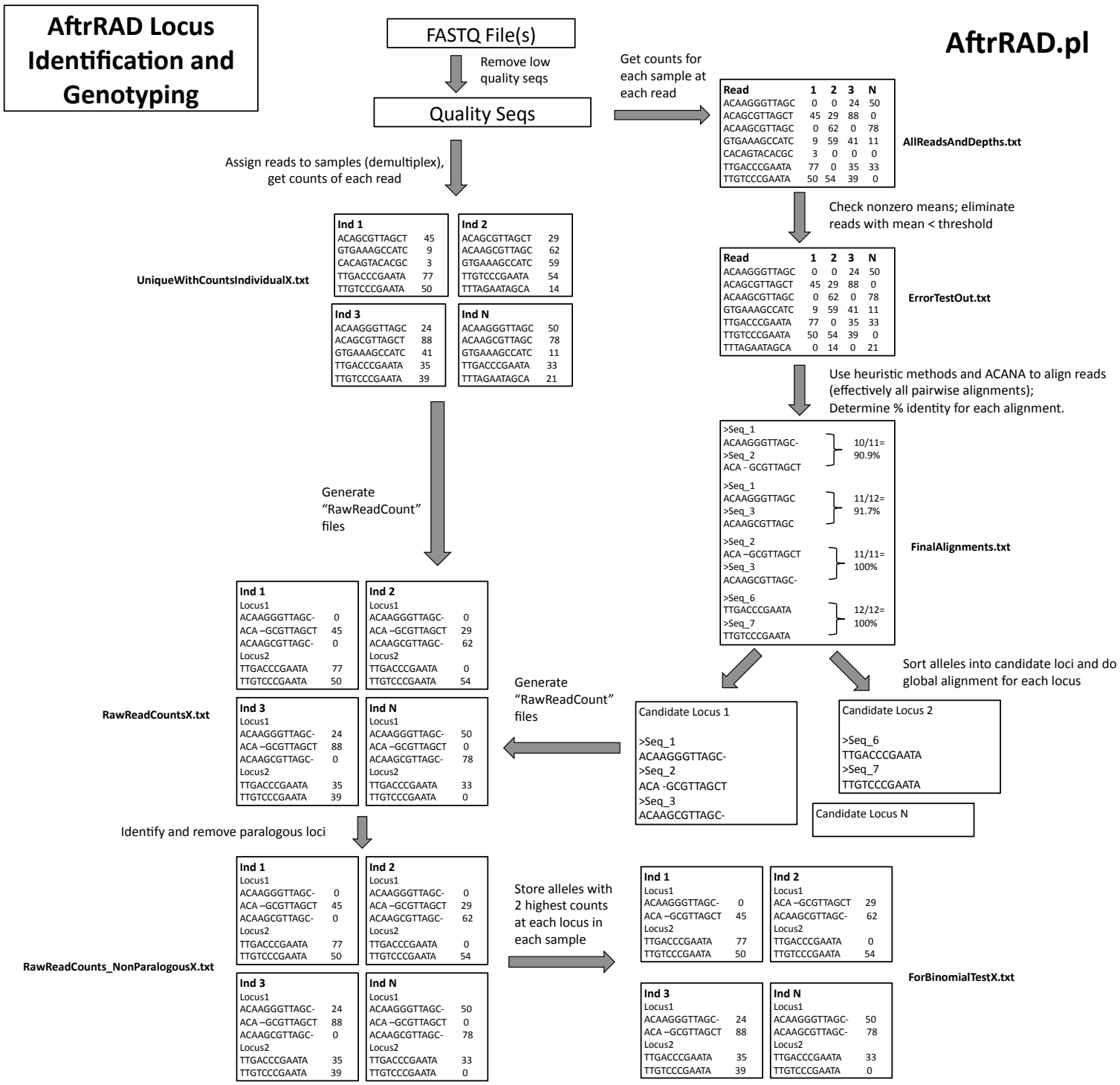

AftrRAD Locus Identification and Genotyping
AftrRAD.pl
FASTQ File(s)
Remove low quality seqs
Quality Seqs
Get counts for each sample at each read
Read 1 2 3 N
ACAAGGGTTAGC 0 0 24 50
ACAGCGTTAGCT 45 29 88 0
ACAAGCGTTAGC 0 62 0 78
GTGAAAGCCATC 9 59 41 11
CACAGTACACGC 3 0 0 0
TTGACCCGAATA 77 0 35 33
TTGTCCCGAATA 50 54 39 0
AllReadsAndDepths.txt
Assign reads to samples (demultiplex), get counts of each read
Check nonzero means; eliminate reads with mean < threshold
Read 1 2 3 N
ACAAGGGTTAGC 0 0 24 50
ACAGCGTTAGCT 45 29 88 0
ACAAGCGTTAGC 0 62 0 78
GTGAAAGCCATC 9 59 41 11
TTGACCCGAATA 77 0 35 33
TTGTCCCGAATA 50 54 39 0
TTTAGAATAGCA 0 14 0 21
ErrorTestOut.txt
UniqueWithCountsIndividualX.txt
Ind 1
ACAGCGTTAGCT 45
GTGAAAGCCATC 9
CACAGTACACGC 3
TTGACCCGAATA 77
TTGTCCCGAATA 50
Ind 2
ACAGCGTTAGCT 29
ACAAGCGTTAGC 62
GTGAAAGCCATC 59
TTGTCCCGAATA 54
TTTAGAATAGCA 14
Ind 3
ACAAGGGTTAGC 24
ACAGCGTTAGCT 88
GTGAAAGCCATC 41
TTGACCCGAATA 35
TTGTCCCGAATA 39
Ind N
ACAAGGGTTAGC 50
ACAAGCGTTAGC 78
GTGAAAGCCATC 11
TTGACCCGAATA 33
TTTAGAATAGCA 21
Use heuristic methods and ACANA to align reads (effectively all pairwise alignments); Determine % identity for each alignment.
>Seq_1
ACAAGGGTTAGC-
>Seq_2
ACA - GCGTTAGCT
10/11= 90.9%
>Seq_1
ACAAGGGTTAGC
>Seq_3
ACAAGCGTTAGC
11/12= 91.7%
>Seq_2
ACA –GCGTTAGCT
>Seq_3
ACAAGCGTTAGC-
11/11= 100%
>Seq_6
TTGACCCGAATA
>Seq_7
TTGTCCCGAATA
12/12= 100%
FinalAlignments.txt
Generate "RawReadCount" files
Sort alleles into candidate loci and do global alignment for each locus
Candidate Locus 1
>Seq_1
ACAAGGGTTAGC-
>Seq_2
ACA -GCGTTAGCT
>Seq_3
ACAAGCGTTAGC-
Candidate Locus 2
>Seq_6
TTGACCCGAATA
>Seq_7
TTGTCCCGAATA
Candidate Locus N
Generate "RawReadCount" files
RawReadCountsX.txt
Ind 1
Locus1
ACAAGGGTTAGC- 0
ACA –GCGTTAGCT 45
ACAAGCGTTAGC- 0
Locus2
TTGACCCGAATA 77
TTGTCCCGAATA 50
Ind 2
Locus1
ACAAGGGTTAGC- 0
ACA –GCGTTAGCT 29
ACAAGCGTTAGC- 62
Locus2
TTGACCCGAATA 0
TTGTCCCGAATA 54
Ind 3
Locus1
ACAAGGGTTAGC- 24
ACA –GCGTTAGCT 88
ACAAGCGTTAGC- 0
Locus2
TTGACCCGAATA 35
TTGTCCCGAATA 39
Ind N
Locus1
ACAAGGGTTAGC- 50
ACA –GCGTTAGCT 0
ACAAGCGTTAGC- 78
Locus2
TTGACCCGAATA 33
TTGTCCCGAATA 0
Identify and remove paralogous loci
Store alleles with 2 highest counts at each locus in each sample
RawReadCounts_NonParalogousX.txt
ForBinomialTestX.txt

Genotype.pl

Ind 1	
Locus1	
ACAAGGGTTAGC-	0
ACA –GCGTTAGCT	45
Locus2	
TTGACCCGAATA	77
TTGTCCCGAATA	50

Ind 2	
Locus1	
ACA –GCGTTAGCT	29
ACAAGCGTTAGC-	62
Locus2	
TTGACCCGAATA	0
TTGTCCCGAATA	54

Ind 3	
Locus1	
ACAAGGGTTAGC-	24
ACA –GCGTTAGCT	88
Locus2	
TTGACCCGAATA	35
TTGTCCCGAATA	39

Ind N	
Locus1	
ACAAGGGTTAGC-	50
ACAAGCGTTAGC-	78
Locus2	
TTGACCCGAATA	33
TTGTCCCGAATA	0

ForBinomialTestX.txt

Perform binomial test at each sample/locus to call genotypes.

	Locus 1	Locus 2	Locus N
Ind1	ACA –GCGTTAGCT	TTGACCCGAATA	
Ind1	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind2	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind2	ACAAGCGTTAGC -	TTGTCCCGAATA	
Ind3	ACAAGGGTTAGC-	TTGACCCGAATA	
Ind3	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind N	ACAA GGGTTAGC-	TTGACCCGAATA	
Ind N	ACAA GCGTTAGC-	TTGACCCGAATA	

Genotypes.txt

Generate matrix containing called SNPs.

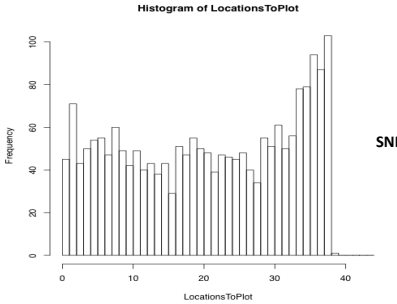
	Locus1	Locus2	Locus3	Locus3	Locus4	Locus5
Ind1	C	A	T	A	G	T
Ind1	C	T	T	A	A	G
Ind2	C	T	T	A	NA	T
Ind2	C	T	T	A	NA	G
Ind3	G	A	C	T	A	T
Ind3	C	T	T	A	A	G
IndN	G	A	NA	NA	NA	T
IndN	C	A	NA	NA	NA	T

Calculate proportion of missing data for each sample.
Option to remove bad samples.

Sample	**% Missing**
Ind1	0
Ind2	0.20
Ind3	0
IndN	0.40

MissingDataProportions.txt

Get and plot the position along the read of each SNP in the dataset.



SNPLocations.pdf

OutputSNPs.pl

	Locus 1	Locus 2	Locus N
Ind1	ACA –GCGTTAGCT	TTGACCCGAATA	
Ind1	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind2	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind2	ACAAGCGTTAGC -	TTGTCCCGAATA	
Ind3	ACAAGGGTTAGC-	TTGACCCGAATA	
Ind3	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind N	ACAA GGGTTAGC-	TTGACCCGAATA	
Ind N	ACAA GCGTTAGC-	TTGACCCGAATA	

Genotypes.txt

Remove individuals identified as bad in Genotype.pl.

	Locus 1	Locus 2	Locus N
Ind1	ACA –GCGTTAGCT	TTGACCCGAATA	
Ind1	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind2	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind2	ACAAGCGTTAGC -	TTGTCCCGAATA	
Ind3	ACAAGGGTTAGC-	TTGACCCGAATA	
Ind3	ACA –GCGTTAGCT	TTGTCCCGAATA	
Ind N	ACAA GGGTTAGC-	TTGACCCGAATA	
Ind N	ACAA GCGTTAGC-	TTGACCCGAATA	

GenotypesUpdate.txt

Choose the threshold proportion of samples to be genotyped at a locus, and the maximum position along reads to call SNPs (see SNP locations plot from Genotypes.pl). SNPs and haplotypes are output in the SNPMatrix_X.Y.txt and Haplotypes_X.Y.txt files, respectively, based on the chosen parameters. X refers to the proportion of samples that must be genotyped at each locus, and Y is the max SNP location.

Seqs in ErrorTestOut.txt and not in the 'AllPoly' file from AftrRAD.pl are monomorphic. Filter these based on the value chosen for the threshold proportion of samples genotyped at each locus (X), and for each locus meeting this threshold, print it, with its counts in each sample to Monomorphics_X.txt.

SNPMatrix_75_33.txt

	Locus1	Locus2	Locus3	Locus3	Locus5
Ind1	C	A	T	A	T
Ind1	C	T	T	A	G
Ind2	C	T	T	A	T
Ind2	C	T	T	A	G
Ind3	G	A	C	T	T
Ind3	C	T	T	A	G
IndN	G	A	NA	NA	T
IndN	C	A	NA	NA	T

Haplotypes_75_33.txt

	Locus1	Locus2	Locus3	Locus5
Ind1	C	A	TA	T
Ind1	C	T	TA	G
Ind2	C	T	TA	T
Ind2	C	T	TA	G
Ind3	G	A	CT	T
Ind3	C	T	TA	G
IndN	G	A	NA	T
IndN	C	A	NA	T

Monomorphics_75.txt

Read	1	2	3	N
GTGAAAGCCATC	9	59	41	11