Divergence

Tags

Tag size

Multi-maps

Chr, strand, start and end positions

HapMap Format (.hmp)

rs#	alleles	chrom	pos	strand	assembly#	center	prot LSID	assay LSID	panel LSID	QCc ode	Sample_100 X:3:A4	Sample_101:8 1PVTA A5	Sample_102:8 1PVTA A6	Sample_103:8 1PVTA A7	Sample_104:8 1PVTA A8	Sample_105:8 1PVTA A9	Sample_106:8 1PVTA A10	Sample_107:8 1PVTA B1	Sample_108:8 1PVTA B2
S10_99 22	G/A	10	9922+		MSU_v6.1	IGD	NA	NA	NA	NA	G	A	N	N	N	N	N	N	N
S10_99 33	C/T	10	9933+		MSU_v6.1	IGD	NA	NA	NA	NA	T	C	N	N	N	N	N	N	N
S10_44 338	G/A	10	44338+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	G	N	N	N	N	A
S10_44 378	T/A	10	44378+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	T	N	N	N	N	A
S10_44 425	A/G	10	44425+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	A	N
S10_44 465	T/G	10	44465+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	T	N
S10_45 726	A/C	10	45726+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	A	M	A	M	A	N	A
S10_45 733	T/G	10	45733+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	K	T	T	K	N	K
S10_45 756	T/G	10	45756+		MSU_v6.1	IGD	NA	NA	NA	NA	T	T	T	K	T	T	T	N	T
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	N	N	N	T	T	A	N
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	A	A	T	T	A	N
S10_57 320	A/C	10	57320+		MSU_v6.1	IGD	NA	NA	NA	NA	A	A	A	M	A	N	N	N	A
S10_70 209	A/T	10	70209+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	N	N	N	N	T	N	T
S10_74 853	A/C	10	74853+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	C	N	N	A	N
S10_74 862	T/C	10	74862+		MSU_v6.1	IGD	NA	NA	NA	NA	T	C	N	N	T	C	N	T	N
S10_74 954	A/G	10	74954+		MSU_v6.1	IGD	NA	NA	NA	NA	A	G	N	A	N	N	N	A	G
S10_74 954	A/G	10	74954+		MSU_v6.1	IGD	NA	NA	NA	NA	A	N	N	N	A	G	G	N	G
S10_75 136	A/C	10	75136+		MSU_v6.1	IGD	NA	NA	NA	NA	A	A	N	M	M	A	A	A	A

HapMap Format (.hmp)

rs#	alleles	chrom	pos	strand	assembly#	center	prot LSID	assay LSID	panel LSID	QCc ode	Sample_100 :81PVTABX X:3:A4	Sample_101:8 1PVTABX A5	Sample_102:8 1PVTABX A6	Sample_103:8 1PVTABX A7	Sample_104:8 1PVTABX A8	Sample_105:8 1PVTABX A9	Sample_106:8 1PVTABX A10	Sample_107:8 1PVTABX B1	Sample_108:8 1PVTABX B2
S10_99 22	G/A	10	9922+		MSU_v6.1	IGD	NA	NA	NA	NA	G	A	N	N	N	N	N	N	N
S10_99 33	C/T	10	9933+		MSU_v6.1	IGD	NA	NA	NA	NA	T	C	N	N	N	N	N	N	N
S10_44 338	G/A	10	44338+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	G	N	N	N	N	A
S10_44 378	T/A	10	44378+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	T	N	N	N	N	A
S10_44 425	A/G	10	44425+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	A	N
S10_44 465	T/G	10	44465+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	T	N
S10_45 726	A/C	10	45726+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	A	M	A	M	A	N	A
S10_45 733	T/G	10	45733+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	K	T	T	K	N	K
S10_45 756	T/G	10	45756+		MSU_v6.1	IGD	NA	NA	NA	NA	T	T	T	K	T	T	T	N	T
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	N	N	N	T	T	A	N
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	A	A	T	T	A	N
S10_57 320	A/C	10	57320+		MSU_v6.1	IGD	NA	NA	NA	NA	A	A	A	M	A	N	N	N	A
S10_70 209	A/T	10	70209+		MSU_v6.1	IGD	NA	NA	NA	NA									

The Hapmap file format is defined by the Human HapMap Project (www.hapmap.org).

HapMap Format (.hmp)

rs#	alleles	chrom	pos	strand	assembly#	center	prot LSID	assay LSID	panel LSID	QCc ode	Sample_100 :81PVTABX X:3:A4	Sample_101:8 1PVTABX A5	Sample_102:8 1PVTABX A6	Sample_103:8 1PVTABX A7	Sample_104:8 1PVTABX A8	Sample_105:8 1PVTABX A9	Sample_106:8 1PVTABX A10	Sample_107:8 1PVTABX B1	Sample_108:8 1PVTABX B2
S10_99 22	G/A	10	9922+		MSU_v6.1	IGD	NA	NA	NA	NA	G	A	N	N	N	N	N	N	N
S10_99 33	C/T	10	9933+		MSU_v6.1	IGD	NA	NA	NA	NA	T	C	N	N	N	N	N	N	N
S10_44 338	G/A	10	44338+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	G	N	N	N	N	A
S10_44 378	T/A	10	44378+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	T	N	N	N	N	A
S10_44 425	A/G	10	44425+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	A	N
S10_44 465	T/G	10	44465+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	T	N
S10_45 726	A/C	10	45726+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	A	M	A	M	A	N	A
S10_45 733	T/G	10	45733+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	K	T	T	K	N	K
S10_45 756	T/G	10	45756+		MSU_v6.1	IGD	NA	NA	NA	NA	T	T	T	K	T	T	T	N	T
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	N	N	N	T	T	A	N
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	A	A	T	T	A	N
S10_57 320	A/C	10	57320+		MSU_v6.1	IGD	NA	NA	NA	NA	A	A	A	M	A	N	N	N	A
S10_70 209	A/T	10	70209+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	N	N

Alleles

Chr

Position

Assembly

Genotypes

HapMap Format (.hmp)

rs#	alleles	chrom	pos	strand	assembly#	center	prot LSID	assay LSID	panel LSID	QCc ode	Sample_100 :81PVTABX X:3:A4	Sample_101:8 1PVTABX A5	Sample_102:8 1PVTABX A6	Sample_103:8 1PVTABX A7	Sample_104:8 1PVTABX A8	Sample_105:8 1PVTABX A9	Sample_106:8 1PVTABX A10	Sample_107:8 1PVTABX B1	Sample_108:8 1PVTABX B2
S10_99 22	G/A	10	9922+		MSU_v6.1	IGD	NA	NA	NA	NA	G	A	N	N	N	N	N	N	N
S10_99 33	C/T	10	9933+		MSU_v6.1	IGD	NA	NA	NA	NA	T	C	N	N	N	N	N	N	N
S10_44 338	G/A	10	44338+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	G	N	N	N	N	A
S10_44 378	T/A	10	44378+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	N	T	N	N	N	N	A
S10_44 425	A/G	10	44425+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	A	N
S10_44 465	T/G	10	44465+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	T	N
S10_45 726	A/C	10	45726+		MSU_v6.1	IGD	NA	NA	NA	NA	N	A	A	M	A	M	A	N	A
S10_45 733	T/G	10	45733+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	K	T	T	K	N	K
S10_45 756	T/G	10	45756+		MSU_v6.1	IGD	NA	NA	NA	NA	T	T	T	K	T	T	T	N	T
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	N	N	N	T	T	A	N
S10_48 453	A/T	10	48453+		MSU_v6.1	IGD	NA	NA	NA	NA	N	T	T	A	A	T	T	A	N
S10_57 320	A/C	10	57320+		MSU_v6.1	IGD	NA	NA	NA	NA	A	A	A	M	A	N	N	N	A
S10_70 209	A/T	10	70209+		MSU_v6.1	IGD	NA	NA	NA	NA	N	N	N	N	N	N	N	N	N

**IUPAC code is used to encode heterozygous sites,
eg. R = A/G**

Binary compression of the files

- **Tag-Counts (TC):**
*.cnt.txt *.cnt
- **Tag-by-taxa (TBT):**
*.tbt.txt *.tbt.bin
- **Tags-on-physical-map (TOPM):**
*.topm.txt *.topm.bin
- **Hapmap**
*.hmp.txt GDPDM blobs

* 64 bp tags were represented as 2 long integers (8 bytes for long in Java).

Bioinformatics Challenges

- **Massive amounts of data**
- **Complex genomes with many unstable parts of a genome**
- **No reference genome**
- **Missing data**
- **Phasing and imputation**

Pipeline Implementation

Where to get the software

- Documentation and software (TASSEL) can be downloaded at <http://www.maizegenetics.net/>.
- GBS Bioinformatics home page
<http://www.maizegenetics.net/gbs-bioinformatics>
- Source code is available at <http://sourceforge.net>
(Project name: TASSEL)
- The Pipeline is installed at Cornell BioHPC lab, and will be available through iPLANT Discovery Environment;


Check List for Running the Pipeline

- ❑ A computer with at least 8GB or more RAM
- ❑ Download TASSEL Standalone from maizegenetics.net
- ❑ BWA (for alignment to reference genome)
- ❑ Netbeans (or other IDE for developers)

The screenshot shows a web browser window displaying the TASSEL website. The browser's address bar shows the URL: www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119. The website header includes the text "Buckler Lab for Maize Genetics and Diversity" and "A USDA-ARS Lab with Cornell's Institute for Genomic Diversity". The main content area is titled "TASSEL" and lists two versions: "Tassel Version 3.0 (Build: September 15, 2011 Requires: Java 1.6)" and "Tassel Version 2.1 (Build: March 15, 2010)". Under the 3.0 version, there are links for "Launch TASSEL 3.0" and "TASSEL 3.0 Standalone". Under the 2.1 version, there are links for "Launch TASSEL 2.1 (Requires: Java 1.6)" and "Launch TASSEL 2.1 (Requires: Java 1.5)". A red box on the right side of the page contains the text: "Download the zip file: TASSEL_x.x _Standalone".

GBS Pipeline on Cornell BioHPC Lab

Step 1: Reserve a machine



Cornell University
Life Sciences Core Laboratories Center
Computational Biology Service Unit

SEARCH CORNELL:

[Pages](#) [People](#) [more options](#)

[HOME](#) [Mission and Services](#) [BioHPC](#) [Staff](#) [Publications and Outreach](#) [CBSU/3CPG BioHPC Lab](#) [Contact us](#) [Forum](#) [Mail list](#)

CBSU / 3CPG BioHPC Laboratory (625 Rhodes Hall)

Reservations for

These workstations can be only accessed remotely via an ssh client.
Total of 26 workstations available

Display reservations from for

New reservation from to for machine

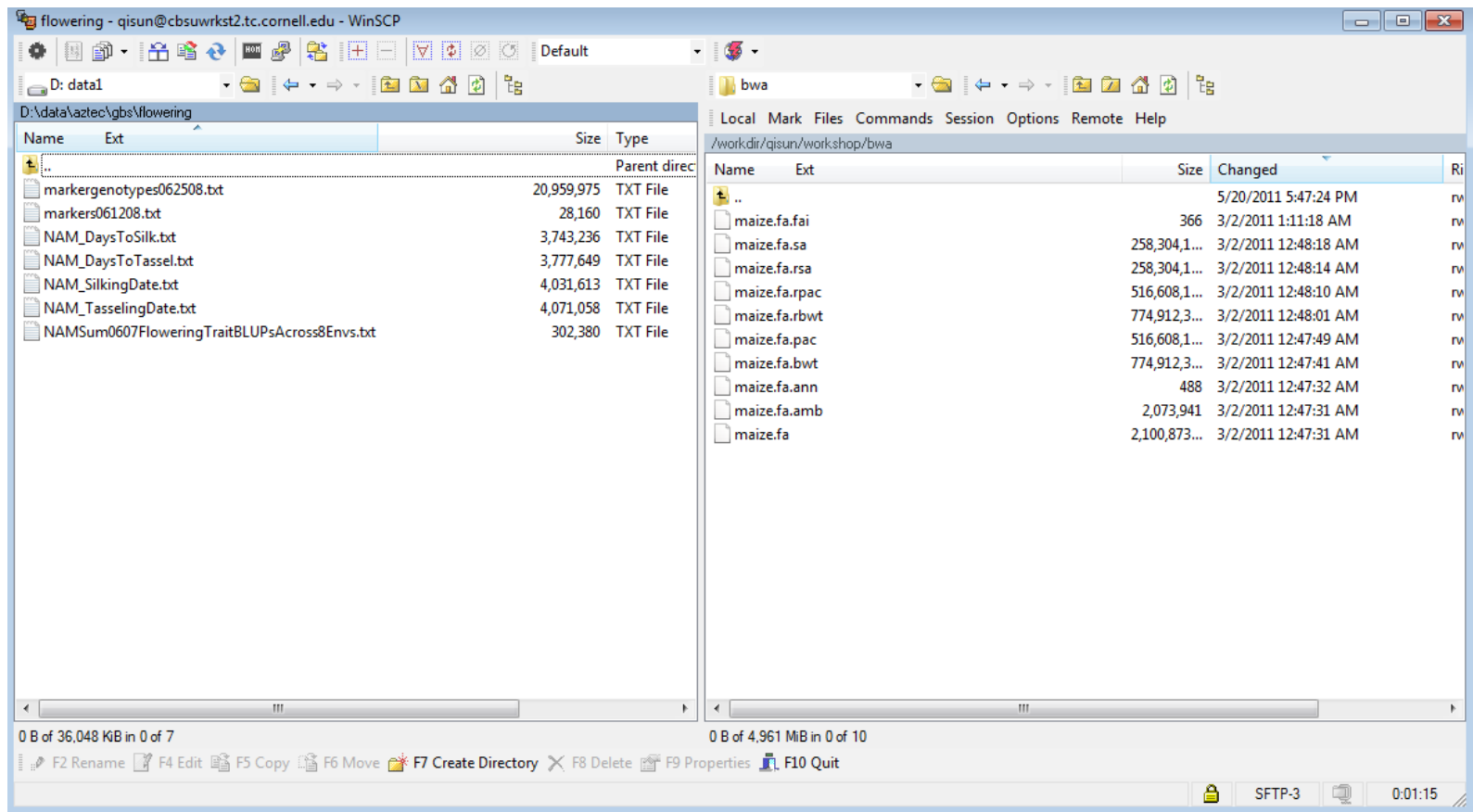
You can delete your reservations by clicking on [x], you can modify your reservations by clicking on them, you can add a new reservation by clicking on the appropriate "AVAILABLE" text or in the box above.

	cbsum1c1b002 Linux	cbsum1c1b003 Linux	cbsum1c1b004 Linux	cbsum1c1b005 Linux	cbsum1c1b006 Linux	cbsum1c1b007 Linux	cbsum1c1b009 Linux	cbsum1c1b010 Linux	cbsum1c1b011 Linux	cbsum1c1b012 Linux	cbsum1c1b013 Linux	cbsum1c1b014 Linux	cbsum1c1b015 Linux	cbsum1c1b016 Linux	cbsum1 Linux	cbsum1c2b001 Linux	cbsum1c2b002 Linux
Sun Sep 18 2011	hs568 UNTIL 11:00 PM AVAILABLE	AVAILABLE	sw54 ALL DAY	AVAILABLE	AVAILABLE	yj55 ALL DAY	AVAILABLE	yj55 ALL DAY	AVAILABLE	jm889 ALL DAY	yj55 ALL DAY	yj55 ALL DAY	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Mon Sep 19 2011	AVAILABLE	AVAILABLE	sw54 UNTIL 10:30 AM AVAILABLE	AVAILABLE	AVAILABLE	yj55 ALL DAY	AVAILABLE	yj55 ALL DAY	AVAILABLE	jm889 ALL DAY	yj55 ALL DAY	yj55 ALL DAY	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Tue Sep 20 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	yj55 ALL DAY	AVAILABLE	yj55 ALL DAY	AVAILABLE	jm889 ALL DAY	yj55 ALL DAY	yj55 ALL DAY	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Wed Sep 21 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	yj55 ALL DAY	AVAILABLE	yj55 ALL DAY	AVAILABLE	jm889 ALL DAY	yj55 ALL DAY	yj55 ALL DAY	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Thu Sep 22 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	yj55 ALL DAY	AVAILABLE	yj55 ALL DAY	AVAILABLE	jm889 ALL DAY	yj55 ALL DAY	yj55 ALL DAY	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Fri Sep 23 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	yj55 ALL DAY	AVAILABLE	yj55 ALL DAY	AVAILABLE	jm889 ALL DAY	yj55 ALL DAY	yj55 ALL DAY	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Sat Sep 24 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	yj55 UNTIL 01:00 PM AVAILABLE	AVAILABLE	yj55 UNTIL 11:00 AM AVAILABLE	AVAILABLE	jm889 ALL DAY	yj55 UNTIL 01:00 PM AVAILABLE	yj55 UNTIL 11:00 AM AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Sun Sep 25 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	jm889 ALL DAY	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Mon Sep 26 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	jm889 UNTIL 11:00 PM AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Tue Sep 27 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Wed Sep 28 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Thu Sep 29 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Fri Sep 30 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Sat Oct 01 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE
Sun Oct 02 2011	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE	AVAILABLE

<http://cbsu.tc.cornell.edu>

GBS Pipeline on Cornell BioHPC Lab

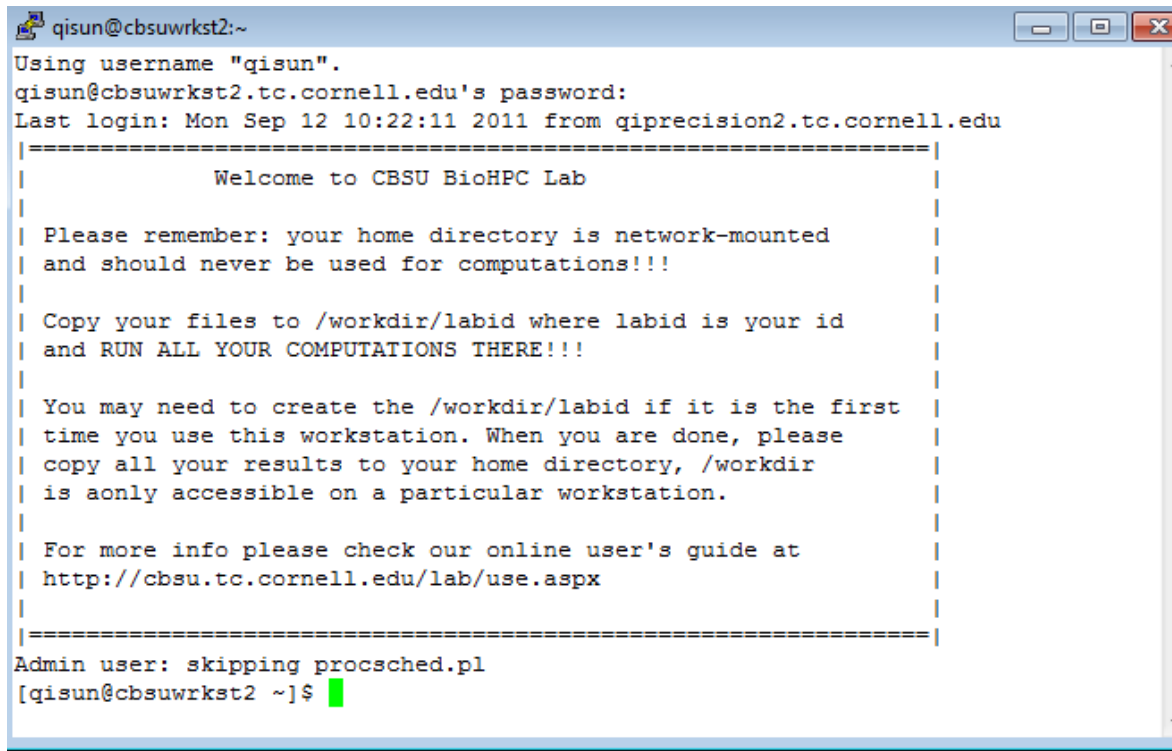
Step 2: Upload files



Fetch (mac), FileZilla (win) or WinSCP (win)

GBS Pipeline on Cornell BioHPC Lab

Step 3: Type the command to run pipeline



```
qisun@cbsuwrkst2:~  
Using username "qisun".  
qisun@cbsuwrkst2.tc.cornell.edu's password:  
Last login: Mon Sep 12 10:22:11 2011 from qiprecision2.tc.cornell.edu  
|-----|  
|           Welcome to CBSU BioHPC Lab           |  
|-----|  
| Please remember: your home directory is network-mounted |  
| and should never be used for computations!!!          |  
|-----|  
| Copy your files to /workdir/labid where labid is your id |  
| and RUN ALL YOUR COMPUTATIONS THERE!!!                |  
|-----|  
| You may need to create the /workdir/labid if it is the first |  
| time you use this workstation. When you are done, please  |  
| copy all your results to your home directory, /workdir   |  
| is aonly accessible on a particular workstation.          |  
|-----|  
| For more info please check our online user's guide at    |  
| http://cbsu.tc.cornell.edu/lab/use.aspx |  
|-----|  
Admin user: skipping procsched.pl  
[qisun@cbsuwrkst2 ~]$
```

Mac: terminal window; PC: Putty

```
tassel/run_pipeline.pl -fork1 -QseqToTagCountPlugin -i . -k rice.key -e  
apeki -endPlugin -runfork1
```

Tutorial:

1. **Documentation of the plugins.**
2. **Step-by-step walkthrough of an exercise project.**

*** Training project data is provided by Chih-Wei Tung & Susan McCouch.**