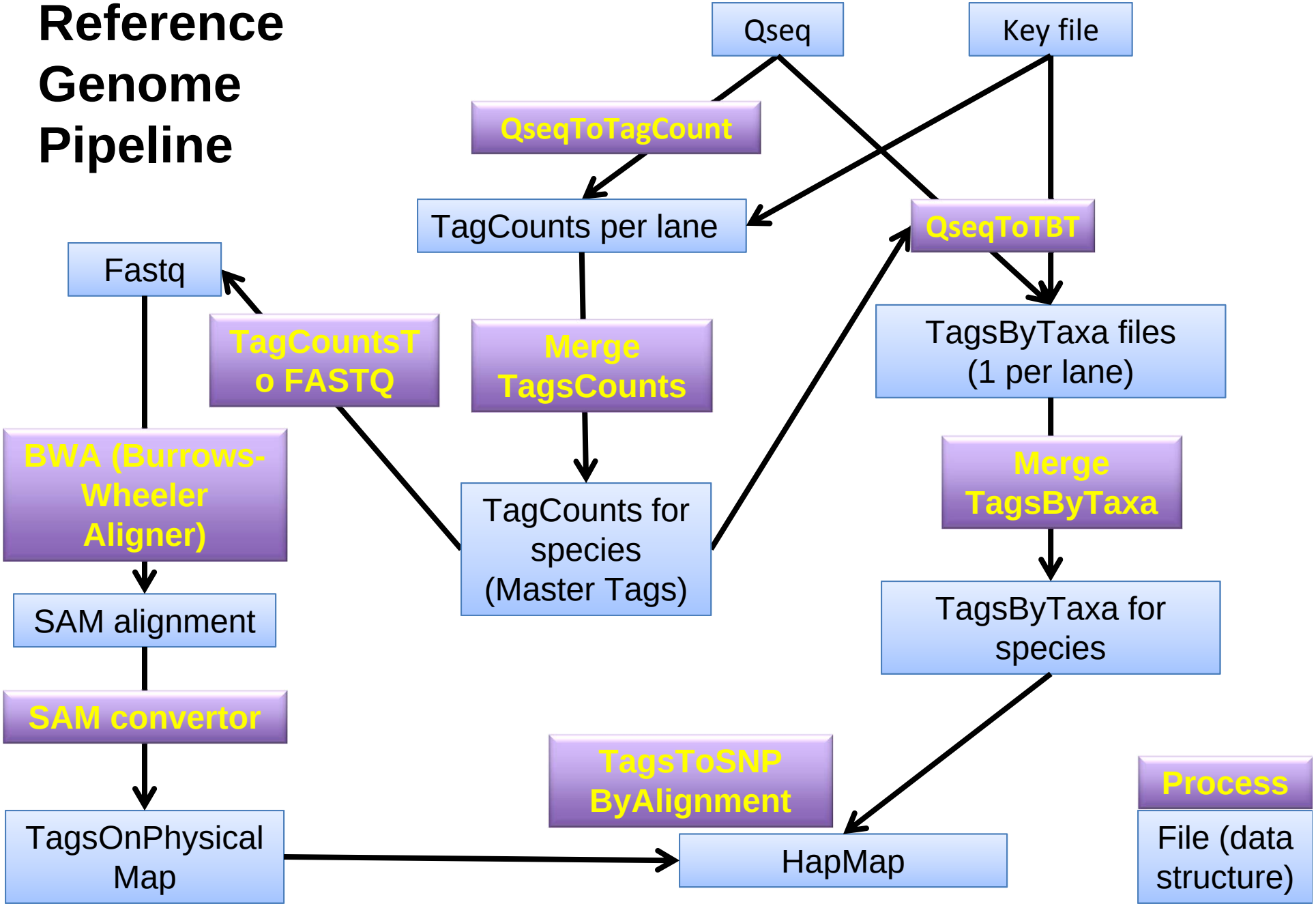
GBS Bioinformatics Pipeline

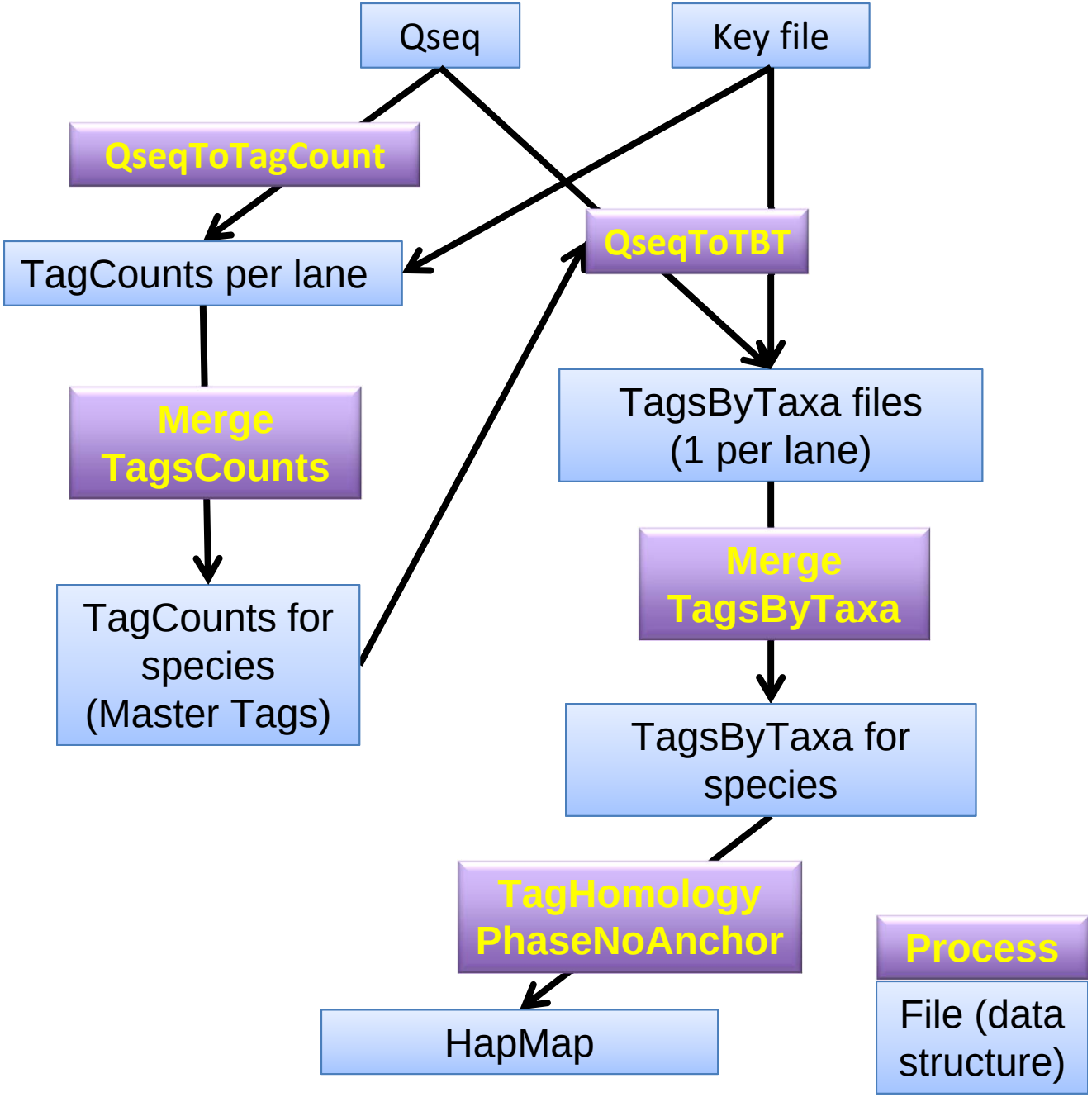
...or, “Where Your Data Go After Sequencing”

James Harriman
Ed Buckler
Jeff Glaubitz

Reference Genome Pipeline



Non-Reference Genome Pipeline



Raw Sequence (Qseq)

HWI-ST397	0	3	68	15896	200039	0	1	GTCGATTCTGCTGACTTCATGGCTTCTGTTGACGACGATGTGGAACGAGCTGTTGTTGAAACTGATGAGGTTG
HWI-ST397	0	3	68	15960	200043	0	1	GAGAATCAGCTTTTCCAACACCTTGAGTTTGAGTATGCGATGACAGTTACTCTTACTGTCCATTGTCAGCATTG
HWI-ST397	0	3	68	15831	200053	0	1	ATGTACTGCACCGTTGCAAGCGAGCACCACCAAGCGGCGGTATGCACTTTGCAATATGTAGCTAGAATAGGAT
HWI-ST397	0	3	68	15867	200049	0	1	CCAGCTCAGCCTGCATTCTTTCAAAAACTTCCAATGCCTCTCTTGGCCTAGCATTTTGGGCATACCCTGTGACC
HWI-ST397	0	3	68	15943	200048	0	1	GATTTTACTGCACATCGGTCTTGTCACACCAGCTATACCTGTAGAGTTGCCTTCCACAGTTGTAGAGATCGGAA
HWI-ST397	0	3	68	15812	200062	0	1	TCACCCAGCATCACGCCCCTTCACATCCAGTAAAACCCCTGAATGATGTGCTGTCACTGTTTGATATACAGTTC
HWI-ST397	0	3	68	15888	200067	0	1	CTTGACTGCCACCATGAATATGTGTTCCAAGTGCCACAAGGACTTGGCCCTGAAGCAAGAACAAGCCAAACTT
HWI-ST397	0	3	68	15969	200067	0	1	CCACAACTGCTCCATCTTTTCCATGAGACATTGCTCCCGCCATTGCACCCTTGGCATCAGCAGAGATCGGAAG
HWI-ST397	0	3	68	15786	200078	0	1	GTATTCTGCACACGAATCAGCTGAGACACCAATTGGGCATGAATCAAATGGCGCCATTGCCGGGGATCGAACC
HWI-ST397	0	3	68	15830	200072	0	1	AATATGCCAGCAGTTAAGAGAGTTCAAGATCCAGGGCTCATATTCAGTCACCTATATCAATTTCGAAATGGATT
HWI-ST397	0	3	68	15863	200073	0	1	CTCCCTGCGGGTGCGCGCGACCCATCTTCAGTTGGAGCGTCTATCGGCGTTGCTGAGATCGGAAGAGCGGTT
HWI-ST397	0	3	68	15762	200088	0	1	TGGTACGTCTGCGGAATGGCGTTTTTTATGCCTTAGTGGTTCGCAGAGCATTTGGCAGCTGAGATGGGAAGAC
HWI-ST397	0	3	68	15903	200085	0	1	GGACCTACTGCCCAAGAACGGCTCACCCATCATCCGCTTTCTTCACCTTCCGTCTTCTTTGGCTGAGATCGGA
HWI-ST397	0	3	68	15921	200082	0	1	GAGAATCAGCGTGTACGGGGCACGGGGTGACTGCTGTTGCGTGCGAGGGCTGAGATCGGAAGAGCGGTTCA
HWI-ST397	0	3	68	15984	200085	0	1	TTCTCCAGCCGCATGGGCCGGAGACCAGAGAGGCCTCCCCAGGATTTGCACGATAGACCACGACTTATGGAC
HWI-ST397	0	3	68	15788	200096	0	1	GCGTCAGCAAATGCCCCAACAGCCAAGTCAGCAATTGCCTCAGCAACTTGGGCCACAAACACCACAGCTGAG
HWI-ST397	0	3	68	15842	200099	0	1	TAGGCCATCAGCTGACTTCCCGGGTGTGGAGAAAAGAGGGCCCCTCACTTCTCTCAAGTGCTGAGATCGGAA
HWI-ST397	0	3	68	15876	200105	0	1	GGACCTACTGCCGGCGGGACGAAAGCGGTTGTTGAATGATGGGGGTCACTAGGCCTTCCAGGGCCTTTAAGC
HWI-ST397	0	3	68	15937	200097	0	1	CTCCCTGTTGAAGCATGTGCAAAAGAGCTTGTTCTCGGCCTTCTTCAAGCCATTCTCTTGGCAGACGGCTTTG
HWI-ST397	0	3	68	15958	200102	0	1	CGCCTTATCTGCCCTCGCCGGTCATGGGGAGTGGTGCCCCTACCTCGGACAAGACAGATGCAGAGATCGGAA
HWI-ST397	0	3	68	15765	200113	0	1	CCAGCTCAGCATGGATCTCTCCTTGATGGACTGAAAGCGCGTGTGCTCCCCTGTGTGATGGAAAGTGGCAGT
HWI-ST397	0	3	68	15912	200114	0	1	CCAGCTCAGCTCAAGCATTGGCTTCCGCTTTGGCATCCTGGAGGGTAAGCTTCTGCTCTTCTCACTAGAGGAC
HWI-ST397	0	3	68	15791	200127	0	1	ACAAACAGCAGAGGTCGCATTGTAGTTAGTCCGGGACTTGCCCAGTTCATTGCTGAGATCGGAAGAGCGGTT
HWI-ST397	0	3	68	15831	200117	0	1	GCTCTACAGCTTCTGGCCAGAATGCTTTTGGCACTTGTTTGTCACAAAGCATGCACTGAACCATATTCATGATA
HWI-ST397	0	3	68	15848	200124	0	1	TTCTCCAGCTGCTACATGCACCGTGGGAAGAAGGTCTGCCCCACATACCCACCAGCCATCGCCCTTCTCACAT
HWI-ST397	0	3	68	15891	200120	0	1	GAGATACAGCTGCGAATTGGGGGTTCCTGTGTTGCGAAGTGGCACTCGTGTGCCAAACTTGGCTACGCAGAG
HWI-ST397	0	3	68	15931	200128	0	1	AAAAGTTCAGCAATACCTGTTGAAGCCAAGCCCTTGTGGTGATTGCCTCGTTCATTGCTGCTGAGATCGGAAG
HWI-ST397	0	3	68	15991	200121	0	1	GAATCTGCTACTAGTGAGCCTTTGTATGGGGACCGAGTTCAGAAGCTCTAACCCTCGTTTTCCCATCTGCTGA
HWI-ST397	0	3	68	15765	200133	0	1	TAGCATGCCTGCTGCAGGAGTTGGTGCCCAGCATTCTCAGGTGTAGTCCAAATTCTGTCTGATACTTATTGTT
HWI-ST397	0	3	68	15810	200133	0	1	TTCAGACAGATGATGCTTGTCAAGGGTCACCATCTTGCATTGCGCTGCGTCACATCCTTAGTGGGAATAGGGC
HWI-ST397	0	3	68	15871	200135	0	1	CTTGCTTCAGCCATGTAGAGTGGTGTTGCTCCTTTACTACCACGAATCATTGGTAACTCCCTGTTCTTATTCAC
HWI-ST397	0	3	68	15974	200136	0	1	TTCAGACAGCCAAACGACGTCTTAGTGGAGAAAATACCTGAGAAAAGTCAAGAAACCAAAACACTAAAAAATG
HWI-ST397	0	3	68	15909	200147	0	1	AGCCTCAGCTTGGTTGCTTGTGGTTGGGGGTGAGGGGGCGGGCGGGAACTTATGTTTGCGCCCCGAGGCGC
HWI-ST397	0	3	68	15946	200152	0	1	CTTGACTGGGCGTGGTGCTGAGGCTACTGCGGAATTGAGGTGTTGTCATCCACCGGATTGGGTCGTAGGGCC
HWI-ST397	0	3	68	15774	200153	0	1	TTCAGACAGCCAACTGAGATGACTCTCATTCTTGGTAGGAACCAATTTCTGAGAGCTTCGTAATGACATCAACT
HWI-ST397	0	3	68	15814	200155	0	1	GAGATACAGCAACAAATGATGTCATTCCTTGCAAAAGCTGTACAAAGCCCTGGTTTCTTAGCTCAGCTGGTAC
HWI-ST397	0	3	68	15850	200154	0	1	GTGTTTGGTCGTGAAAGTGGACCTCTTTCAGGTGCAGGTGCGAGTAGAAGGAGGTCCCAGAGACGTGCGGCT
HWI-ST397	0	3	68	15870	200157	0	1	GAGAAACCGCAGAATGATAGCAAAAAGCGCGTTACAGGAGATATTAAGAAAAGGAGACTTGCAATGCAGGAGT
HWI-ST397	0	3	68	15984	200158	0	1	CGTCAACTGCATGAAGGAGGTTGTCTGGCCGTTGGAGGAGTGATTTTGGAAGGCTGAGATCGGAAGAAAGGT

Assignment to Samples

Barcode sequences from a key file are compared to barcode sequences in the reads, in order to associate reads with the samples from which they originate.

Parameters:

Users supply a plate map and staff members supply DNA barcodes. These are combined into a table of barcodes by sample (barcode key file).

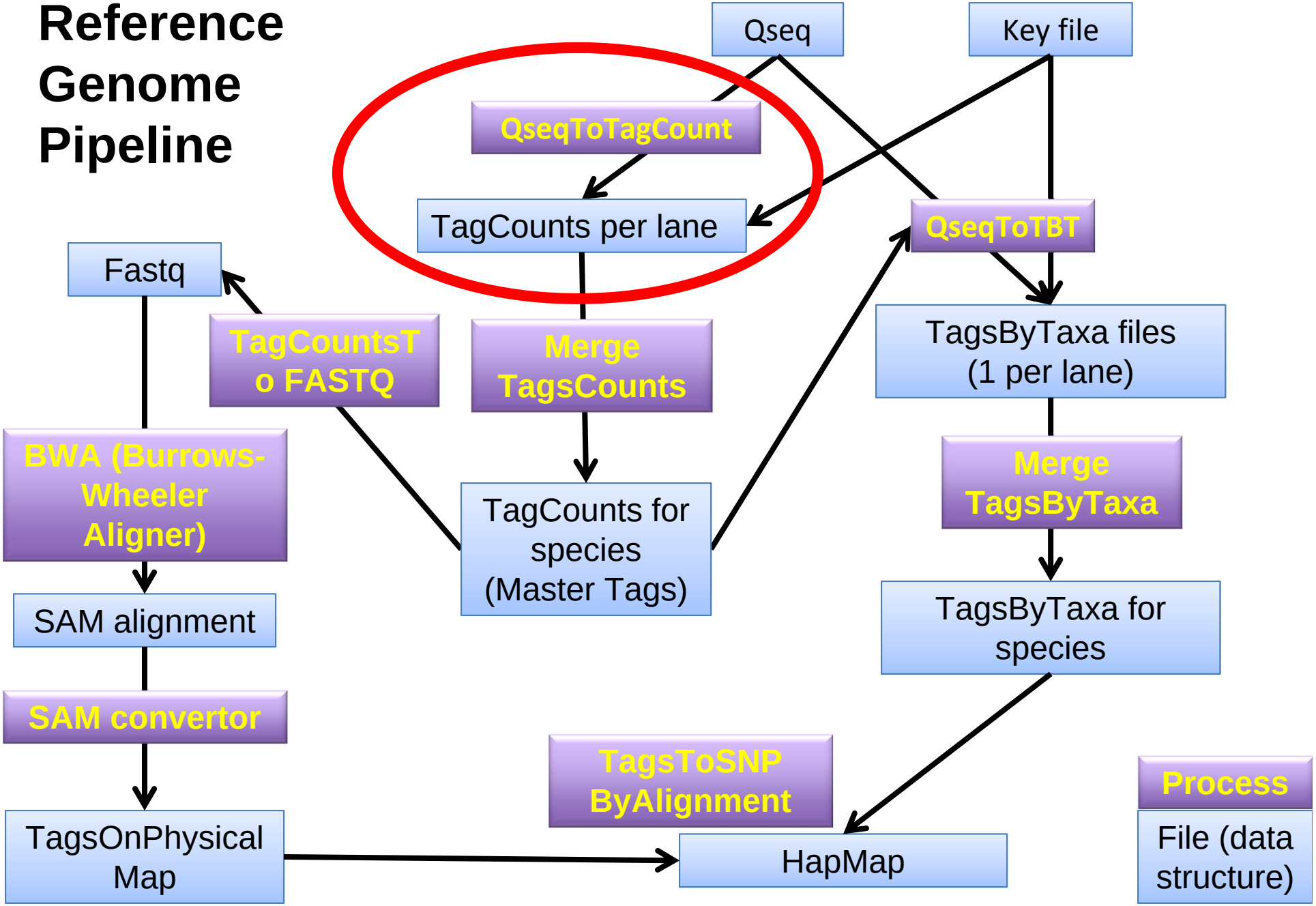
Plate Map

Project Details		Sample Details								
Project Name	Source Lab	Plate Name	Well	Sample Name	Pedigree	Stock Number	DNA Conc	Sample Volume	Sample DNA mass	Prepa
BREAD	CIMMYT	Kassa CIMMYT ONE	A01	KE_Maize1	(PI655994xPI655998)S4	V479-29	32.0	70	2238	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A02	KE_Maize10	(PI655994xPI655998)S4	J221-11	30.0	70	2100	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A03	KE_Maize20	(PI655994xPI655998)S4	V476-211/1	49.8	70	3489	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A04	KE_Maize29	(PI655994xPI655998)S4	V278-101	51.9	70	3630	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A05	KE_Maize40	(PI655994xPI655998)S4	V445-35	56.0	70	3920	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A06	KE_Maize49	(PI655994xPI655998)S4	V547-127	53.2	70	3725	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A07	KE_Maize58	(PI655994xPI655998)S4	V547-40	53.0	70	3710	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A08	KE_Maize68	(PI655994xPI655998)S4	V479-160	50.5	70	3537	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A09	KE_Maize78	(PI655994xPI655998)S4	V547-88	45.1	70	3158	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A10	KE_Maize86	(PI655994xPI655998)S4	V547-10	44.0	70	3080	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A11	KE_Maize94	(PI655994xPI655998)S4	V479-84	98.2	70	6874	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	A12	KE_Maize103	(PI655994xPI655998)S4	V547-65	43.1	70	3016	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B01	KE_Maize2	(PI655994xPI655998)S4	V479-32	53.4	70	3736	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B02	KE_Maize13	(PI655994xPI655998)S4	V445-27	31.1	70	2174	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B03	KE_Maize22	(PI655994xPI655998)S4	V481-75	38.0	70	2660	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B04	KE_Maize30	(PI655994xPI655998)S4	V278-102/1	50.5	70	3534	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B05	KE_Maize41	(PI655994xPI655998)S4	V547-23	51.4	70	3599	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B06	KE_Maize51	(PI655994xPI655998)S4	V366-134	92.0	70	6440	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B07	KE_Maize59	(PI655994xPI655998)S4	V366-47	107.0	70	7490	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B08	KE_Maize69	(PI655994xPI655998)S4	V479-65	111.0	70	7770	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B09	KE_Maize79	(PI655994xPI655998)S4	V547-133	109.0	70	7630	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B10	KE_Maize87	(PI655994xPI655998)S4	V547-174	113.0	70	7910	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B11	KE_Maize95	(PI655994xPI655998)S4	V479-139	82.0	70	5740	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	B12	KE_Maize104	(PI655994xPI655998)S4	V547-179	43.0	70	3010	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	C01	KE_Maize4	(PI655994xPI655998)S4	V479-43	107.0	70	7490	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	C02	KE_Maize14	(PI655994xPI655998)S4	V555-99/1	105.0	70	7350	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	C03	KE_Maize23	(PI655994xPI655998)S4	J184-6	106.0	70	7420	Kassa
BREAD	CIMMYT	Kassa CIMMYT ONE	C04	KE_Maize31	(PI655994xPI655998)S4	V366-51	108.0	70	7560	Kassa

Example DNA Barcode Key

Flowcell	Lane	Barcode	Sample	PlateName	Row	Column
434GFAAXX	2	CTCC	M0001	IBM1	A	1
434GFAAXX	2	TGCA	M0012	IBM1	A	2
434GFAAXX	2	ACTA	M0021	IBM1	A	3
434GFAAXX	2	GTCT	M0029	IBM1	A	4
434GFAAXX	2	GAAT	M0038	IBM1	A	5
434GFAAXX	2	GCGT	M0046	IBM1	A	6
434GFAAXX	2	TGGC	M0057	IBM1	A	7
434GFAAXX	2	CGAT	M0067	IBM1	A	8
434GFAAXX	2	CTTGA	M0080	IBM1	A	9
434GFAAXX	2	TCACC	M0090	IBM1	A	10
434GFAAXX	2	CTAGC	M0099	IBM1	A	11
434GFAAXX	2	ACAAA	M0113	IBM1	A	12
434GFAAXX	2	TTCTC	M0003	IBM1	B	1
434GFAAXX	2	AGCCC	M0013	IBM1	B	2
434GFAAXX	2	GTATT	M0022	IBM1	B	3
434GFAAXX	2	CTGTA	M0030	IBM1	B	4
434GFAAXX	2	AGCAT	M0039	IBM1	B	5
434GFAAXX	2	ACTAT	M0047	IBM1	B	6
434GFAAXX	2	GAGAAT	M0058	IBM1	B	7
434GFAAXX	2	CCAGCT	M0068	IBM1	B	8
434GFAAXX	2	TTCAGA	M0081	IBM1	B	9
434GFAAXX	2	TAGGAA	unknown	IBM1	B	10

Reference Genome Pipeline

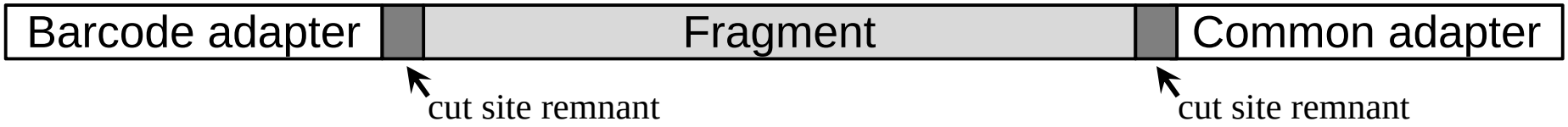


QSeqToTagCount

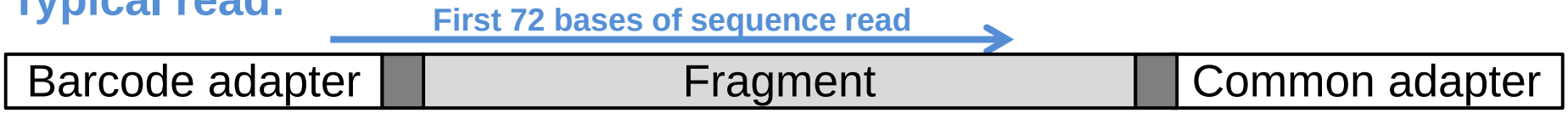
Processes a Qseq file to obtain a list of tags (alleles) present in the sampled individuals

- Ignores reads with N in first 72 bases
- Keeps reads matching one of the barcodes & the cut site remnant
- Trims off the barcode
- Trims reads to 64 bases
- Truncates reads that:
  1. Read into the common adapter (restriction fragments < 64bp)
  2. Have a internal cut site (partial digestion or chimeras)
- Counts how many times each unique 64 base read (tag) occurred in the qseq file

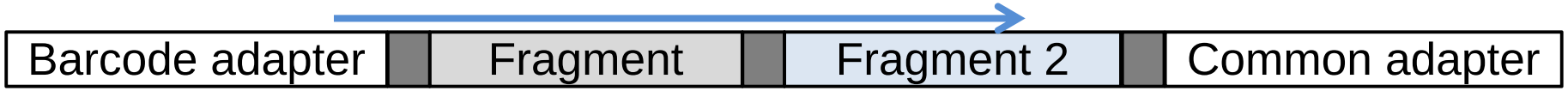
GBS Restriction Fragment Structure (not to scale)



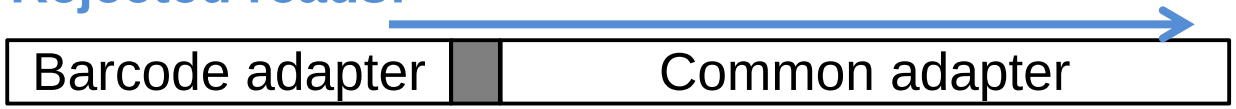
Typical read:



Trimmed reads (trimmed at second cut site):



Rejected reads:



Sequence Processing

Raw sequence data is processed into unique 64-bp sequences.

For example, raw reads:

```
CTCCCAGCCCTCGGCGGTCAAACCACCCGGTCATCCATGCACCAAGGCCTGCGTGCGGGCTTGGTGTCATCGTACGC
GTTGAACAGCCCTCGGCGGTCAAACCACCCGGTCATCCATGCACCAAGGCCTGCGTGCGGGCTTGGTGTCATCGTACGC
```

Become a sequence tag:

```
CAGCCCTCGGCGGTCAAACCACCCGGTCATCCATGCACCAAGGCCTGCGTGCGGGCTTGGTGTCATCGTACGC  64  2
```

Parameters:

Restriction enzyme

Different enzymes have different sequence motifs (remnant cut sites)

Barcodes

Acceptable reads must match one of the barcode sequences.

Minimum count for a tag to be retained

This gives investigators the option to ignore singleton or rare reads.

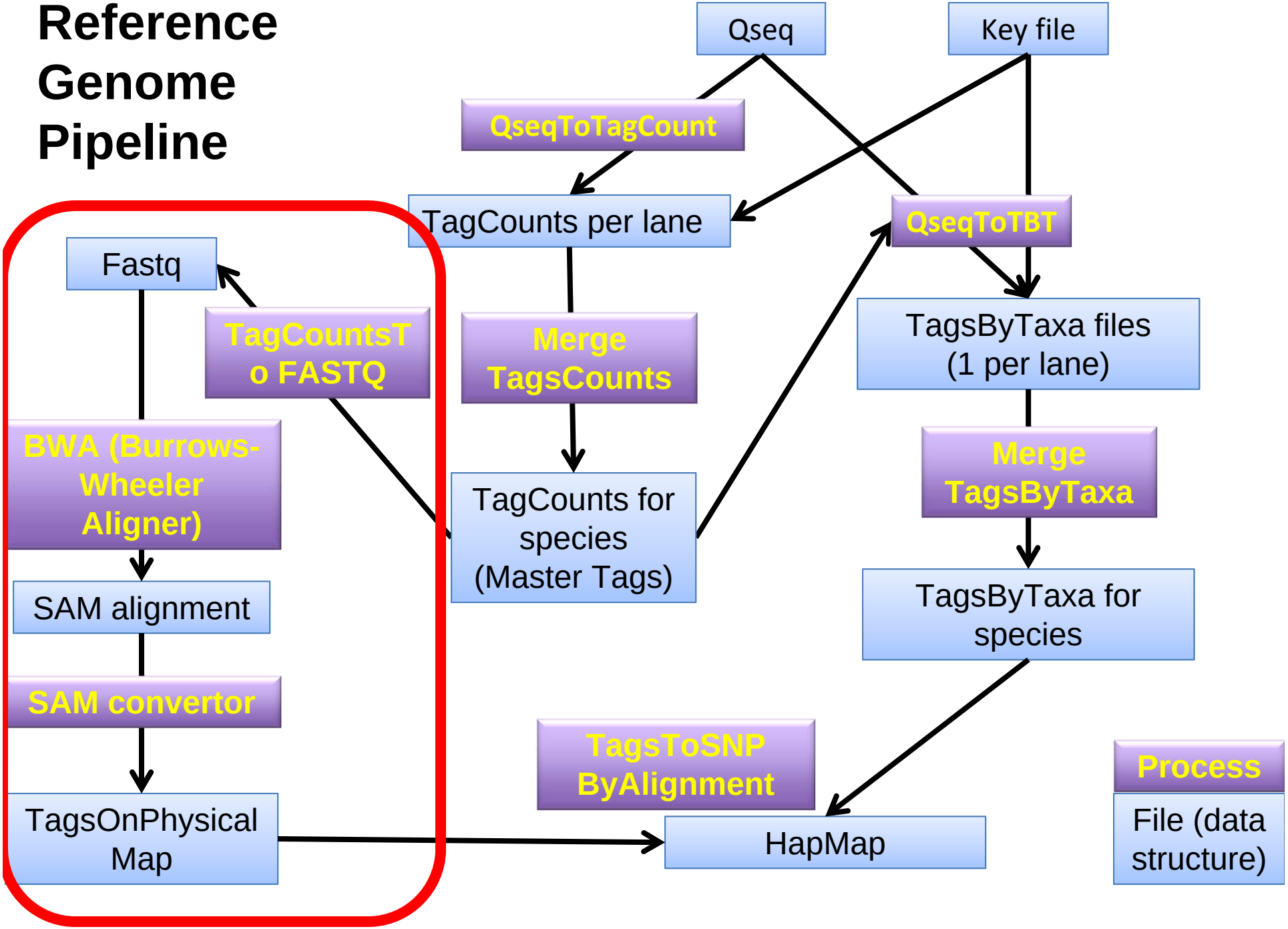
TagCounts File

Number of Tags

Max Size of Tag x 32bp

26442466	2	Tag Sequence	Length (bp)	Count
CAGCAAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGAC			64	1
CAGCAAAAAAAAAAAAAAAAAAAACCAAGAATTTTATGTTTCCTACCTCCAACCCCAGGACTTT			64	1
CAGCAAAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCCTATACCTCATCCCACAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAAAACCAAGTAATTTTATTTCTCATACCTCATACCACAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAAAACCCAAGAAATTTGATGTCTCAAACCCCAACACACAGGCTT			64	1
CAGCAAAAAAAAAAAAAAAAAAAACCCAAGAAATTTTTTGTCTCAAACCCCAACCCCCAGGCCT			64	1
CAGCAAAAAAAAAAAAAAAAAAAAGGGGTTTTGAATAAAAAAAACTGAAGGATCTTAAATCTAC			64	1
CAGCAAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTTTCATACCTCATACCACAGGACT			64	1
CAGCAAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGACT			64	2
CAGCAAAAAAAAAAAAAAAAAAACCAAAAAATTTTATGTCTCAAACCCCAAACCCCAGGGCTTC			64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAATAATTTGATGTCTCATACCTCATACCACAGGGCTTC			64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAGAAATTTGATGTCTCATACCTCATACCACAGGACTTC			64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAGAAATTTTGGCACTCAAGCCCAAAACCACAGATCTTC			64	1
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTGTTGTCTCATACCTCATACCACAGAACTTC			64	1
CAGCAAAAAAAAAAAAAAAAAAACCCAAAAAATTTTTTTTTCCAACCCCAAAACCCAAGGCTTC			64	1
CAGCAAAAAAAAAAAAAAAAAAACCCAAGAAATTTTTTTTCCCAAACCCCAAACCCCAGGCTTT			64	1
CAGCAAAAAAAAAAAAAAAAAAAGGGATAGGGAAGATGGGGGAGAGTGGCGGCCACGCATGGAA			64	1
CAGCAAAAAAAAAAAAAAAAAACAACAAGGAATTTGGGTATTCATTCCCCATACCCCAGGATTT			64	1
CAGCAAAAAAAAAAAAAAAAAACACAAAAAAATTTGTTTTCTCAACCCCAAAACCAAAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCAAAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCACAGGACTT			64	2
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCCCAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAACACCAAGGAATTGAATCTCTCACACCTTAAAACACCGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATTTCTCATACCTCATACCAAAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAACACCAATTATTTGAAAGATCATTACCCTATACCACGGGGTTC			64	1
CAGCAAAAAAAAAAAAAAAAAACCAAAAAAATTTGATGTCTCATACCCCATACCACAGGACTCCC			64	1
CAGCAAAAAAAAAAAAAAAAAACCAAAAAATTTTATTTCTCATACCCCAAACCCCAGGACTTCC			64	1
CAGCAAAAAAAAAAAAAAAAAACCAAAGAATTTTATGTCTCATACCTCAAACCAAAGGACTTCC			64	1
CAGCAAAAAAAAAAAAAAAAAACCAAATAAATTTGTTGCTCATACCCCAAACCACAGGGCTTTC			64	1
CAGCAAAAAAAAAAAAAAAAAACCAAGCAATTTGATTCCACTTAATCTATCCCACAGAACTTCC			64	1
CAGCAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCTCATACCTCATACCACAGGACTTCC			64	1
CAGCAAAAAAAAAAAAAAAAAACCCAAAAAATTTTTTGTTTCCCTAACCCCAAAACCACGGACT			64	1
CAGCAAAAAAAAAAAAAAAAAACCCAATGAATTTGTAGTGCCAAACCCCAAACCAACGGACTTT			64	1
CAGCAAAAAAAAAAAAAAAAAACCCCAAGAAATTTGATGTCTCATACCCCAAACCCCAGGACTT			64	1
CAGCAAAAAAAAAAAAAAAAAAGACCAGGTAATTATTGCTCACATACATCAAACTCCAATTGCC			64	1
CAGCAAAAAAAAAAAAAAAAAAGCGCCTAACGTTTCAAAATGAATGAGTTGCCAACCAAGGACT			64	1
CAGCAAAAAAAAAAAAAAAAAAGGGTTAGGAAAGATGGGTGGGAGGGGCGGGCCTGCTTGAAAT			64	1

Reference Genome Pipeline



Unique Reads (FASTQ)

```
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGAC
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAAACCAAGAATTTTATGTTTCCTACCTCCAACCCCAGGACTTT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTGATGTCCTATACCTCATCCCACAGGACTT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAAACCAAGTAATTTTATTTCTCATACCTCATACCACAGGACTT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAAACCCAAGAAATTTGATGTCTCAAACCCCAACACACAGGCTT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAAACCCAAGAAATTTTTTGTCTCAAACCCCAACCCCCAGGCCT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAAAGGGGTTTTGAATAAAAAAAACTGAAGGATCTTAAATCTAC
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTTTCATACCTCATACCACAGGACT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=2
CAGCAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGACT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
CAGCAAAAAAAAAAAAAAAAAACCAAAAAATTTTATGTCTCAAACCCCAAACCCCAGGGCTTC
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
@length=64count=1
```

BWA (Burrows-Wheeler Aligner)

- **Aligns the tags in FASTA format to the reference genome**
- **Parameters:**
  - **Similarity of read sequence and genome sequence. This controls the tradeoff between number of SNPs and confidence in the alignment. Default is 4 edits per sequence.**
  - **Gap penalty. This controls sensitivity to indels. Default is no indels within 5bp of the read ends.**
- **Outputs a SAM Alignment**
- **There are many other aligners. BWA is fast and memory efficient, but may not be appropriate for your species**

Generic Alignment (SAM)

length=64count=1	0	7	6994125	37	55M2I7M	*	0	0	CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCA
length=64count=2	0	7	6994125	37	54M2I8M	*	0	0	CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCA
length=64count=1	0	7	6994125	37	53M2I9M	*	0	0	CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCC
length=64count=1	0	7	6994125	37	54M2I8M	*	0	0	CAGCAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCA
length=64count=1	0	7	6994125	37	55M2I7M	*	0	0	CAGCAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATTTCTCATACCTCATACCA
length=64count=4	0	7	6994125	37	4M3D47M2I11M	*	0	0	CAGCAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAG
length=64count=1	16	17	14761759	25	64M	*	0	0	CCTTTCTTGGCCTGGTTCTCACTCATCTGGGCTTGGAATTGAGAACCGTTTTTTTT
length=64count=7	16	18	1517944	25	64M	*	0	0	GCCCGTCTACACGCTTGTGTCCCATGCCCGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=1	16	18	1517944	25	64M	*	0	0	GCCCGTCTACACGTTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=1	16	18	1517944	25	64M	*	0	0	GCCCGTCTACAGGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=4	16	18	1517944	25	64M	*	0	0	GCCCGTCTACCCGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=2	16	18	1517944	25	64M	*	0	0	GCCCGTCTCCACGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=53	16	18	1517944	37	64M	*	0	0	GCCCGTCTACACGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=1	16	18	1517944	25	64M	*	0	0	CCCCGTCTACACGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=1	16	18	1517944	25	64M	*	0	0	GCCCGTCTACACCCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTTT
length=64count=1	0	10	10388735	37	58M1I5M	*	0	0	CAGCAAAAAAAAAAAAATAGAACTTAGAAACTTATACCGTGGGACACGTCAAGTGAC
length=64count=1	0	2	714861	37	64M	*	0	0	CAGCAAAAAAAAAAACCAAAGATCGACTTGCAACATCTGGATGGAAACAACAAACA
length=64count=11	16	19	13463035	37	49M1I14M	*	0	0	TGCCCGTCTACACGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTTT
length=58count=1	0	2	14032437	37	4M1I59M	*	0	0	CAGCAAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGC
length=64count=1	0	2	14032437	37	4M1I59M	*	0	0	CAGCAAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGC
length=64count=1	16	19	13463036	37	48M2I14M	*	0	0	GCCCGTCTACACGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCCTTTTTTT
length=64count=1	0	6	20542400	37	64M	*	0	0	CAGCAAAAAAAAAAATCCTCTCCTCATACGCTCCTCCCAGCTTGCACTAACGGCCA
length=64count=1	16	5	15019027	37	49M1I14M	*	0	0	CCCATTGTTGTATCTTGATTGCAGACTCCCCCTCATCACTCTTTCTGCATTTTTTT
length=64count=3	16	5	15019027	37	49M1I14M	*	0	0	ACCATTGTTGTATCTTGATTGCAGACTCACCCTCATCACTCTTTCTGCATTTTTTT
length=64count=1	16	5	15019027	37	49M1I14M	*	0	0	CCCATTGTTGTATCTTGATTGCAGACTCACCCCCATCACTCTTTCTGCATTTTTTT
length=64count=1	16	5	15019027	37	49M1I14M	*	0	0	ACCATTGTTGTATCTTGATTGCAGACTCTGCCTCATCACTCGTTCTGCATTTTTTT
length=64count=1	0	6	20542400	37	4M1I59M	*	0	0	CAGCAAAAAAAAAACATCCTCTCCTCATACGCTCCTCCCAGCTTGCACTAACGGCC
length=64count=1	0	8	18851188	37	64M	*	0	0	CAGCAAAAAAAAAAGAGAGGCCTAAAAAGGGTAATGAAGGCAAAAGTGCCCTTCTT
length=64count=5	16	19	13463034	23	64M	*	0	0	CTGCCCGTCTACACGCTTGTGTCCCATGCACGCAAGCCGCCCCATCCCTCTTTTTT
length=64count=1	0	5	6176480	37	64M	*	0	0	CAGCAAAAAAAAAAGCCCAATCTAGACCCTATCTTCTAATAGCGAATAAGAAAAGG
length=64count=7	0	5	6176480	37	64M	*	0	0	CAGCAAAAAAAAAAGCCCAATCTAGAGCCTATCTTCTAATAGCGAATAAGAAAAGG
length=57count=31	0	2	14032437	25	64M	*	0	0	CAGCAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGCT
length=64count=4	0	2	14032437	25	64M	*	0	0	CAGCAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGCT
length=64count=1	0	2	14032437	25	64M	*	0	0	CAGCAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGTT
length=64count=1	16	5	15019027	37	16M1I47M	*	0	0	TCCATTGTTGTATCTTCGATTGCAGACTCACCCTCATCACTCTTTCCGCCTTTTTT

SAMConverter & TagsOnPhysicalMap (TOPM)

- **TOPM file contains all the information needed to interpret tags present in a species:**
  - **Tag Sequence**
  - **Physical position (chromosome, position & strand)**
  - **Divergence from reference**
  - **Polymorphisms (up to 4 variant alleles per tag)**
  - **Genetic mapping support (not implemented yet)**

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

1020631	2	4					
CAGAGATAAACGCAAGCAGAAGCAGGAGGAGGGAGGCAAAGATGACAAGGCCATGGGCAGTGAC	64	1	11		35	28230922	0
CAGAGATAAACTAAAGCATCGATCCATCAAAATGAATTATGTAATTGCAAAAGTCAATTATGCC	64	1	11		11	3832148	0
CAGAGATAAAGAGAGAGAAAAACTCAAAAAGAAAAGCATATCATGCAGGCGACGCAGAGAGATA	64	1	6	-1	27993074	27993011	0
CAGAGATAAATCCAAACGTTTTCTGTTCGGTAGGGTCTGTCAGCGAGGCAGAAAAAAAAAAAAA	51	1	7	1	3250207	3250257	0
CAGAGATAAATTTGAACGGTAGTTAACTAGCAAATTCGTCTATATTAAATTCAATAGTATGACA	64	1	2	1	34487829	34487892	0
CAGAGATAACTAGTTAATTAGTTGCATGCAAGTGTGTGCTCTTGTGCTATTGATTAAGCTGGTA	64	1	2	1	1249312	1249375	0
CAGAGATAACTTATATATTTGCATCCAAAGATAATCCAGCCGACAGCCCGACATATGTAAGACG	64	1	5	-1	24463120	24463057	0
CAGAGATAAGCTTTAGGGCACGAATCATATCTTCAACAGCATCCTCTATAGCACGCAATGGCAG	64	1	3	-1	27908515	27908452	0
CAGAGATAAGGATACCTACTGTGGATAAATACACAGAATATACACGTCGACGTGACCGGACATC	64	1	1	1	41245835	41245898	0
CAGAGATAATAGTAGGATTCTGCGCAATGCAATCGAGTTTCACGCCTCTGTTCTTCAGTTCTTG	64	1	6	-1	28001431	28001	0
CAGAGATAATATCATCAGTACCTTCTTTATTTCCCAGAACGCCATCACCATCAGCTGAAAAAAA	57	1	6	-1	80		2
CAGAGATAATGCTCTTACACGGTACAAGGAGAAGAAGAAGAGAAGAAAGTAAGTGCATCTTACT	64	2	*	*	*	*	0
CAGAGATAATTAAATATGTTTAGTTTTGACAGGTGAAGGAACAATAGTTAGCTTATCTGTCATG	64	1	1	1			0
CAGAGATAATTGACAATG CGTAGGCATATTGACAACTTCAAAAATTATGCACAAAAATGTCC A	64						

Max #mismatches

Divergence

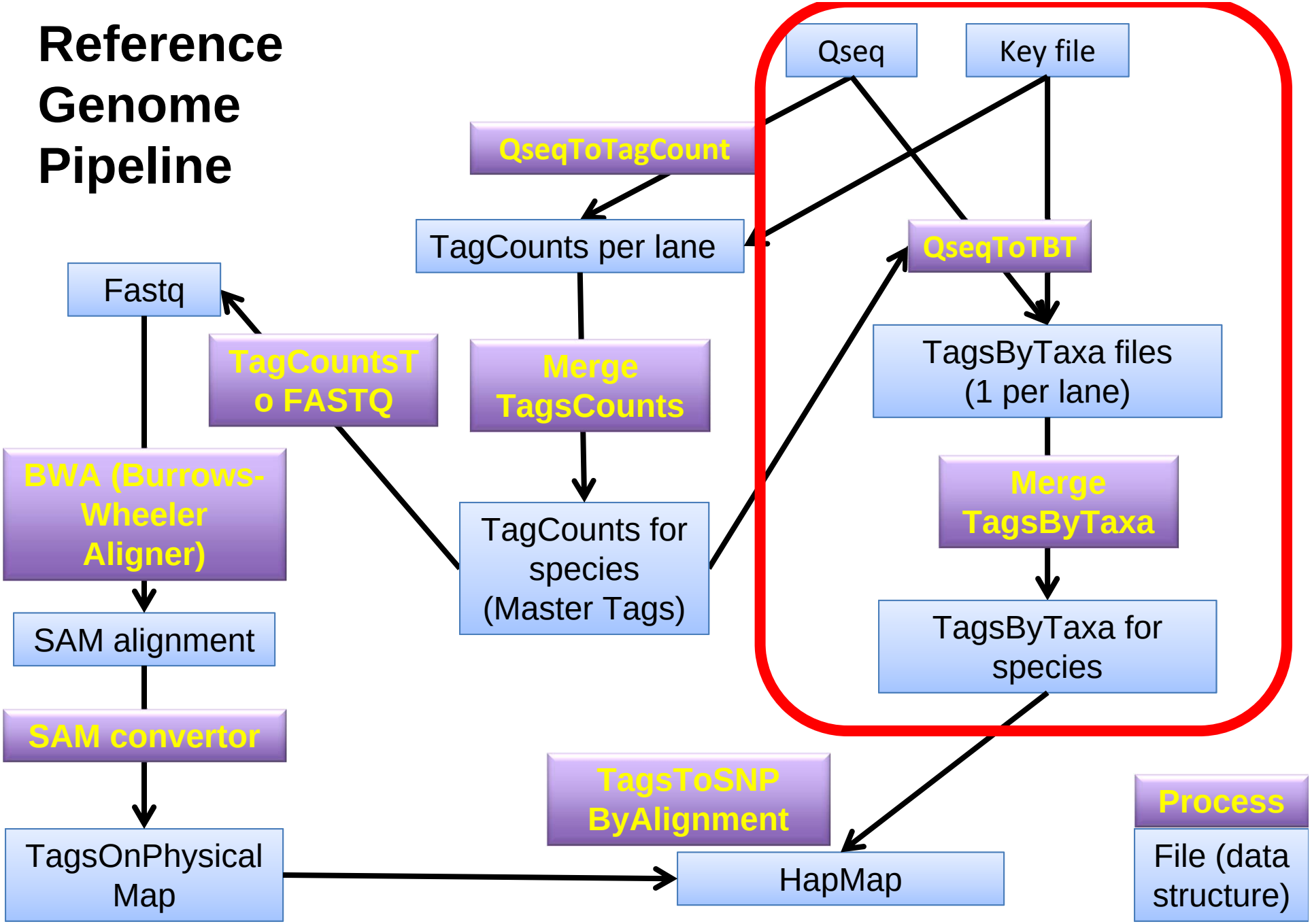
Tags

Tag size

Multi-maps

Chr, strand, start and end positions

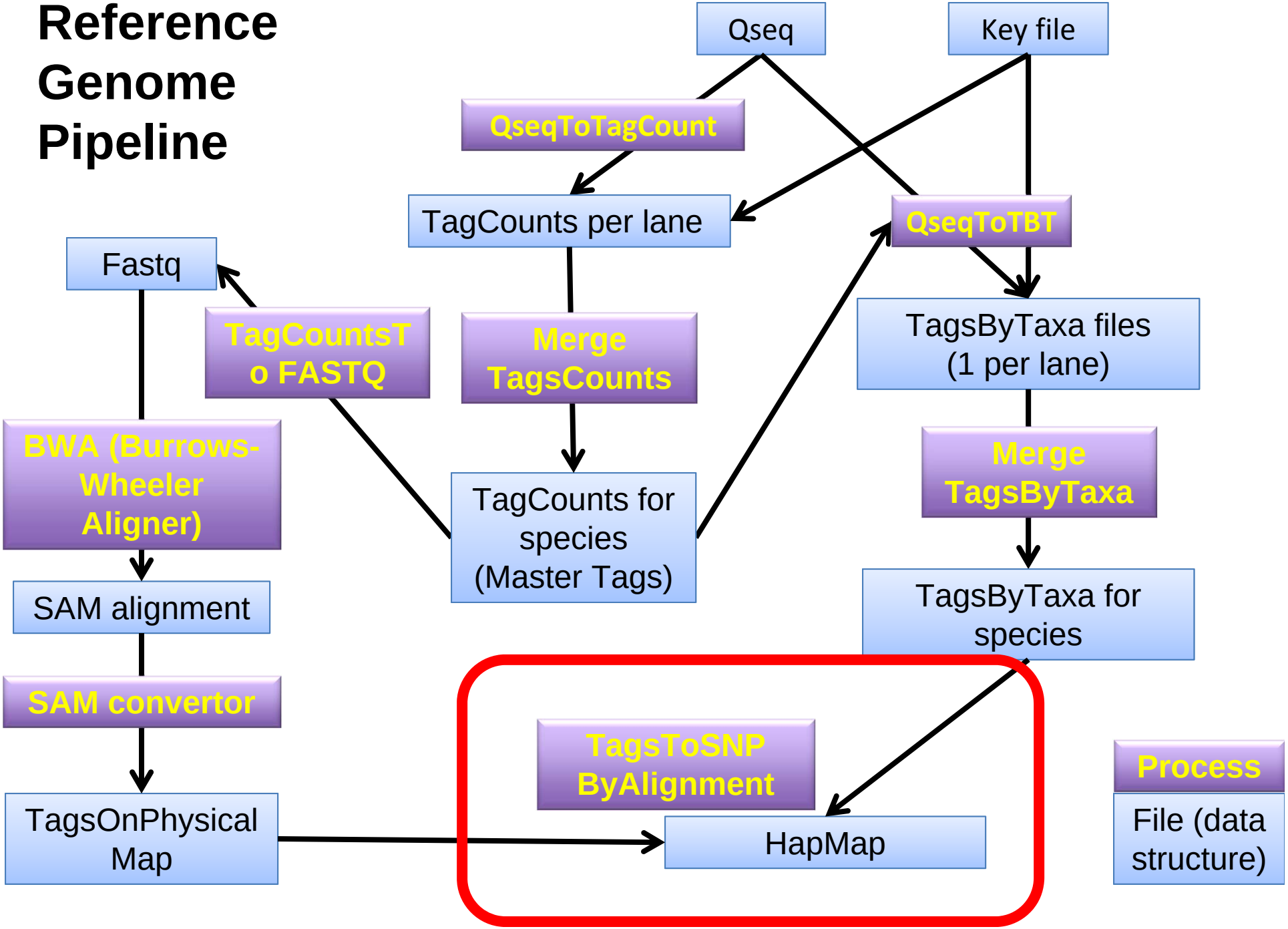
Reference Genome Pipeline



Tags by Taxa

6040401 2	88								
08.0731-5	chardonnay	08.0731-19	08.0731-29	08.0731-6	08.0731-24	08.0731-37	08.0731-15	08	
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCAAAGGACTT				64	0	0	1	0	0	0	0
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCACAGGACTT				64	0	1	0	0	0	0	0
CAGCAAAAAAAAAAAAAAAAAACACCAAGAAATTTGATGTCTCATACCTCATACCCCAGGACTT				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGACTT				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAAAAAAACACCAAGTAATTTGATTTCTCATACCTCATACCAAAGGACTT				64	1	0	0	0	0	0	0
CAGCAAAAAAAAAAAAAAAACACCAAGTAATTTGATGTCTCATACCTCATACCACAGGACTTCCC				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAACGGTTCTCAATTCCAAGCCCAGATGAGTGAGAACCAGGCCAAGAAAGG				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGGGCATGGGACACAAGCGTGTAGACGGGC				64	0	1	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAACGTGTAGACGGGC				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCCTGTAGACGGGC				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGGGTAGACGGGC				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTGGAGACGGGC				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTGTAGACGGGC				64	0	1	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTGTAGACGGGG				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGGGTGTAGACGGGC				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAATAGAACTTAGAAACTTATACCGTGGGACACGTCAAGTGACTGCTGATG				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAACCAAAGATCGACTTGCAACATCTGGATGGAAACAACAAACAAACAAAGA				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTGTAGACGGGCA				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGCTGAAAAAA				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGCTGGAGGGA				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAAGGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTGTAGACGGGC				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAATCCTCTCCTCATACGCTCCTCCCAGCTTGCACTAACGGCCAACAGATTT				64	0	0	1	0	0	0	0
CAGCAAAAAAAAAAATGCAGAAAGAGTGATGAGGGGGAGTCTGCAATCAAGATACAACAATGGG				64	0	1	0	0	0	0	0
CAGCAAAAAAAAAAATGCAGAAAGAGTGATGAGGGTGAGTCTGCAATCAAGATACAACAATGGT				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAATGCAGAAAGAGTGATGGGGGTGAGTCTGCAATCAAGATACAACAATGGG				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAATGCAGAACGAGTGATGAGGCAGAGTCTGCAATCAAGATACAACAATGGT				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAACATCCTCTCCTCATACGCTCCTCCCAGCTTGCACTAACGGCCAACAGATT				64	0	1	0	0	0	0	0
CAGCAAAAAAAAAAGAGAGGCCTAAAAAGGGTAATGAAGGCAAAAGTGCCCTTCTTAGCTGTAG				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAGAGGGATGGGGCGGCTTGCGTGCATGGGACACAAGCGTGTAGACGGGCAG				64	0	0	0	0	0	0	0
CAGCAAAAAAAAAAGCCCAATCTAGACCCTATCTTCTAATAGCGAATAAGAAAAGGCCCCATCC				64	0	0	0	0	1	0	0
CAGCAAAAAAAAAAGCCCAATCTAGAGCCTATCTTCTAATAGCGAATAAGAAAAGGCCCCATCC				64	0	1	0	0	1	0	0
CAGCAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGCTGAAAAAAA				64	1	1	0	0	1	0	0
CAGCAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGCTGGAGAGAT				64	1	0	0	0	0	0	0
CAGCAAAAAAAAAAGCTATGAACCATCGGGGGAGAGGTGAGAAATGTTGATTGGTTGCAGAGAA				64	0	0	0	0	1	0	0

Reference Genome Pipeline



TagsToSNPByAlignment

Tags that align to the same chromosomal position & strand are aligned against one another

SNPs and small indels are identified from each tag alignment

The SNP genotypes are recorded for each DNA sample in a HapMap format genotype file (one file per chromosome)

Parameters:

- **Minimum minor allele frequency**
- **Min. locus coverage (proportion of Taxa with a genotype)**
- **Minimum F (inbreeding coefficient) (off by default)**
- **Allow 3rd or 4th alleles at a SNP (off by default)**
- **Call indels (as single base gaps) (off by default)**

HapMap Format

rs#	alleles	chrom	pos	strand	DNA1:633Y5AAXX:2:C9			DNA2:633Y5AAX	
S1_2100	A/G	1	2100	+	N	N	N	N	N
S1_2163	T/C	1	2163	+	N	N	N	N	N
S1_13837	T/G	1	13837	+	N	N	N	N	N
S1_14606	C/T	1	14606	+	N	N	C	N	N
S1_20601	T/A	1	20601	+	T	N	N	N	N
S1_68332	C/T	1	68332	+	N	N	N	N	N
S1_68596	A/T	1	68596	+	A	N	N	N	N
S1_69309	G/A	1	69309	+	N	G	N	N	N
S1_79955	T/G	1	79955	+	N	T	G	T	T
S1_79961	T/G	1	79961	+	N	T	T	T	T
S1_80647	C/T	1	80647	+	N	N	N	N	N
S1_81274	T/G	1	81274	+	N	N	N	N	N
S1_108834	G/A	1	108834	+	N	N	N	N	N
S1_112345	T/G	1	112345	+	N	N	N	N	N
S1_115359	C/T	1	115359	+	N	N	N	N	N
S1_115362	T/C	1	115362	+	N	N	N	N	N
S1_115405	G/A	1	115405	+	G	G	A	N	N
S1_115516	T/G	1	115516	+	N	N	T	N	N
S1_116694	A/G	1	116694	+	N	A	G	N	N
S1_119016	C/T	1	119016	+	N	N	N	N	C
S1_155366	T/C	1	155366	+	N	T	N	N	N

Why can GBS be complicated? Tools for filtering, error correction and imputation.

Edward Buckler

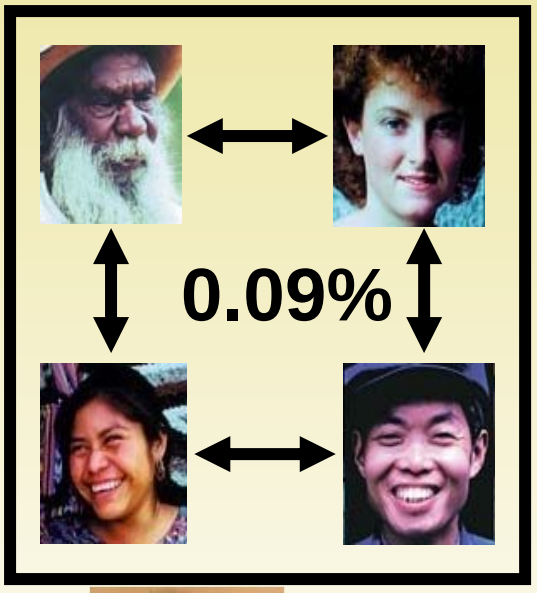
USDA-ARS

Cornell University

http://www.maizegenetics.net

Maize has more molecular diversity than humans and apes combined

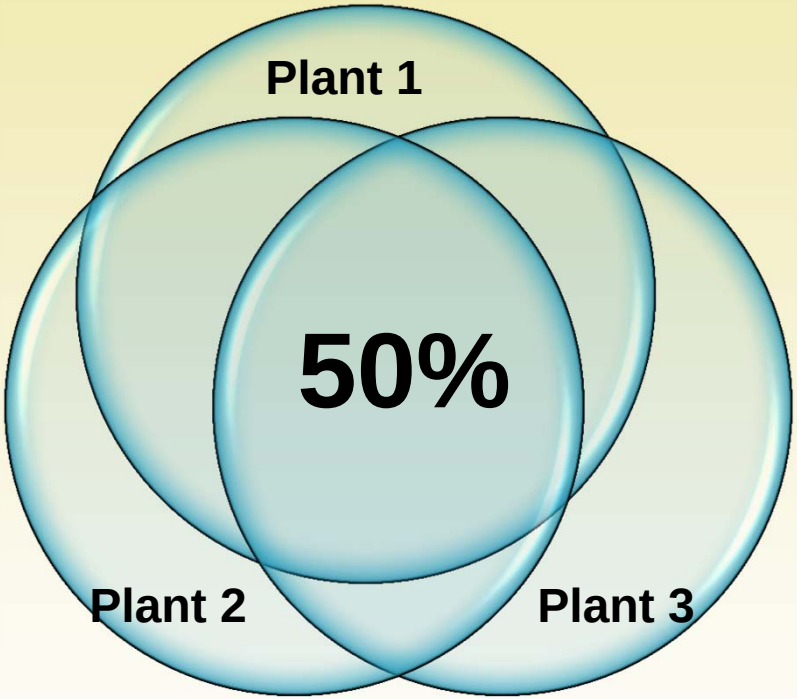
1.34%



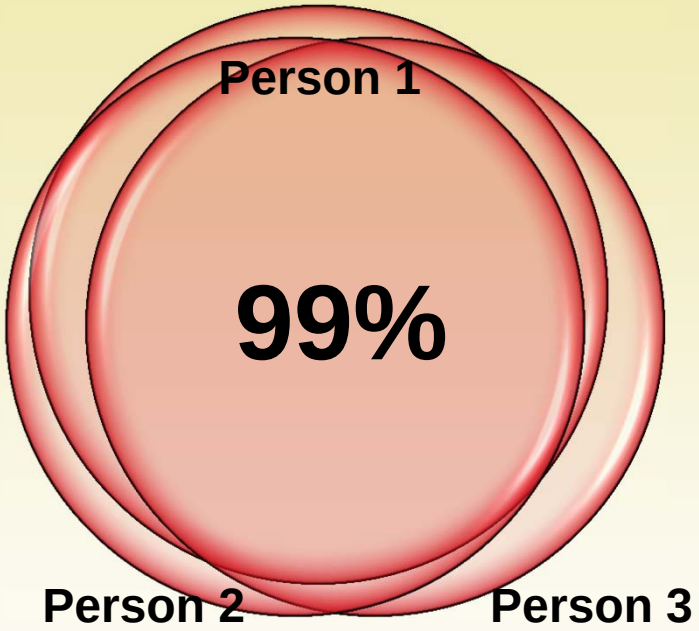
1.42%

Silent Diversity (Zhao PNAS 2000; Tenallion et al, PNAS 2001)

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



Maize

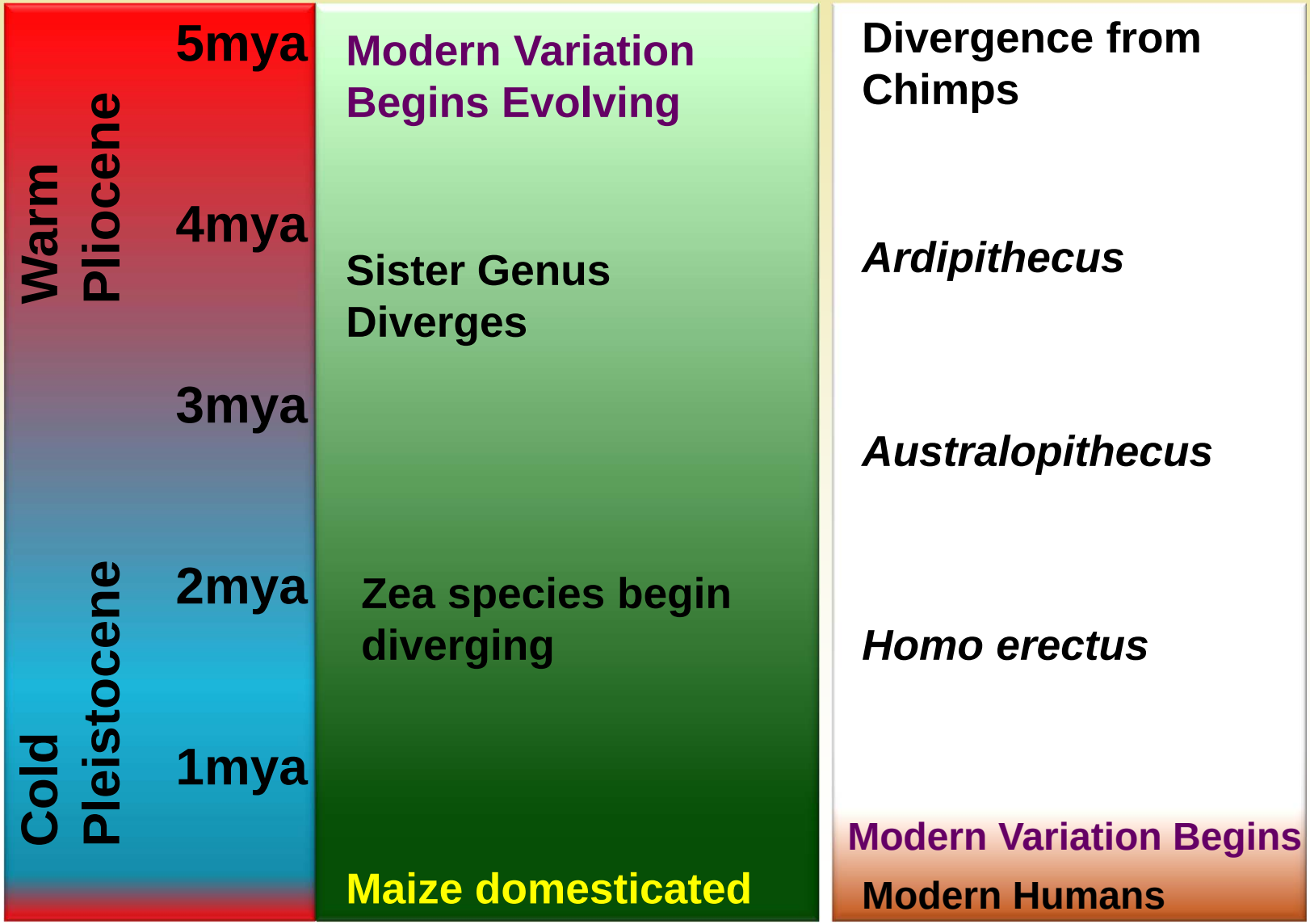


Humans

Fu & Dooner 2002, Morgante et al. 2005, Brunner et al 2005

Numerous PAVs and CNVs - Springer, Lai, Schnable in 2010

Maize genetic variation has been evolving for 5 million years

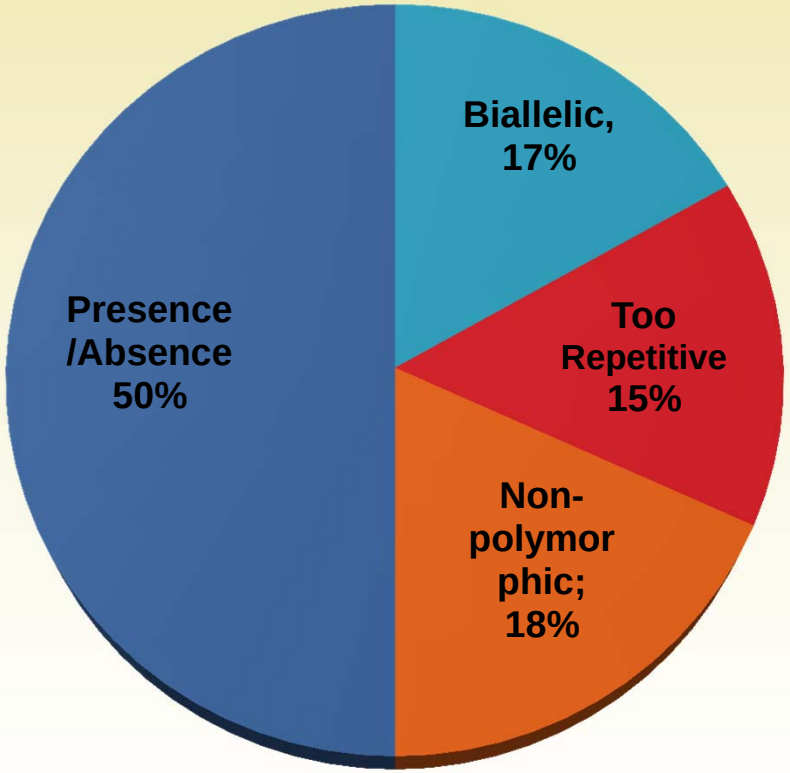


What are our expectations with GBS?

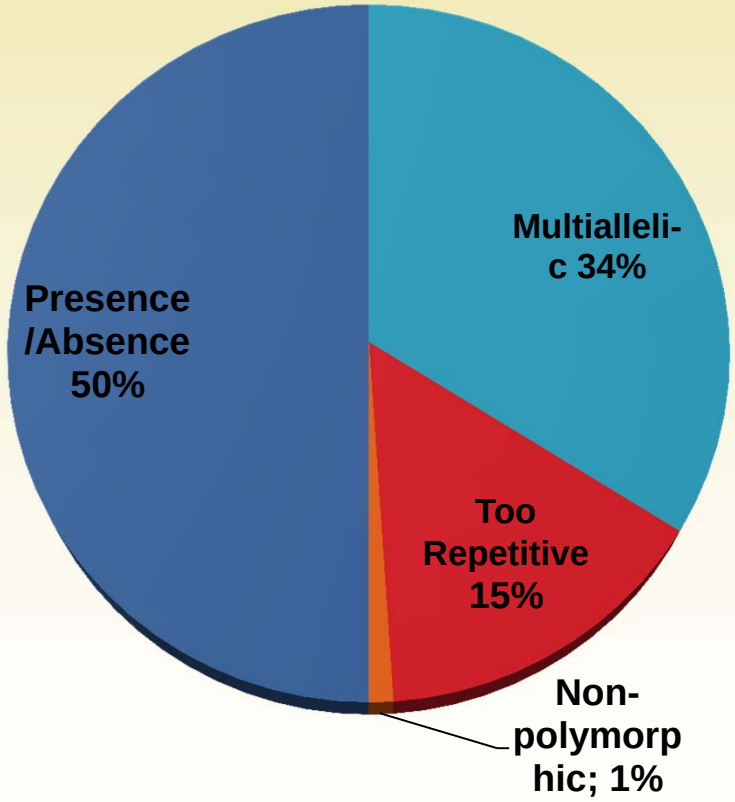
High Diversity Ensures High Return on Sequencing

- **Proportion of informative markers**
  - **Highly repetitive – 15% not easily informative**
  - **Half the genome is not shared between two maize line**
    - **Potentially all of these are informative with a large enough database**
  - **Low copy shared proportion (1% diversity)**
    - **Bi-parental information = (1-0.01)^64bp = 48% informative**
    - **Association information = (1-0.05)^64bp= 97% informative**

Expectation of marker distribution



Biparental population



Across the species

Sequencing Error

Illumina Basic Error Rate is ~1%

- Error rates are associated with distance from start of sequence
  - Bad – GBS puts these all at the same position
  - Good – Reverse reads can correct
  - Good – Error are consistent and modelable

Reads with errors

- **Perfect sequences:**

  0.99^{64}=52.5% of the 64bp sequences are perfect

  47.5 are NOT perfect

  The errors are autocorrelated so the proportion of perfect sequence is a little higher, and those with 2 or more is also higher.

Do we see these errors?

- **Assume 10,000 lines genotyped at 0.5X coverage**

Base	Type	Read # (no SNP)	Read # (w/ SNP)
A	Major	4950	4900
C	Minor	17	67 (50 real)
G	Error	17	17
T	Error	17	17

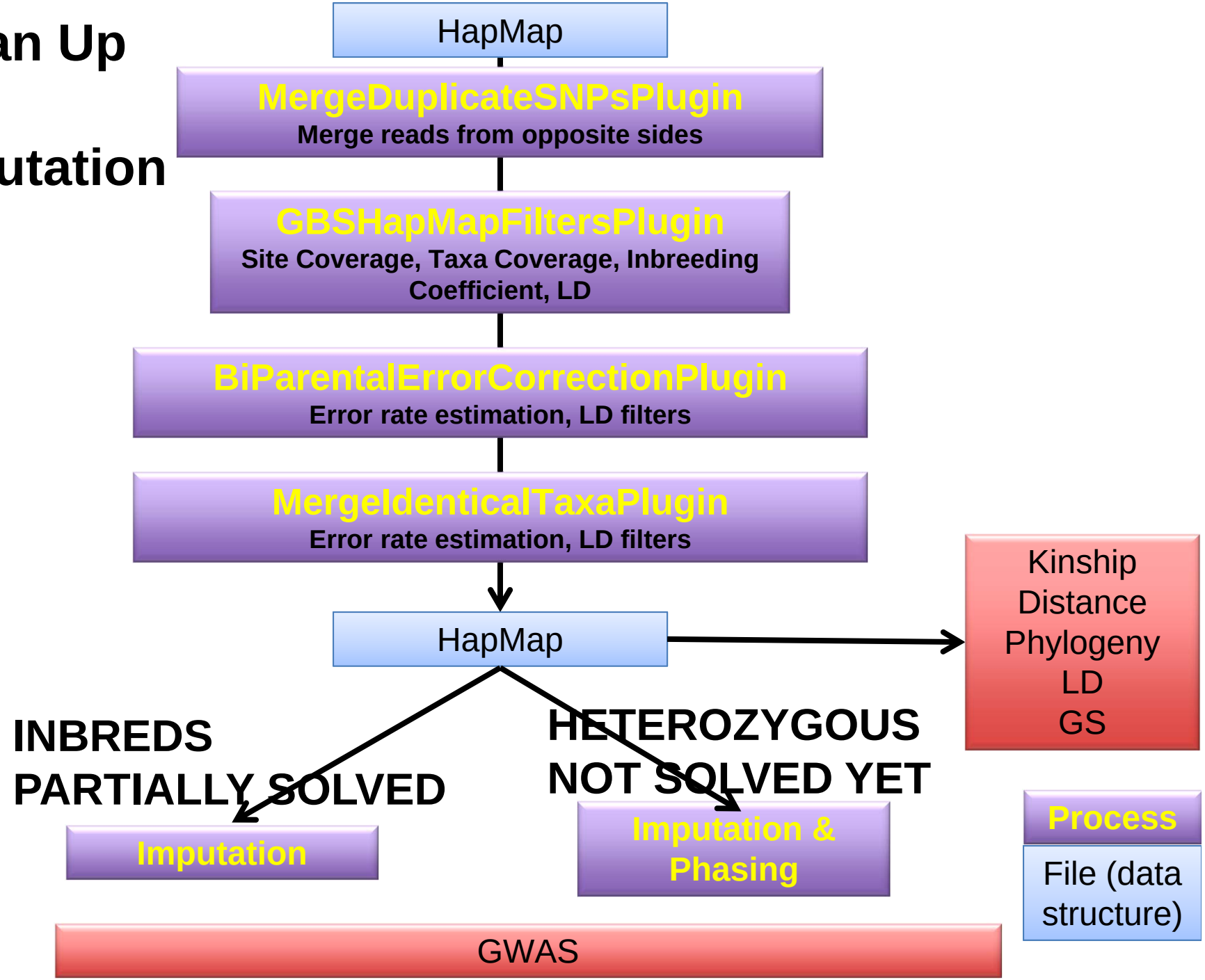
Do Errors Matter?

- **Yes –Imputation, Haplotype reconstruction**
- **Maybe – GWAS for low frequency SNPs**
- **No – GS, genetic distance, mapping on biparental populations**

Expectations of Real SNPs

- **Vast majority are biallelic**
- **Homozygosity is predicted by inbreeding coefficient**
- **Allele frequency is constrained in structured populations**
- **In linkage disequilibrium with neighboring SNPs**

Clean Up and Imputation



Filters in TagsToSNPByAlignmentMTPlugin

- Only calls bi-allelic (hard coded now)
  - Two most common alleles used
- Inbreeding coefficient (-mnF)
  - If have inbred samples definitely use, very powerful for errors and paralogues
- Minimum minor allele frequency (-mnMAF)
  - Very important if do not have other tools for filtering (bi-parental populations or LD)
  - Set for >=1% if no other filter method present

MergeDuplicateSNPsPlugin

- When restriction sites are less than 128bp apart, we may read SNP from both directions (strands)
- ~13% of all sites
- Fusing increases coverage
- Fixes errors
- -misMat = set maximum mismatch rate
- -callHets = mismatch set to hets or not

GBSHapMapFiltersPlugin

- **Basic filters for coverage of sites, taxa inbreeding coefficient, and LD**
- **-mnTCov = minimum taxa coverage (e.g.0.05)**
- **-mnSCov = minimum site coverage, proportion of taxa with call (e.g. 0.10)**
- **-mnMAF = minimum minor allele frequency (e.g. 0.01)**

GBSHapMapFiltersPlugin

- **-mnF = minimum inbreeding coefficient (e.g. 0.9) – Don’t use with outcrossers**
- **-hLD = require that sites are in high local LD, currently parameters are hard coded, so difficult to tune without using the code.**
  - **Tests a sliding window of 100 surrounding sites, and looks for a Bonferonni corrected P<0.01**
  - **Useful but can be slow option.**
  - **More work needed here.**

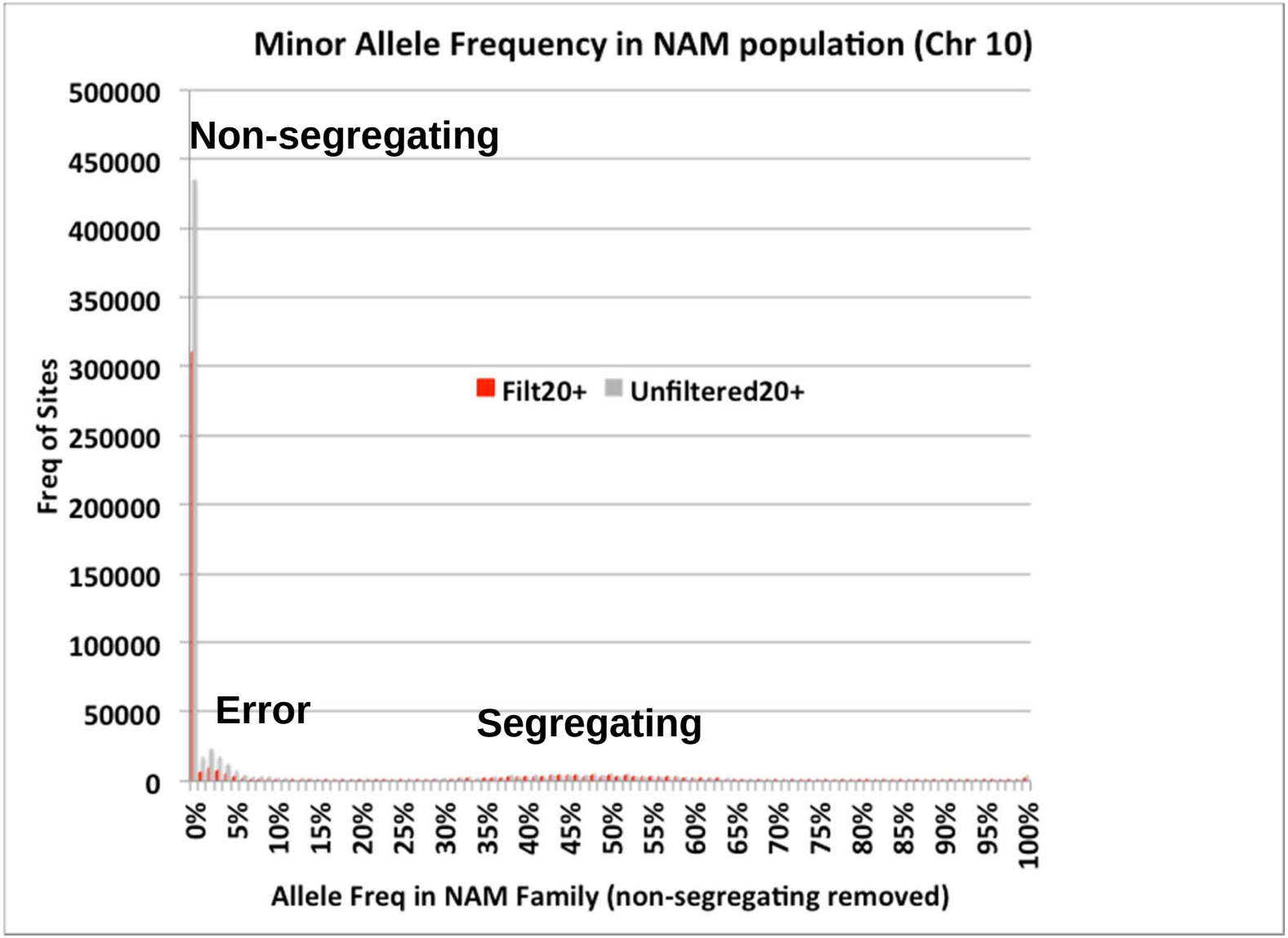
Biparental populations

Limited range of alleles, expected allele frequencies, high LD

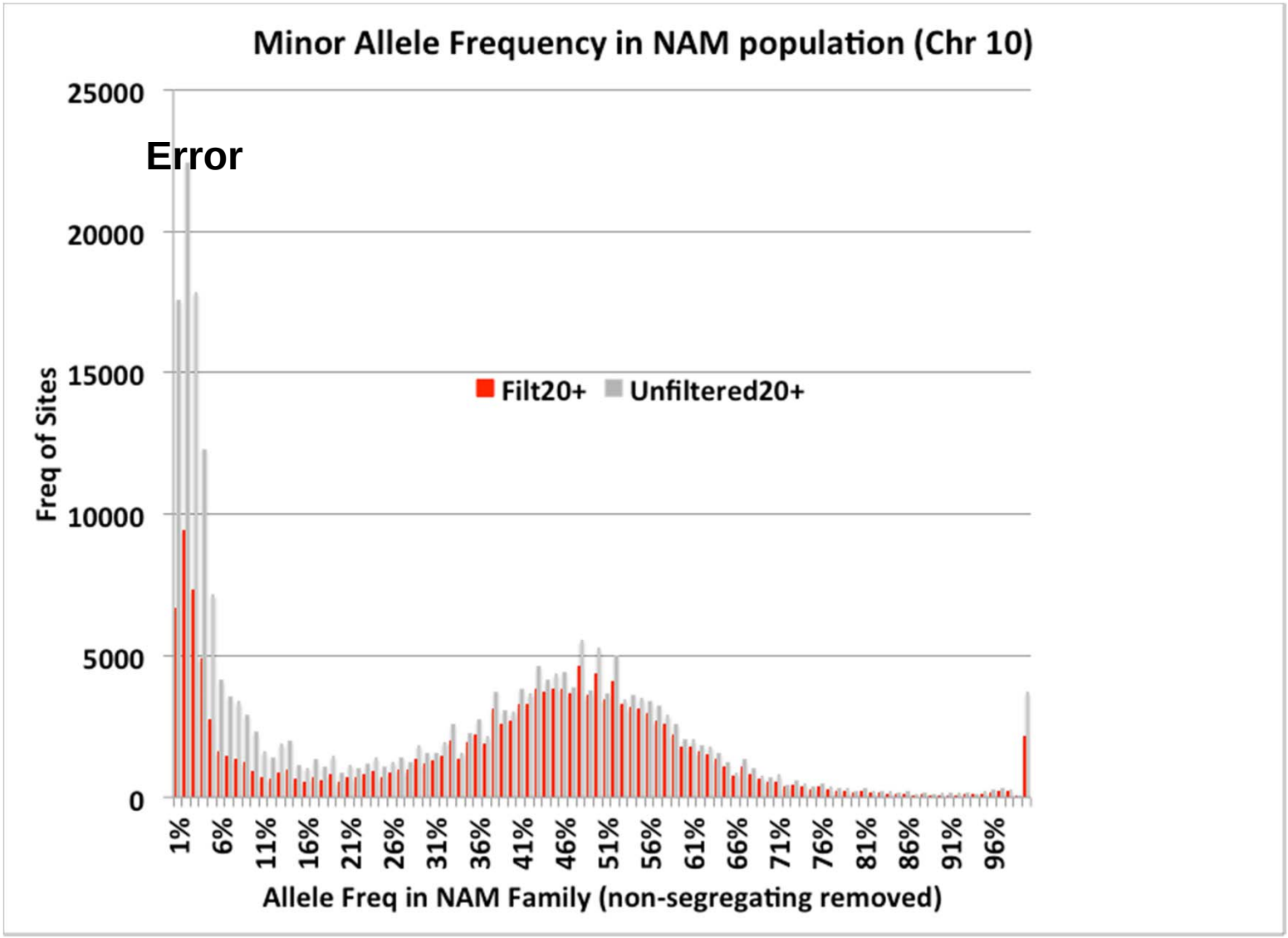
Maize RIL population expectations

- **Allele frequency 0% or 50%**
- **Nearby sites should be in very high LD (r^2>50%)**
- **Most sites can be tested if multiple populations are available**

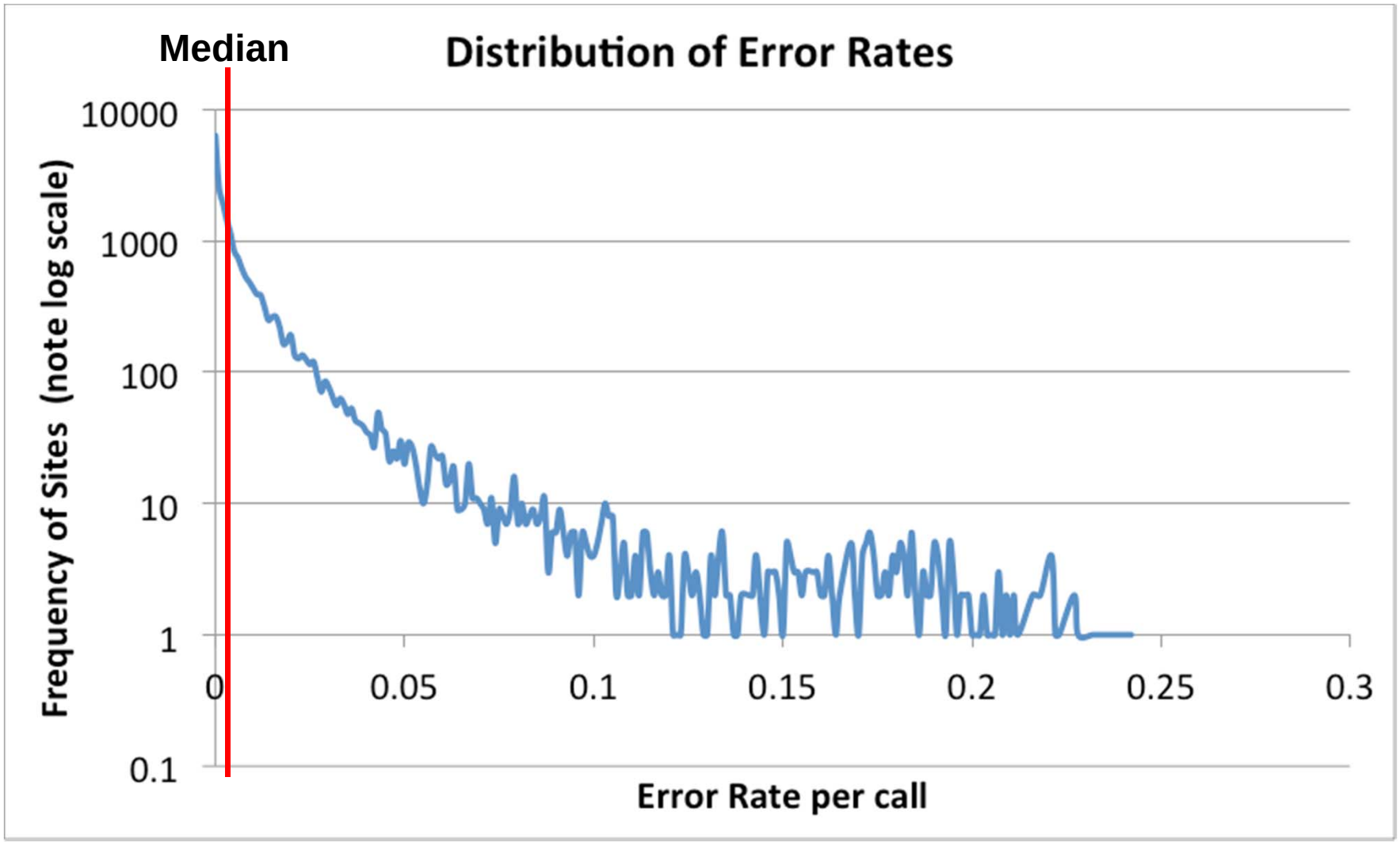
Bi-parental populations allow identification of error, and non-Mendelian segregation



Bi-parental populations allow identification of error, and non-Mendelian segregation



Median error rate is 0.004, but there is a long tail of some high error sites



BiParentalErrorCorrectionPlugin

- **-popM = REGEX population identification(e.g. “Z[0-9]{3}”)**
- **-popF = population File (not implemented) instead of popM option**
- **-mxE = maximum error rate (e.g. 0.01); calculated from non-segregating populations**

BiParentalErrorCorrectionPlugin

- **-mnD = distortion from expectation (e.g. 2.0); the test uses both the binomial distribution and this distortion to classify segregation.**
- **-mnPLD = minimum linkage disequilibrum r^2= 0.5; this is calculated within each population, and then the median across segregating populations is used**

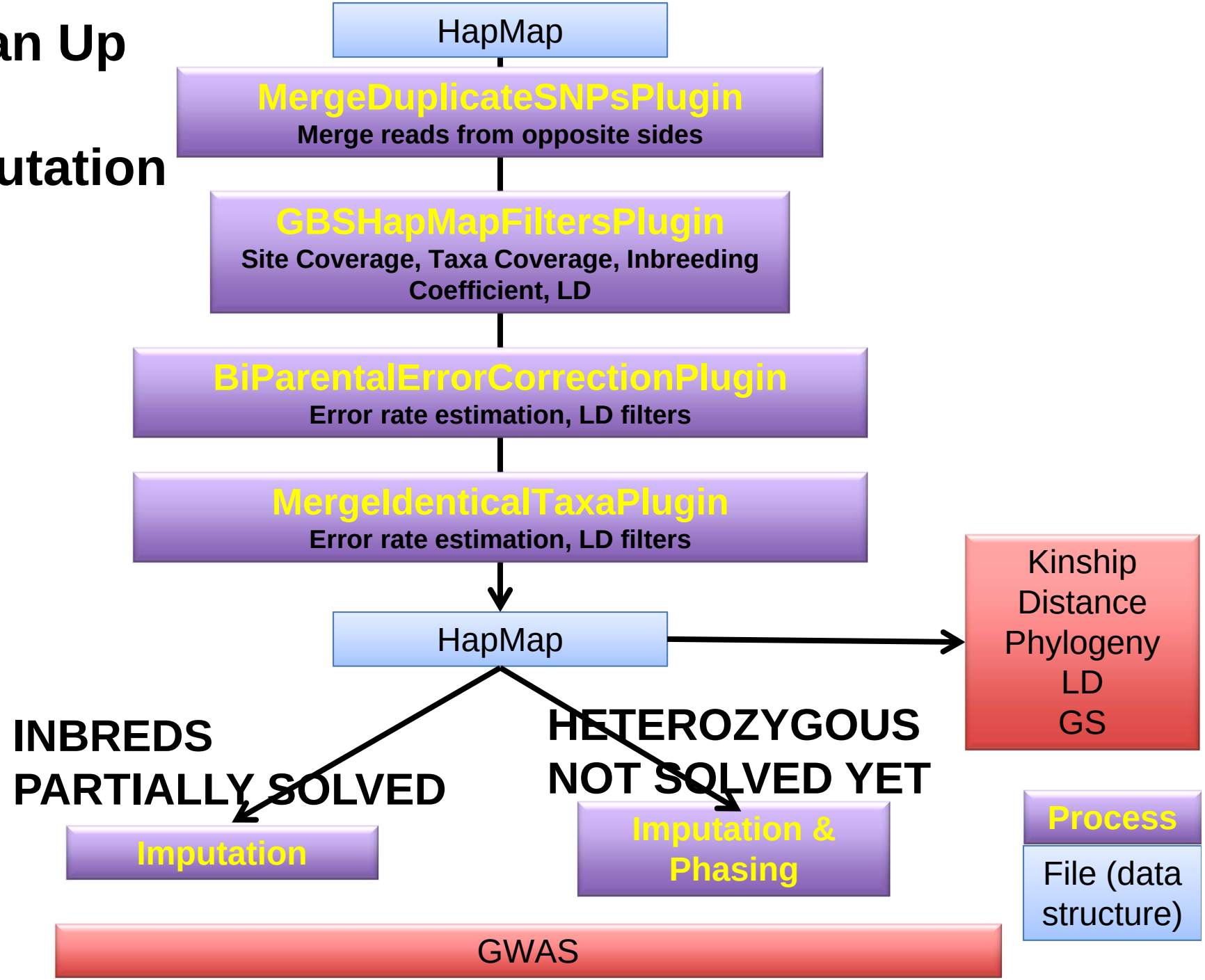
MergeIdenticalTaxaPlugin

- Fuse taxa with the same name. Useful for checks and duplicated runs. Also useful in determining error rates
- -xHets = exclude heterozygotes calls (e.g. true)
- -hetFreq= frequency between hets and homozygous calls (e.g. 0.76)

Product of Filtering

- **After filters, in maize we find 0.0018 error rate**
- **SNPs in wrong location <~1%. Lower in other species.**

Clean Up and Imputation



Missing Data

Two major sources:

- **Sampling**
  - **Low coverage often used in big genomes with inbred lines**
  - **Differential coverage caused by fragment size biases**
- **Biological**
  - **Region on genome not shared between lines**
  - **Cut site polymorphisms**

We want to impute the missing sampling but not the biological

Standard Imputation

- **Lots of algorithms: FastPhase, NPUTE, BEAGLE, etc.**
- **These are appropriate for high coverage loci, inbreds, and regions where biological missing is a rare condition**
- **Some can be slow for sample sizes that we have.**

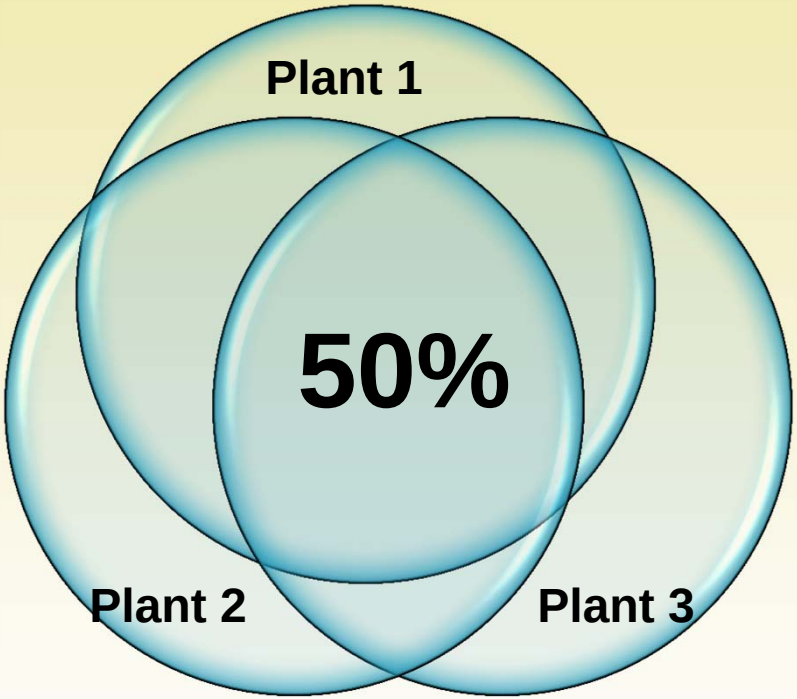
FastImputationBitFixedWindow

- **Imputation approach focused on speed and large sets of taxa with some closely related individuals.**
- **Nearest neighbor approach, fixed window sizes**
- **Strengths: Very accurate <1% error, much faster than other algorithms 100X**
- **Weakness: Not good a recombination junctions, heterozgyosity**
- **Code in TASSEL – not plugin, but available**

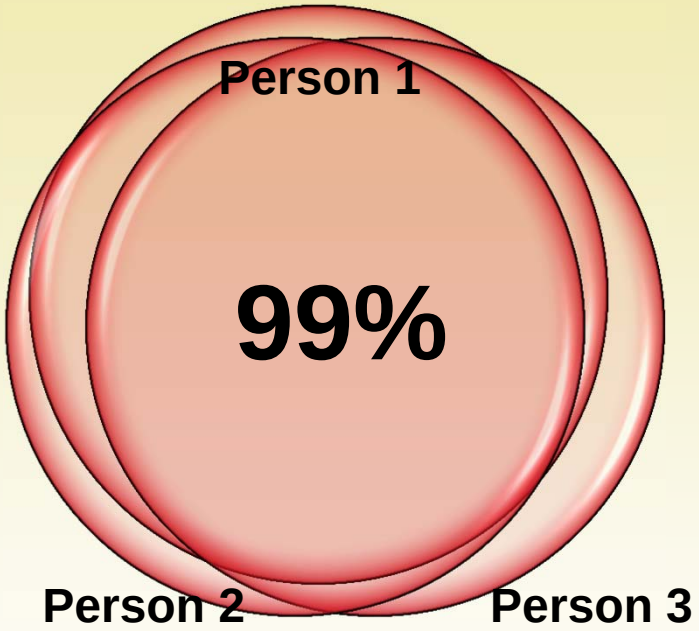
Peter Bradbury Imputation

- Aimed a GBS and biparental populations
- Uses heuristic approaches to get boundaries between crossovers and heterozygous regions.
- Works well on Maize NAM inbred lines, and probably others.
- Available as TASSEL plugin soon.

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



Maize



Humans

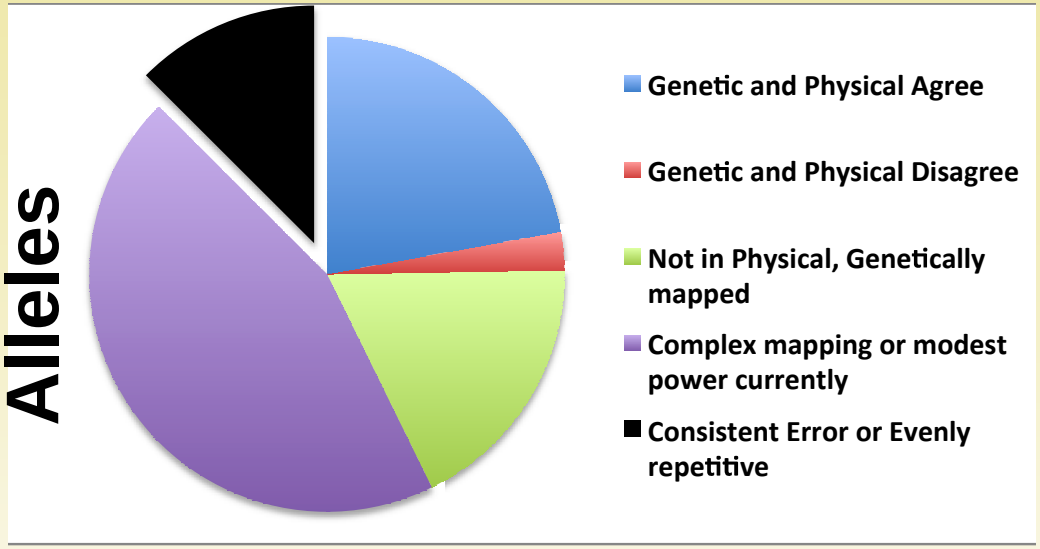
Fu & Dooner 2002, Morgante et al. 2005, Brunner et al 2005

Numerous PAVs and CNVs - Springer, Lai, Schnable in 2010

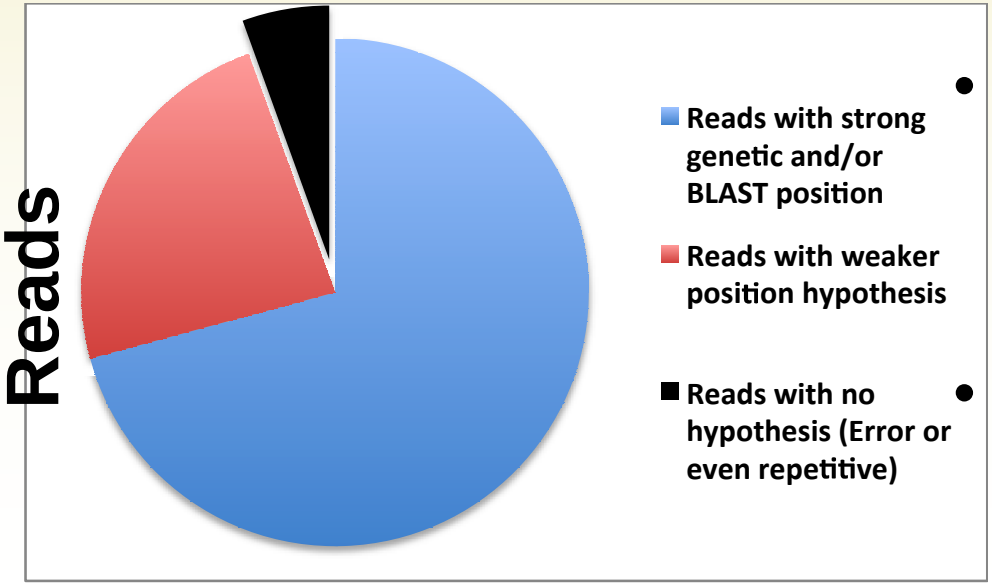
Mapping all the alleles (TagCallerAgainstAnchor)

- Most maize alleles have no position on the reference map
- Map allele presence (TagsByTaxa) versus a anchor SNP map (HapMap)
- 8.7M alleles were mapped in <24 hours using 100 CPU cluster

Physical and genetic mapping of 8.7 million GBS alleles



- Only 29% of alleles are simple - physical and genetic agree
- 55% of alleles are easily genetically mappable



- Many complex alleles are rarer, so 71% of alleles are genetic and/or physically interpretable.
- With more samples and better error models perhaps 90% will be useable

Using the Presence/Absence Variants

- In species like maize, this is the majority of the data
- Less subject to sequencing error
- Need imputation methods to differentiate between missing from sampling and biologically missing

Future

- Need better integration of Whole Genome Sequence data with pipeline
  - Add information on premature cut sites or mutated cut sites
- Paired-End read information

GBS To Do

- **Imputation of linkage pops**
  - **use Peter's code for now** (not a plugin yet)
- **Imputation of GAPs**
  - **Bi-parental imputation of 10% of sites started**
  - **Need more support for gaps in TASSEL**
    - **Current sites filtering tosses (Terry to fix)**
- **Production calling directly with TOPM**
  - **Allow inclusion of low quality sequences (50%+ coverage increase)**
- **Imputation of hets**
- **Imputation of small regions**
- **Included paired ends in TOPM**
- **Genetically map all the PAV tags**

GBS Bioinformatic Pipelines