Motif Loss Results

Species: mice
Run version: 2014_10_24
Saturday 25th October, 2014, 11:58

Changelog

Date refers to first version with changes. If there is a plus after, then all subsequent versions have the change

- **2014_10_24** Added filtering folder for mice (ArrayNotFiltered95.0 used in all prior runs)
- **2014_10_01** Changed clustering method - increase difficulty such that current threshold for iteration is Bonferonni for this round and all other previous rounds. Also, only allow a motif to be included in one cluster per repeat background
- **2014_10_01** BUGFIX - Fixed major repeat name problem - assigning motifs to repeats was fine, but got name wrong between assigning and reporting. Also affected generation of QQ plots in this summary document
- **2014_09_22** Added number of motif clusters in summary section
- **2014_09_22** Add new AT to GC plots, and details of making the plots
- **2014_09_18** Give the option to specify how to compare against other motifs for making 2x2 tables
- **2014_09_18** Don't perform motif loss counting if the the repeat background changes during the motif loss region
- **2014_09_17** For mice, set regions of the genome as uncallable if insufficient number of samples are callable in that region
- **2014_09_16** Clustering - changed it so motifs are only eligible to be added to a cluster if they have the best p-value among lineages for that motif
- **2014_09_15** added removal of A) nearby SNPs on the same lineage and B) too many SNPs in general locally, as well as check against expectations
- **2014_09_15** added proper support for mice for smaller number of musculus lineages
- **2014_09_13** BUGFIX - x axis in left plot for comparison between two test p-values

Important parameters

variable	value	explanation
mouseFilter	ArrayNotFilteredAnnot295.0	what VQSR filtering was used (mice only)
pThresh	1.5244×10^{-4}	p-value threshold used for clustering
mrle	10	analyze motifs if the Maximum Run of a certain nucleotide is Less tha Equal to this number
ndge	0	analyze motifs if the Nucleotide Diversity (ie number of A,C,G,T) bas Greater than or Equal to this number
gcW	100	for AT to GC p-value testing, window left and right of current motif in w we count AT to GC changes
gcW2	1000	window over which to plot AT to GC plot
gcW3	10	smoothing for AT to GC plot
cgte	10	for a p-value to be generated, each Cell must be Greater Than or Equal to number
rgte	50	for a p-value to be generated, each Row must have Greater than or Equ this number
nR1	12	if there are ge this many SNPs in nD1 bp, remove all SNPs
nD1	50	see above
nR2	7	if there are ge this many SNPs down any individual lineage in nD2 bp, re all SNPs
nD2	50	see above
removeNum	0	how many samples are allowed to be uncallable for a region to still be cal

Methods

SNP Filtering

SNPs are filtered if they have any missingness among the lineages, do not agree with the species tree, are multi-allelic, or are heterozygous among homozygous animals.

Subsequently, SNPs are removed if they are too close together. This is the last step in the procedure, ie after removing SNPs for reasons listed above. I removed any SNPs if there were nR1 SNPs within nD1 bases, and, following this, I removed lineage specific SNPs down any lineage if there were nR2 SNPs within nD2 bases. For example, if nR1 was 10 and nD1 was 50, and there was a cluster of 14 SNPs within 50 bases, all of the offending SNPs were removed.

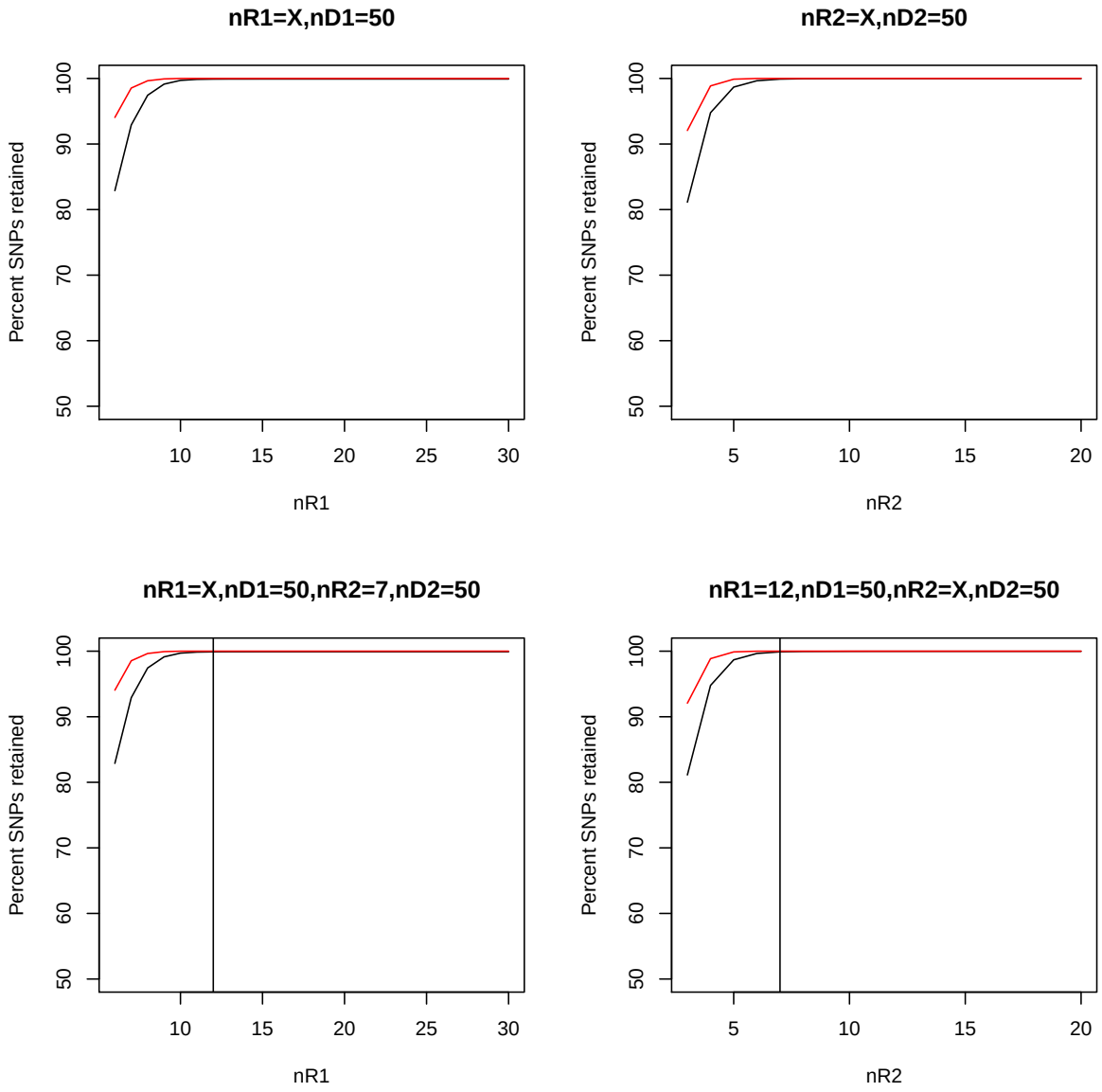
To get a sense of how many SNPs this removed for given parameter settings, I checked how many SNPs were filtered for a range of parameter settings for the smallest chromosome. I also compared this against expectations. To get an expectation, I simulated a pseudo-chromosome of results. I calculated the expected branch lengths of each lineage given the emprical data to this point, ie lineage 1 has 1 percent divergence against MRCA of the set of all lineages, lineage 2 has 0.5 percent divergence against MRCA, etc. Then, I decided whether each base was mutated according to the total branch length of the tree, and then, given a mutation, what branch it occured on with probabilities equal to each lineages share of the total tree length.

Figure 1 shows the results.

Testing - general

3 tests are run based on motifs which are lost at every SNP down a lineage. Note that where there are multiple SNPs within a motif size (ie two SNPs 5 bases away), I considered all motifs lost over that range. If there were any missing bases within that range, the motifs were not counted towards the testing.

Figure 1: Effect of filtering on real data (black) and estimation based on empirical values (red)



Further, if the repeat background changed during that range, the motifs were not counted towards the testing.

Test 1 - Lineage - Motif loss within a lineage versus other lineages

Here I look at the number of motifs lost down a lineage versus all other lineages. The 2×2 test is therefore

	lineage	all other lineages
motif	n_1	n_2
motifs with same CpG, GC	n_3	n_4

Test 2 - Shared - Motif loss within a lineage versus ancestral counts

Here, I look at the number of motifs lost down a lineage versus the number present in its ancestor in the species tree. Then, we test the number of motifs lost vs the number which are not lost. The 2×2 test is therefore

	lost down this lineage	not lost down this lineage
motif	n_1	n_2
motifs with same CpG, GC	n_3	n_4

Test 3 - AT to GC - Motif loss AT to GC counts versus local AT bases

Let $gcW = 100$ be the distance to test for increased presence of AT to GC bases. Then around every motif that is lost, we count three numbers

- The number of AT to GC changes in nearest gcW bases away from the motif, not including the motif itself. For example, with $K = 10$, for a motif from $1,000$ to $1,009$ inclusive, then the following 100 bp are from $1,010$ to $1,109$, inclusive. So an A to a G at $1,109$ is within range while an A to a C at $1,110$ is out of range.
- The number of AT bases within gcW bases, not including motif
- The number of callable bases within gcW bases, not including motif

We are currently not using the third value (the number of callable bases), only the total number of AT bases. The 2×2 test per motif is

	AT to GC bases	not missing and not AT to GC bases
motif	n_1	n_2
motifs with same CpG, GC	n_3	n_4

Motif Filtering

Motifs were kept if the longest run of a specific nucleotide was less than or equal to $mrle = 10$. Motifs were also kept if the nucleotide diversity (ie number of A,C,G,T) bases was greater than or equal to $ndge = 0$

For each test seperately, motifs were removed as well as individual results masked if the counts were too low. Motif level results were not calculated if the counts among all lineages for a motif was less than $rgte = 50$. Lineage specific results were also masked (ie not calculated) in each test seperately if the motif lineage value was less than $cgte = 10$ (ie the n_1 entry in the contingency tables).

Clustering

Given a set of motifs which pass Bonferonni correction for a given lineage for a given test for a given repeat background which are most significant in this lineage versus all other lineages, start with the motif with the most significant p-value. Consider as eligible all motifs on the repeat background for that test which are the "most significant" (ie lowest p-value) for that motif and beat a threshold which increases in difficulty as the iterative process goes on (defined below). Add to the cluster all motifs within a certain distance (defined below), keeping track of alignment. Continue recursively for each added motif until

exhausted. Build a position weight matrix by collapsing all clustered motifs, counting each base, adding 0.5 to all pwm cell entries, and dividing by the column totals to have each entry scaled between 0 and 1.

Let K be the motif length we are interested in. We define as acceptably close for clustering all motifs which align perfectly and are off by 1 base (sum of $K * 4$), or that are off by 1 base in the alignment left or right (2 for left vs right) with any new base in the gap (4 bases) with any one of the remaining $K - 1$ bases allowed to change as well (4). In total there are $(K-1) * 4 * 4 * 2 + K * 4 = 328$ possible acceptably close motifs.

For the first iteration, take a p-value threshold of the number of motifs to be searched, or $(K-1) * 4 * 4 * 2 + K * 4 = 328$. For each subsequent iteration, take a p-value threshold where the deminator is the number of tests already performed plus the number to be searched on this iteration. For example, if on the first iteration 3 close motifs were found to meet the iteration 1 p-value threshold, then on the second iteration the p-value threhsold would be 0.05 divided by the number of motifs searched on the first iteration, 328, plus the 3×328 to be searched on the second iteration.

For plotting, we summed the collapsed motif clusters counts by base and added 0.5 to each count. Then, for a motif cluster of length m, letting $j \in \{1, 2, 3, 4\}$ be the four nucleotides and $i \in \{1, 2, ..., m\}$ be the position within the motif cluster, we set $H_i = \sum_{j=1}^{4} f_{i,j} \log(f_{i,j})$, and define the height of each base as $h_{i,j} = H_i f_{i,j}$. Bases are then ordered from smallest to largest entropy and plotted. We used elements of the seqLogo R package to draw the PWM.

AT to GC plots

To better visualize the localization of AT to GC changes surrounding a motif cluster, we plotted whether bases changed from AT to GC or vice versa, with respect to their distance from the motif. We first scanned through the chromosomes to find instances where motifs in the motif cluster were lost. To ensure we weren't oversampling SNP changes due to similar motifs, we limited ourselves to counting only a single motif loss instance among a run of motif losses each one within the length of the motif cluster in distance from each other.

Next, taking care to get both the correct strand as well as the position of the motif within the cluster of motifs correct, we catalogued both the position and base composition of any changes within a neighbourhood of 1000 bases. By summing across all loss instances of motifs in the cluster, and normalizing to the local sequence context, we could plot any type of ancestral to derived base change. Smoothing was done over 21 bases, ie taking the value at the flanking bases and over the prior and aft 10 bases.

These plots also feature a PWM for the forward and reverse forms of the motif, as well as a series of line plots which show the number of motifs and their p-values for the motifs in the cluster. The middle line is for the test under consideration, while p-values for the other two tests are highlighted above and below, with grey lines linking the same motif (motifs with undefiend p-values on the other two tests are omitted from the plots for those tests and are not linked). These are stratefied into those which are Bonferonni significant on their test to the right of the red line, those which are between the initial clustering p-value threshold and the Bonferonni thresold in the middle, and those which do not meet the initial clustering p-value threhsold on the left. Numbers of motifs falling into each category are given as well.

Some summary numbers

```
## Aligned Genome (Gbp)
##      Total Pass QC Fail QC Pass QC Non Repeat
## [1,] 2.472   1.645  0.8274              1.114
```

```
## Number of Derived Mutations down a specific lineage
##        FAM        AMS    Spretus         AM     WSBEiJ    CASTEiJ
## 17,935,113  8,439,998 11,800,414  4,728,452  6,066,739  6,178,547
##     PWKPhJ
##  6,347,677
## Branch length as percent of alignable genome
##       FAM   AMS Spretus    AM WSBEiJ CASTEiJ PWKPhJ
## [1,] 1.09 0.513   0.717 0.287  0.369   0.376  0.386
## Branch length compared to ancestral of all lineages in SNPs
##      PWKPhJ     CASTEiJ    WSBEiJ     Spretus    FAM
## [1,] 19,516,127 19,346,997 19,235,189 20,240,412 17,935,113
## Branch length compared to ancestral as percent of alignable genome
##      PWKPhJ CASTEiJ WSBEiJ Spretus  FAM
## [1,]  1.186   1.176  1.169    1.23 1.09
```

```
## Number of significant motifs and motif clusters which are most significant down a given lineage
## Number of singificant results (and clusters) per lineage and test
##        FAM       AMS      Spretus  AM       PWKPhJ CASTEiJ CASTEiJ.PWKPhJ
## at     174 (15)  12 (7)   531 (13) 124 (13) 2 (2)  3 (3)   0 (0)
## lin    38 (8)    91 (16)  406 (19) 413 (9)  1 (1)  6 (3)   0 (0)
## shared 1656 (56) 424 (53) 965 (42) 550 (17) 3 (3)  10 (3)  1 (1)
##        WSBEiJ WSBEiJ.CASTEiJ
## at     19 (6) 0 (0)
## lin    10 (2) 1 (1)
## shared 26 (5) 1 (1)
## FAM
##        nonRepeat (CA)n  (TG)n  B1_Mus1 B1_Mus2 RSINE1 B2_Mm2 B3     B1_Mm
## at     167 (11)  0 (0)  0 (0)  0 (0)   0 (0)   0 (0)  0 (0)  0 (0)  0 (0)
## lin    29 (5)    0 (0)  0 (0)  0 (0)   0 (0)   5 (1)  0 (0)  0 (0)  0 (0)
## shared 1268 (4)  88 (2) 80 (1) 44 (7)  31 (2)  19 (4) 21 (7) 16 (2) 15 (4)
##        ID_B1 Others
## at     0 (0) 7 (4)
## lin    1 (1) 3 (1)
## shared 8 (2) 66 (21)
## AMS
##        nonRepeat RSINE1 B1_Mus1 B1_Mus2 B1_Mm  B3    MTC   B2_Mm2 B4A
## at     6 (4)     1 (1)  0 (0)   0 (0)   0 (0)  0 (0) 0 (0) 0 (0)  4 (1)
## lin    56 (8)    18 (1) 1 (1)   0 (0)   0 (0)  5 (1) 4 (1) 0 (0)  1 (1)
## shared 243 (9)   41 (2) 30 (1)  24 (6)  21 (6) 8 (2) 9 (3) 12 (7) 7 (3)
##        ORR1C2 Others
## at     0 (0)  1 (1)
## lin    3 (1)  3 (2)
## shared 5 (1)  24 (13)
## Spretus
##        nonRepeat B3A    Lx3C   B4A    B3     RMER17C (TG)n  AT_rich RSINE1
## at     479 (5)   23 (1) 0 (0)  13 (1) 3 (1)  0 (0)   0 (0)  0 (0)   2 (1)
## lin    360 (8)   12 (1) 11 (1) 3 (2)  0 (0)  8 (1)   2 (1)  0 (0)   4 (1)
## shared 821 (6)   32 (2) 12 (1) 5 (3)  15 (1) 9 (1)   13 (6) 13 (1)  6 (1)
##        URR1A Others
```

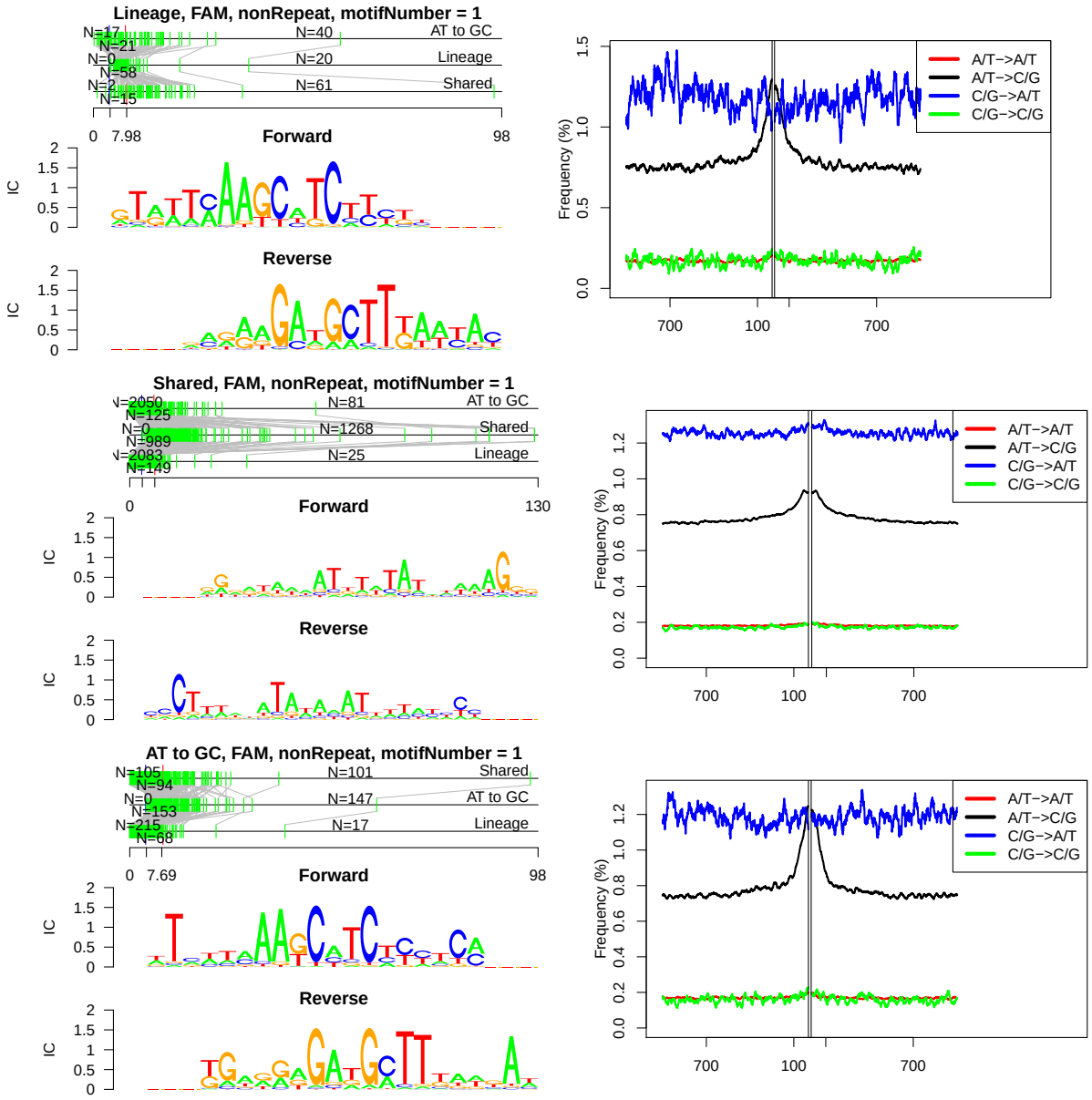
```
## at     0 (0) 11 (4)
## lin    2 (1) 4 (3)
## shared 8 (1) 31 (19)
## AM
##        nonRepeat ID_B1  RSINE1 B3    B4A   AT_rich B1_Mus2 (TCTA)n B1_Mm
## at     115 (8)   0 (0)  0 (0)  7 (3) 0 (0) 0 (0)   0 (0)   0 (0)   0 (0)
## lin    393 (4)   7 (2)  7 (1)  5 (1) 1 (1) 0 (0)   0 (0)   0 (0)   0 (0)
## shared 500 (5)   17 (2) 15 (2) 8 (2) 3 (2) 2 (1)   2 (1)   2 (1)   1 (1)
##        B3A   ID4
## at     1 (1) 1 (1)
## lin    0 (0) 0 (0)
## shared 0 (0) 0 (0)
## PWKPhJ
##        nonRepeat ID_B1
## at     1 (1)     1 (1)
## lin    1 (1)     0 (0)
## shared 3 (3)     0 (0)
## CASTEiJ
##        nonRepeat MTD   (TCTA)n
## at     3 (3)     0 (0) 0 (0)
## lin    5 (2)     1 (1) 0 (0)
## shared 8 (1)     1 (1) 1 (1)
## CASTEiJ.PWKPhJ
##        nonRepeat
## shared 1 (1)
## WSBEiJ
##        nonRepeat (TG)n
## at     19 (6)    0 (0)
## lin    10 (2)    0 (0)
## shared 25 (4)    1 (1)
## WSBEiJ.PWKPhJ
## [1] "No significant results"
## WSBEiJ.CASTEiJ
##        RSINE1
## lin    1 (1)
## shared 1 (1)
```

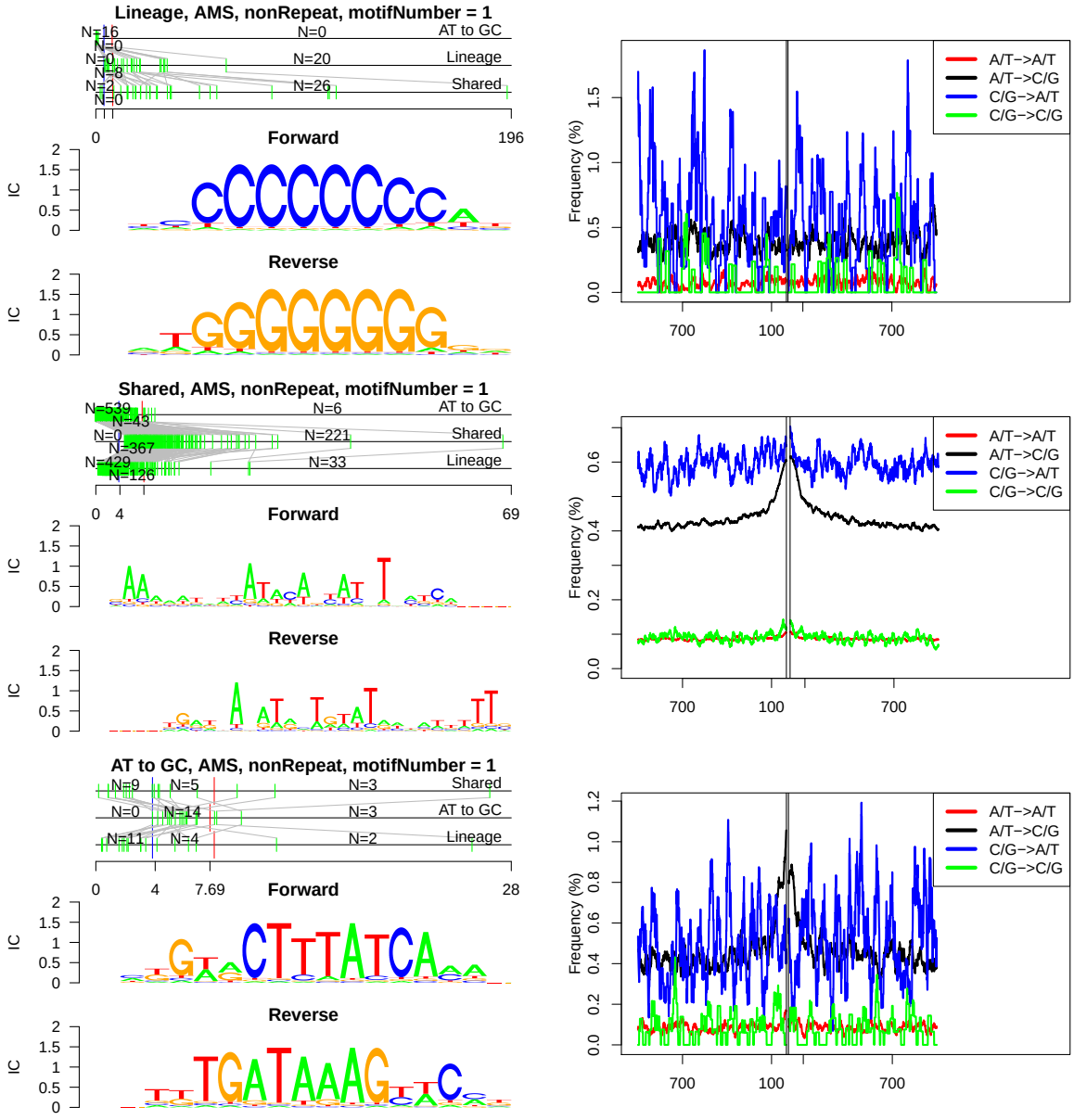
```
## Features of top associated motifs
## Test: lin
## AT content
##
##                     0   1   2    3    4    5    6    7   8   9  10
##   not significant 0.0 0.6 3.2 10.0 20.1 25.6 21.6 12.4 4.9 1.3 0.3
##   significant     0.3 1.1 1.2  3.9 11.4 28.0 27.0 14.2 3.7 6.1 3.1
## Number of CpGs
##
##                      0    1   2 3 4
##   not significant 69.6 28.3 2.1 0 0
##   significant     99.8  0.2 0.0 0 0
```

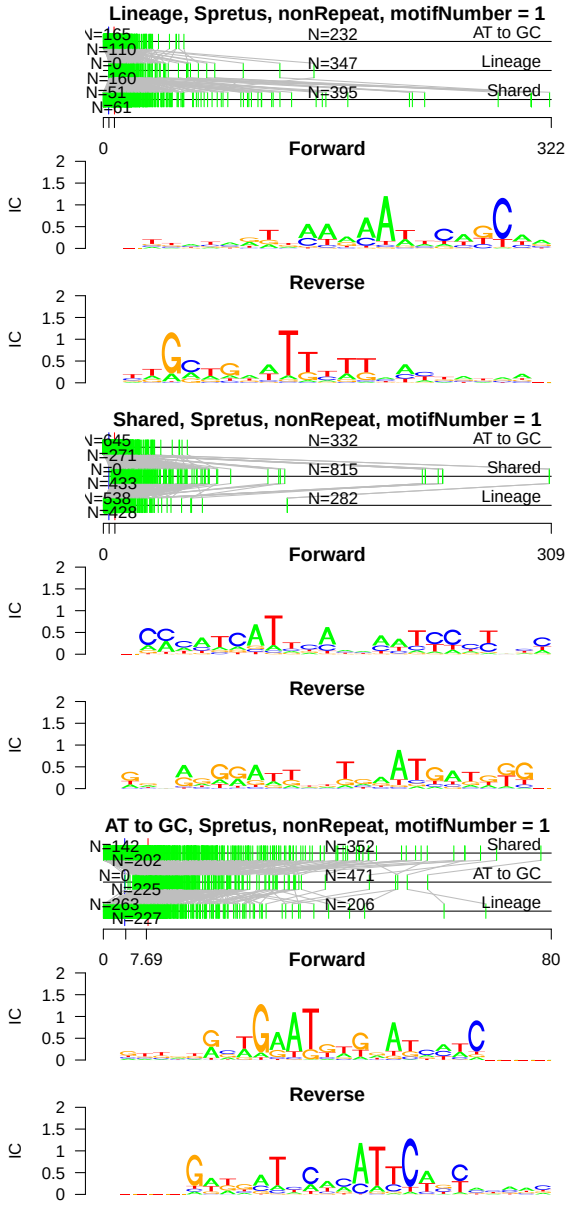
```
## Maximum run length
##
##                     1    2    3   4   5   6   7 8 9  10
##   not significant 7.2 54.7 28.3 7.6 1.8 0.3 0.1 0 0 0.0
##   significant     8.2 59.7 15.4 2.1 1.4 2.2 5.8 4 1 0.2
## Nucleotide diversity
##
##                     1   2    3    4
##   not significant 0.0 1.9 23.4 74.7
##   significant     0.2 8.8 34.3 56.6
## Test: shared
## AT content
##
##                     0   1   2    3    4    5    6    7    8   9  10
##   not significant 0.0 0.6 3.2 10.1 20.2 25.7 21.6 12.4  4.8 1.2 0.3
##   significant     0.2 1.1 1.8  3.7  7.7 16.1 19.8 20.0 16.6 9.9 3.1
## Number of CpGs
##
##                      0    1   2 3 4
##   not significant 69.4 28.5 2.1 0 0
##   significant     97.7  2.1 0.3 0 0
## Maximum run length
##
##                     1    2    3    4   5   6   7   8   9 10
##   not significant  7.2 54.9 28.4  7.6 1.7 0.3 0.0 0.0 0.0  0
##   significant     10.5 37.6 25.7 13.6 4.8 3.6 3.2 0.8 0.2  0
## Nucleotide diversity
##
##                   1    2    3    4
##   not significant 0  1.8 23.2 74.9
##   significant     0 11.3 45.4 43.2
## Test: at
## AT content
##
##                   0   1   2   3    4    5    6    7   8   9  10
##   not significant 0 0.3 2.5 9.2 19.4 25.8 22.8 13.4 5.1 1.3 0.3
##   significant     0 0.0 0.4 3.4  9.8 25.6 30.8 20.7 5.9 1.7 1.6
## Number of CpGs
##
##                      0    1   2 3
##   not significant 68.1 31.8 0.1 0
##   significant     99.1  0.9 0.0 0
## Maximum run length
##
##                     1    2    3   4   5   6 7   8   9  10
##   not significant 7.2 55.2 28.2 7.4 1.6 0.3 0 0.0 0.0 0.0
##   significant     5.5 56.8 27.9 3.4 1.1 1.4 2 1.3 0.4 0.1
## Nucleotide diversity
##
```

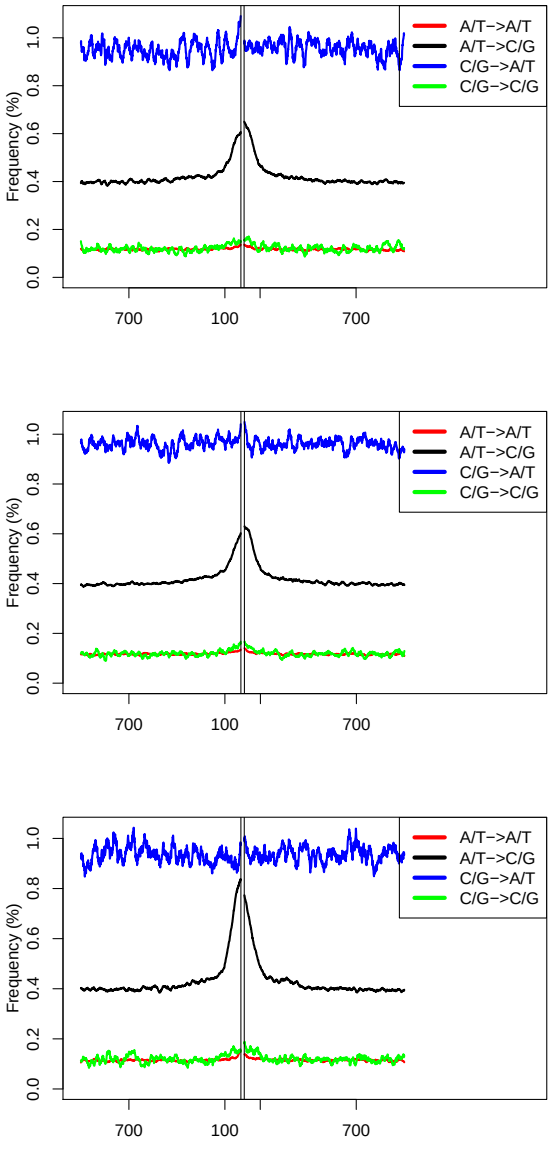
```
##                       1   2    3    4
##   not significant 0.0 1.3 22.3 76.4
##   significant     0.1 5.9 36.1 57.9
```

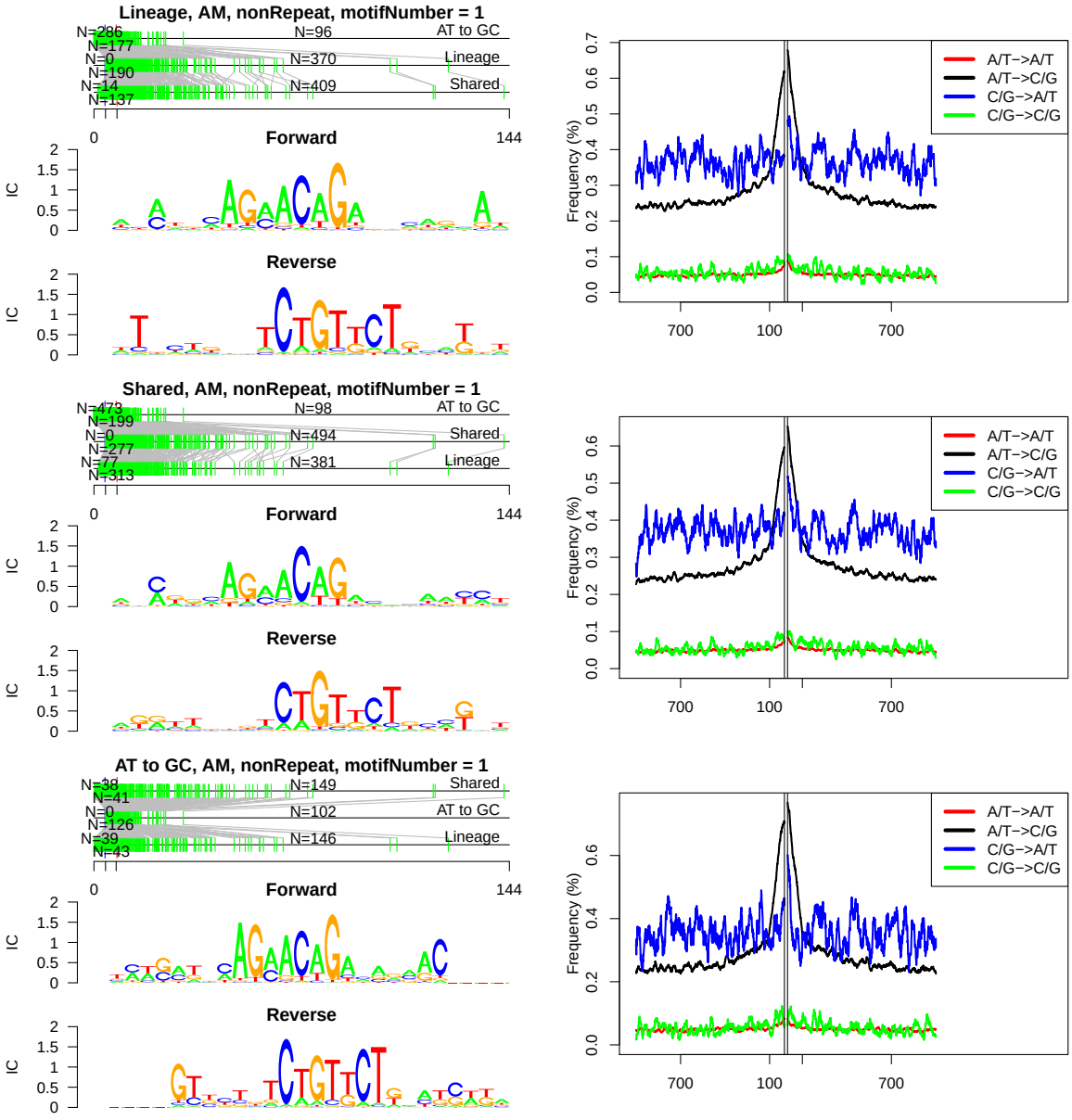
Top non-repeat clusters for AT, Shared and Lineage tests

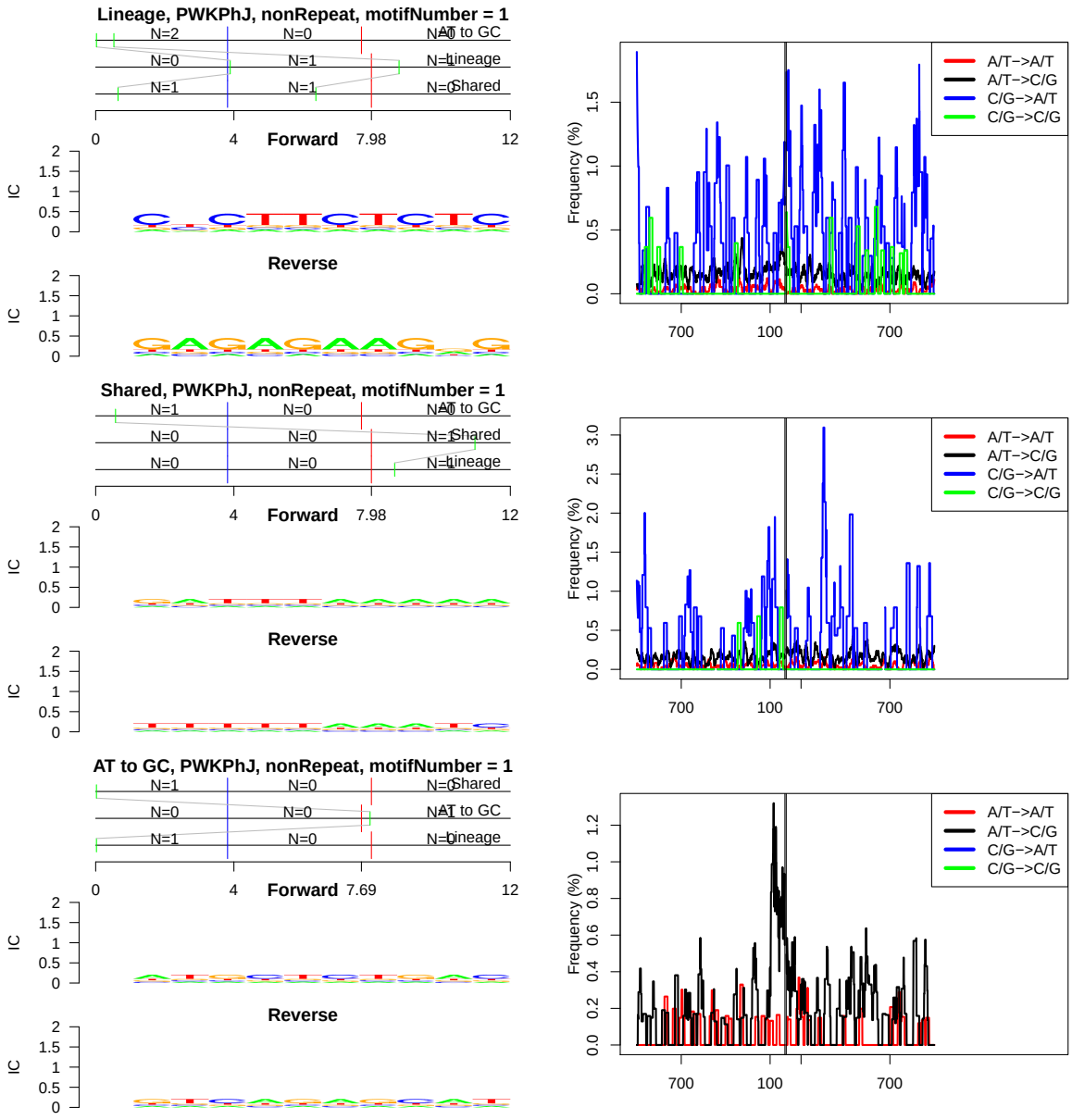


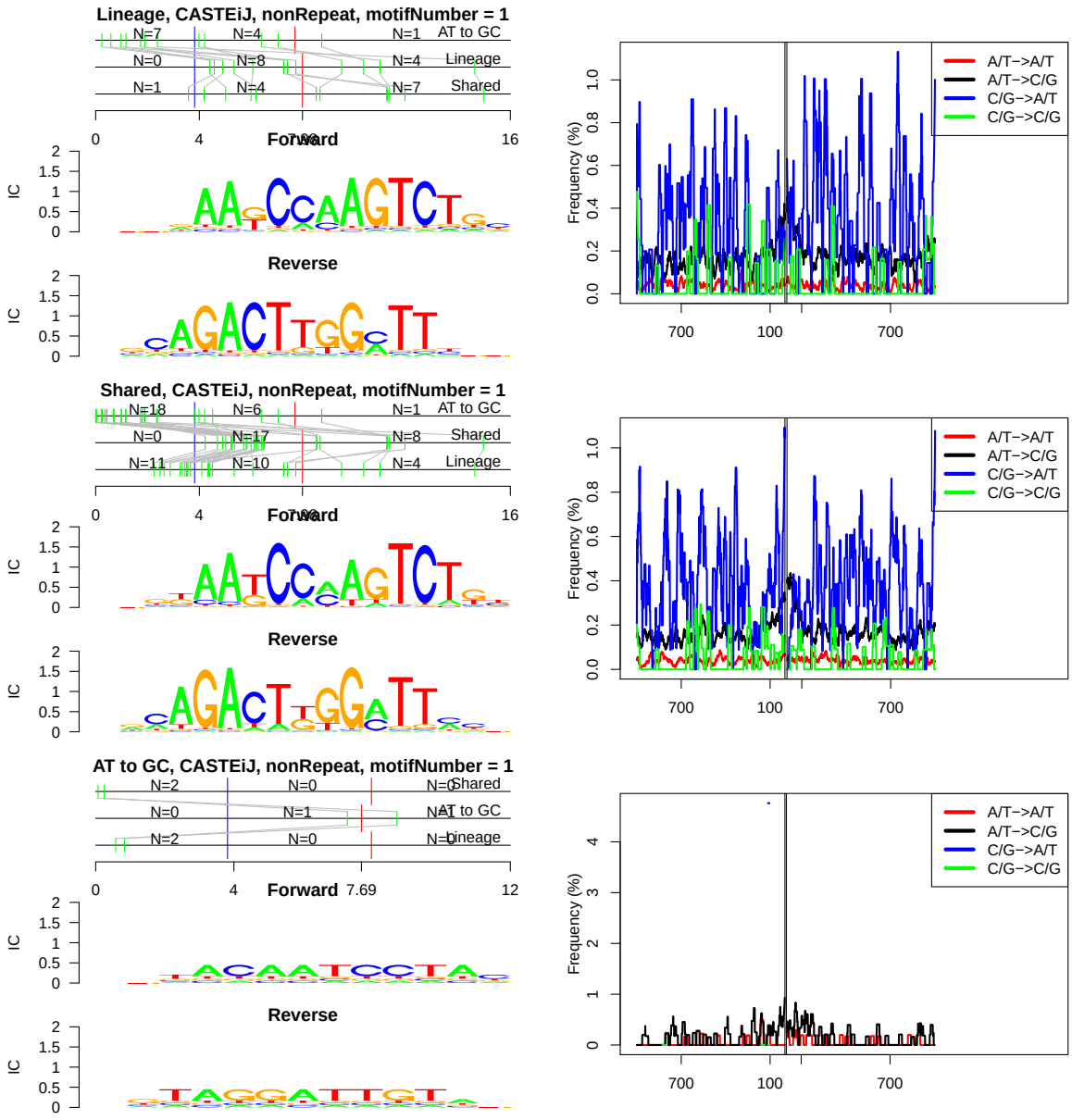

Lineage, AMS, nonRepeat, motifNumber = 1
N=16
N=0
N=0
N=8
N=2
N=0
N=0
N=20
N=26
AT to GC
Lineage
Shared
0
196
Forward
Reverse
IC
A/T->A/T
A/T->C/G
C/G->A/T
C/G->C/G
Frequency (%)
700
100
700
Shared, AMS, nonRepeat, motifNumber = 1
N=539
N=43
N=0
N=367
N=429
N=126
N=6
N=221
N=33
AT to GC
Shared
Lineage
0
4
69
Forward
Reverse
AT to GC, AMS, nonRepeat, motifNumber = 1
N=9
N=5
N=0
N=14
N=11
N=4
N=3
N=3
N=2
Shared
AT to GC
Lineage
0
4
7.69
28
Forward
Reverse

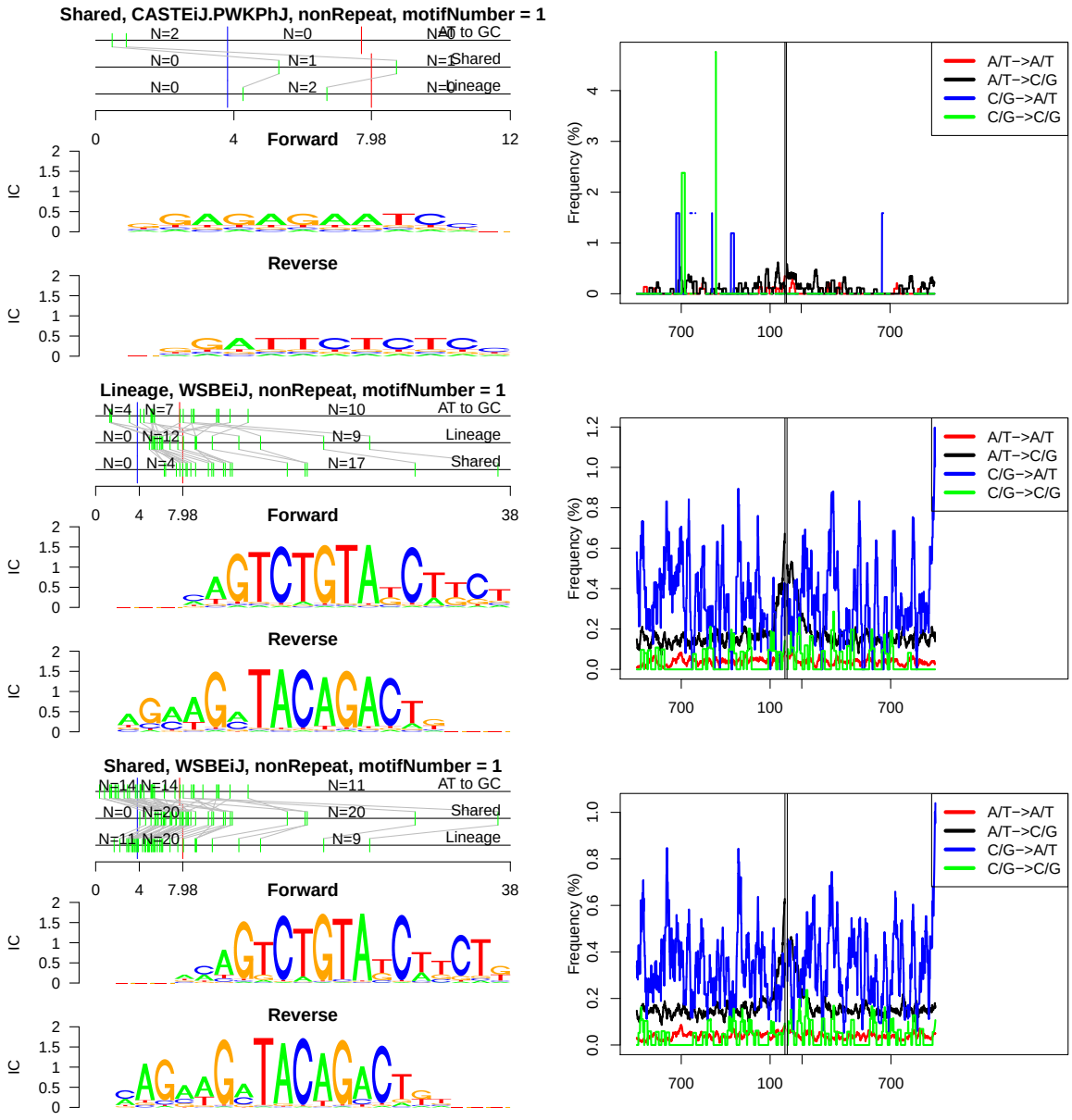

Lineage, Spretus, nonRepeat, motifNumber = 1
N=165
N=110
N=0
N=160
N=51
N=61
N=232
N=347
N=395
AT to GC
Lineage
Shared
0
322
Forward
Reverse
IC
Shared, Spretus, nonRepeat, motifNumber = 1
N=645
N=271
N=0
N=433
N=538
N=428
N=332
N=815
N=282
AT to GC
Shared
Lineage
0
309
Forward
Reverse
IC
AT to GC, Spretus, nonRepeat, motifNumber = 1
N=142
N=202
N=0
N=225
N=263
N=227
N=352
N=471
N=206
Shared
AT to GC
Lineage
0
7.69
80
Forward
Reverse
IC

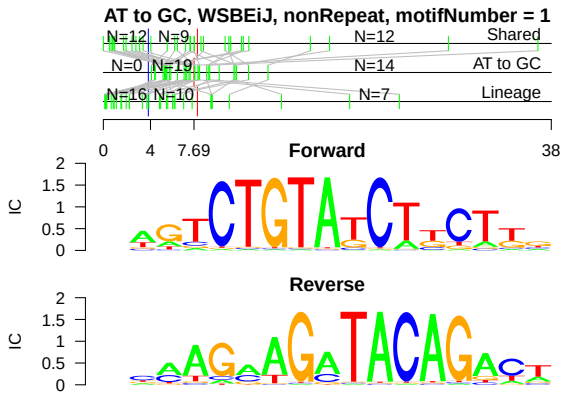

A/T->A/T
A/T->C/G
C/G->A/T
C/G->C/G
Frequency (%)
0.0
0.2
0.4
0.6
0.8
1.0
700
100
700

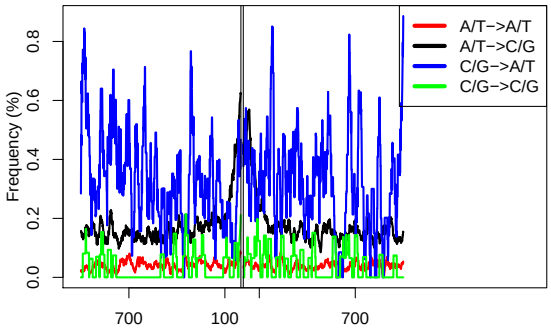

Lineage, AM, nonRepeat, motifNumber = 1
N=286
N=177
N=0
N=190
N=14
N=137
N=96
N=370
N=409
AT to GC
Lineage
Shared
0
144
Forward
Reverse
IC
Shared, AM, nonRepeat, motifNumber = 1
N=473
N=199
N=0
N=277
N=77
N=313
N=98
N=494
N=381
AT to GC, AM, nonRepeat, motifNumber = 1
N=38
N=41
N=0
N=126
N=39
N=43
N=149
N=102
N=146
Frequency (%)
A/T->A/T
A/T->C/G
C/G->A/T
C/G->C/G
700
100
700


Lineage, PWKPhJ, nonRepeat, motifNumber = 1
N=2
N=0
N=0
AT to GC
N=0
N=1
N=1
Lineage
N=1
N=1
N=0
Shared
0
4
Forward
7.98
12
IC
Reverse
Frequency (%)
A/T->A/T
A/T->C/G
C/G->A/T
C/G->C/G
700
100
700
Shared, PWKPhJ, nonRepeat, motifNumber = 1
AT to GC, PWKPhJ, nonRepeat, motifNumber = 1
Forward 7.69

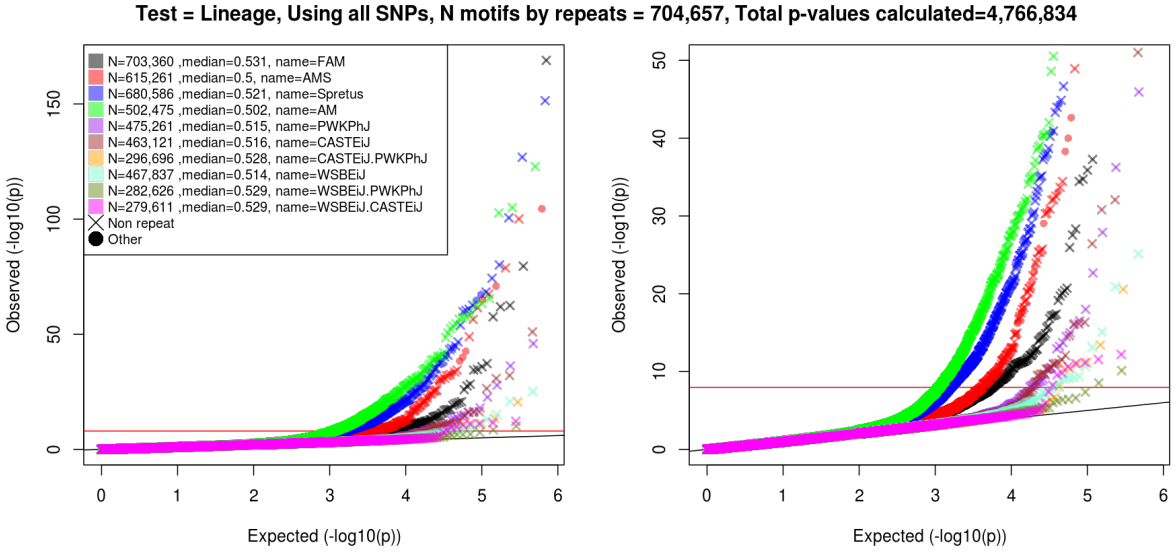

Lineage, CASTEiJ, nonRepeat, motifNumber = 1
N=7
N=4
N=1
AT to GC
N=0
N=8
N=4
Lineage
N=1
N=4
N=7
Shared
0
4
16
Forward
IC
Reverse
Frequency (%)
A/T−>A/T
A/T−>C/G
C/G−>A/T
C/G−>C/G
700
100
700
Shared, CASTEiJ, nonRepeat, motifNumber = 1
N=18
N=6
N=1
AT to GC
N=0
N=17
N=8
Shared
N=11
N=10
N=4
Lineage
AT to GC, CASTEiJ, nonRepeat, motifNumber = 1
N=2
N=0
N=0
Shared
N=0
N=1
N=1
AT to GC
N=2
N=0
N=0
Lineage
7.69
12

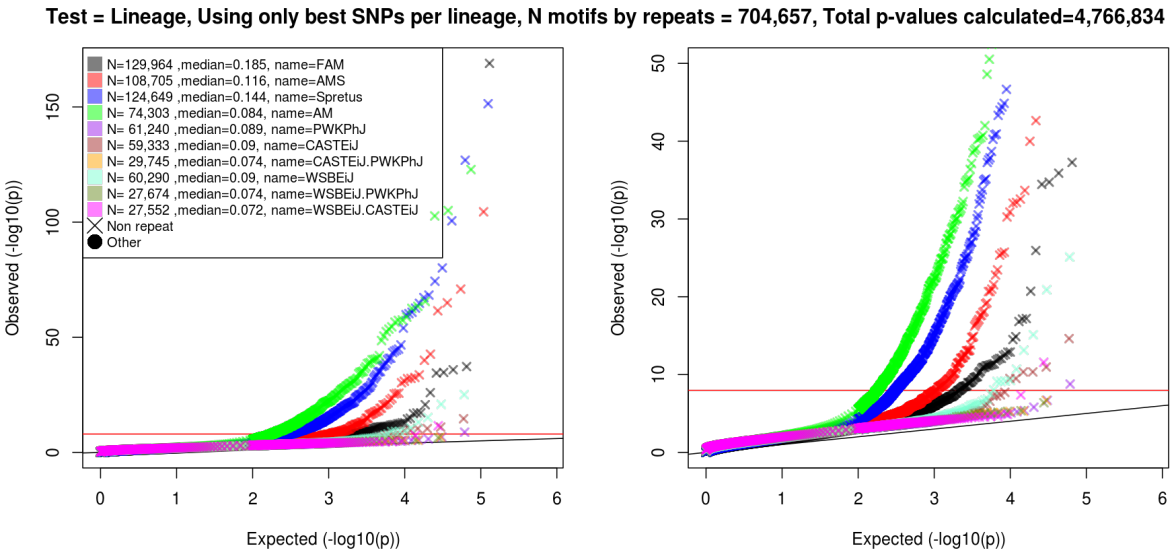

Shared, CASTEiJ.PWKPhJ, nonRepeat, motifNumber = 1
N=2
N=0
N=0
N=0
N=1
N=1
N=0
N=2
N=0
AT to GC
Shared
Lineage
0
4
7.98
12
Forward
Reverse
IC
A/T->A/T
A/T->C/G
C/G->A/T
C/G->C/G
Frequency (%)
700
100
700
Lineage, WSBEiJ, nonRepeat, motifNumber = 1
N=4
N=7
N=0
N=12
N=0
N=4
N=10
N=9
N=17
AT to GC
Lineage
Shared
0
4
7.98
38
Forward
Reverse
Shared, WSBEiJ, nonRepeat, motifNumber = 1
N=14
N=14
N=0
N=20
N=11
N=20
N=11
N=9
AT to GC
Shared
Lineage
0
4
7.98
38
Forward
Reverse

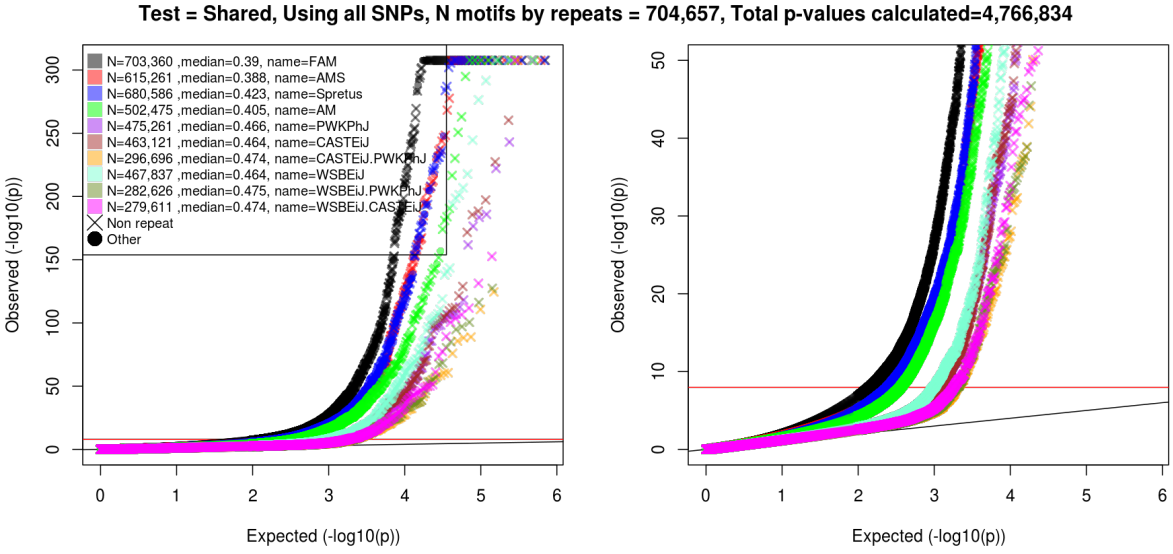

AT to GC, WSBEiJ, nonRepeat, motifNumber = 1
N=12
N=9
N=12
Shared
N=0
N=19
N=14
AT to GC
N=16
N=10
N=7
Lineage
0
4
7.69
38
Forward
IC
Reverse
IC

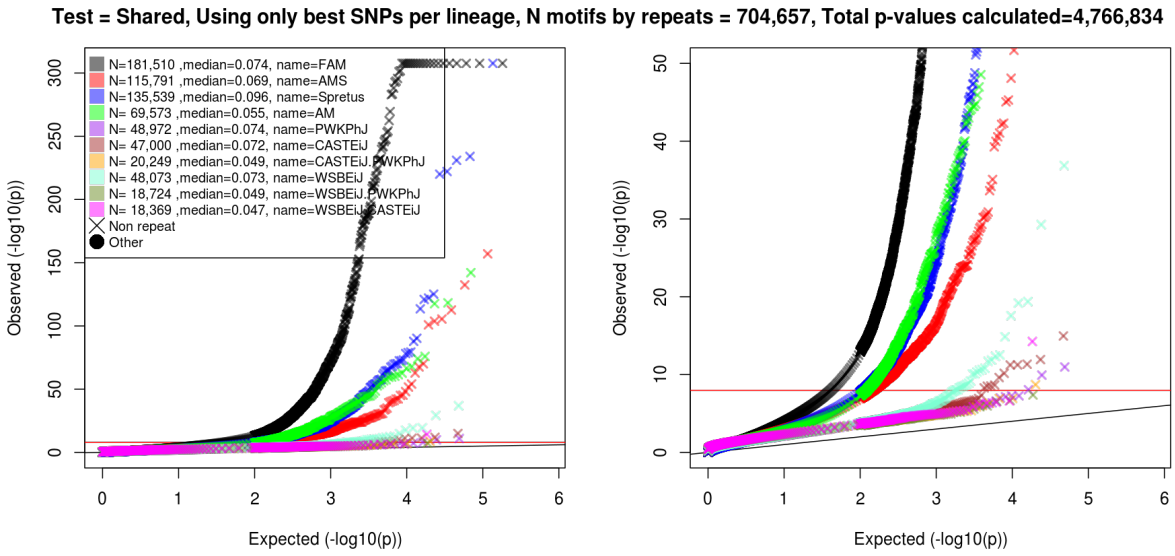

A/T->A/T
A/T->C/G
C/G->A/T
C/G->C/G
Frequency (%)
0.0
0.2
0.4
0.6
0.8
700
100
700

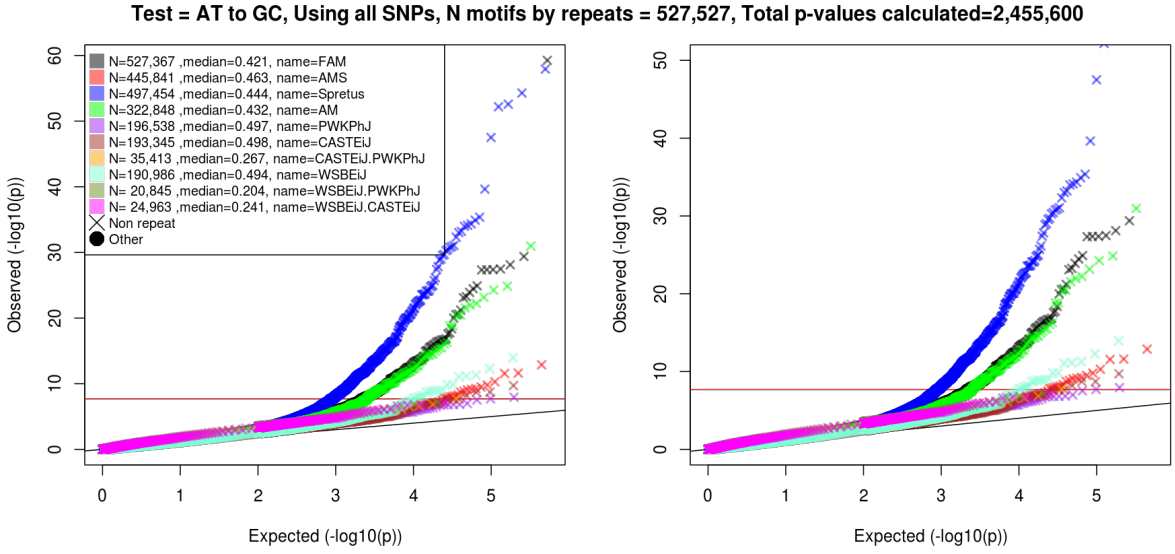
QQ plots

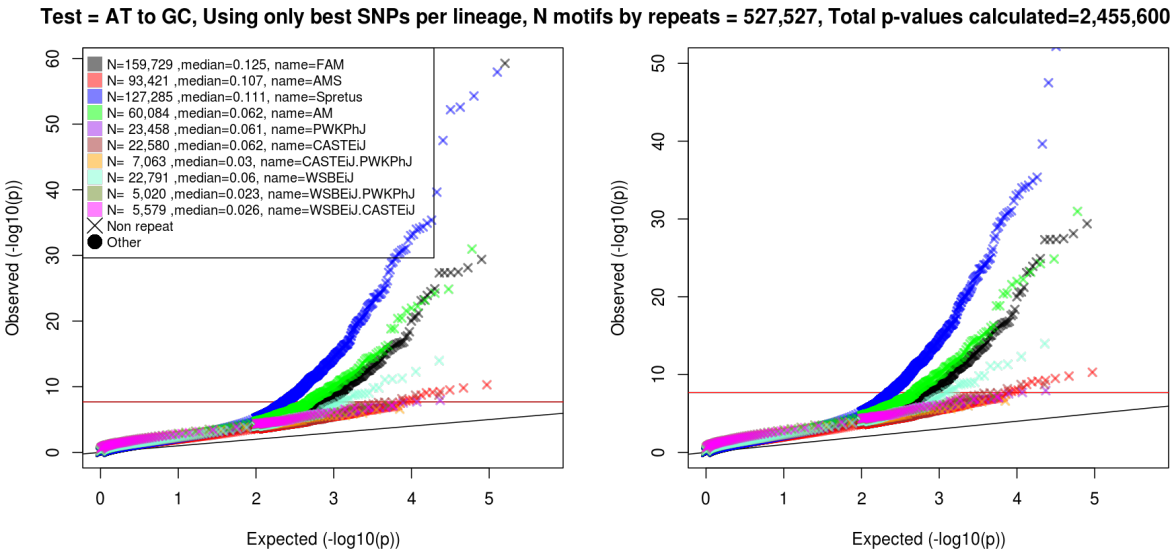

Test = Lineage, Using all SNPs, N motifs by repeats = 704,657, Total p-values calculated=4,766,834
N=703,360 ,median=0.531, name=FAM
N=615,261 ,median=0.5, name=AMS
N=680,586 ,median=0.521, name=Spretus
N=502,475 ,median=0.502, name=AM
N=475,261 ,median=0.515, name=PWKPhJ
N=463,121 ,median=0.516, name=CASTEiJ
N=296,696 ,median=0.528, name=CASTEiJ.PWKPhJ
N=467,837 ,median=0.514, name=WSBEiJ
N=282,626 ,median=0.529, name=WSBEiJ.PWKPhJ
N=279,611 ,median=0.529, name=WSBEiJ.CASTEiJ
Non repeat
Other
Observed (-log10(p))
Expected (-log10(p))
Observed (-log10(p))
Expected (-log10(p))


Test = Lineage, Using only best SNPs per lineage, N motifs by repeats = 704,657, Total p-values calculated=4,766,834
N=129,964 ,median=0.185, name=FAM
N=108,705 ,median=0.116, name=AMS
N=124,649 ,median=0.144, name=Spretus
N= 74,303 ,median=0.084, name=AM
N= 61,240 ,median=0.089, name=PWKPhJ
N= 59,333 ,median=0.09, name=CASTEiJ
N= 29,745 ,median=0.074, name=CASTEiJ.PWKPhJ
N= 60,290 ,median=0.09, name=WSBEiJ
N= 27,674 ,median=0.074, name=WSBEiJ.PWKPhJ
N= 27,552 ,median=0.072, name=WSBEiJ.CASTEiJ
Non repeat
Other
Observed (-log10(p))
Expected (-log10(p))

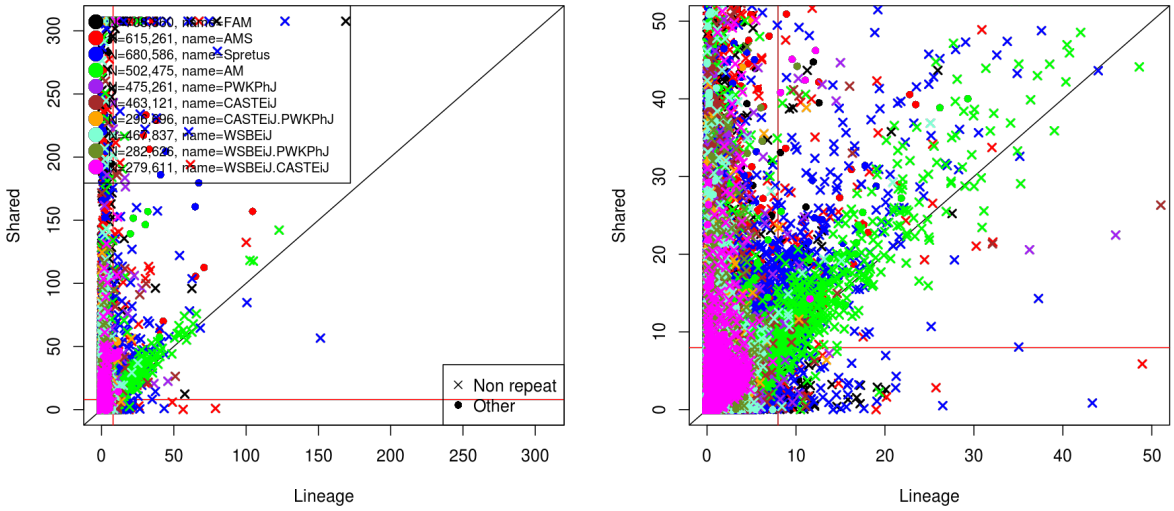

Test = Shared, Using all SNPs, N motifs by repeats = 704,657, Total p-values calculated=4,766,834
N=703,360 ,median=0.39, name=FAM
N=615,261 ,median=0.388, name=AMS
N=680,586 ,median=0.423, name=Spretus
N=502,475 ,median=0.405, name=AM
N=475,261 ,median=0.466, name=PWKPhJ
N=463,121 ,median=0.464, name=CASTEiJ
N=296,696 ,median=0.474, name=CASTEiJ.PWKPhJ
N=467,837 ,median=0.464, name=WSBEiJ
N=282,626 ,median=0.475, name=WSBEiJ.PWKPhJ
N=279,611 ,median=0.474, name=WSBEiJ.CASTEiJ
Non repeat
Other
Observed (-log10(p))
Expected (-log10(p))

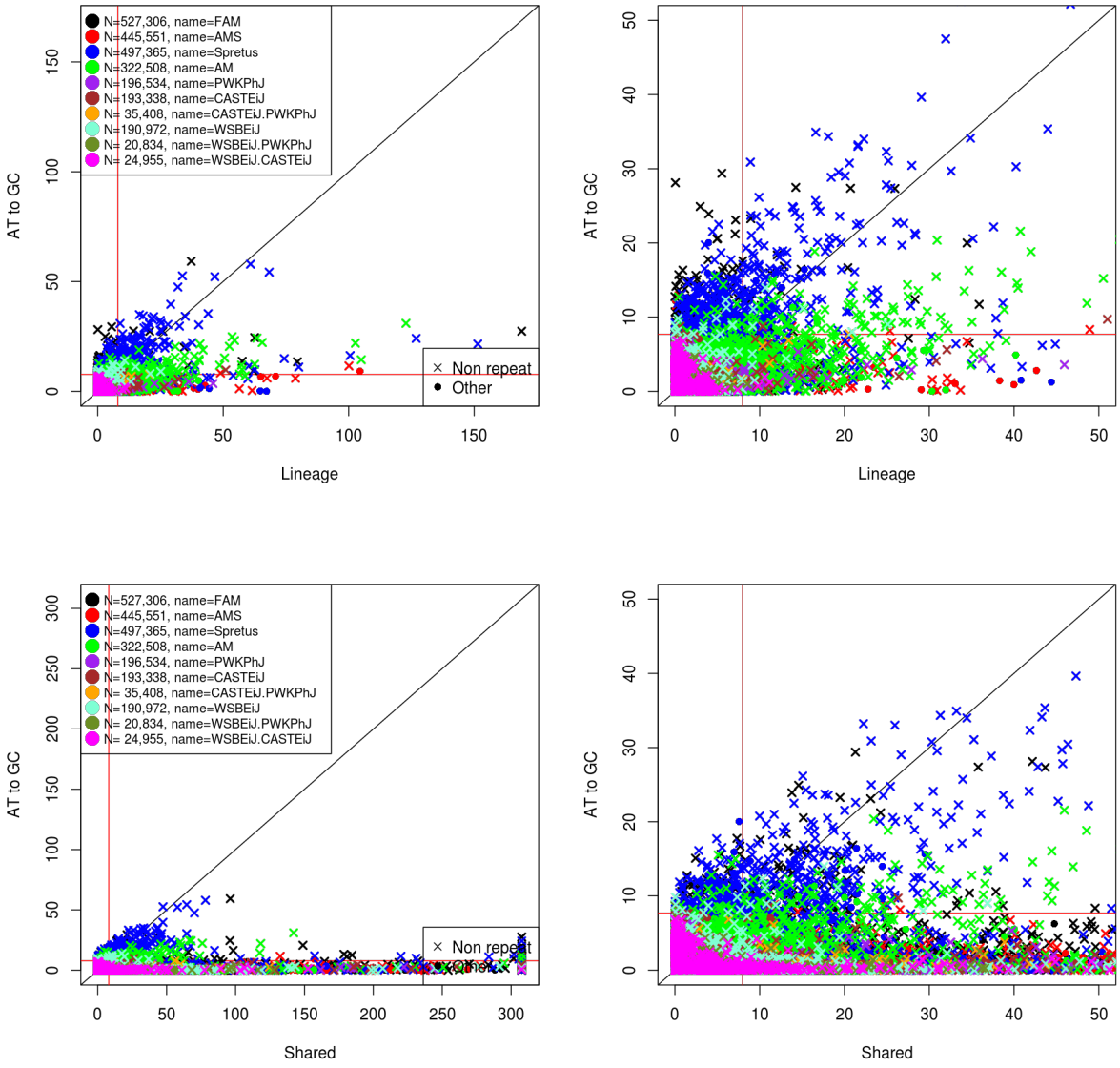

Test = Shared, Using only best SNPs per lineage, N motifs by repeats = 704,657, Total p-values calculated=4,766,834
N=181,510 ,median=0.074, name=FAM
N=115,791 ,median=0.069, name=AMS
N=135,539 ,median=0.096, name=Spretus
N= 69,573 ,median=0.055, name=AM
N= 48,972 ,median=0.074, name=PWKPhJ
N= 47,000 ,median=0.072, name=CASTEiJ
N= 20,249 ,median=0.049, name=CASTEiJ.PWKPhJ
N= 48,073 ,median=0.073, name=WSBEiJ
N= 18,724 ,median=0.049, name=WSBEiJ.PWKPhJ
N= 18,369 ,median=0.047, name=WSBEiJ.CASTEiJ
Non repeat
Other
Observed (-log10(p))
Expected (-log10(p))


Test = AT to GC, Using all SNPs, N motifs by repeats = 527,527, Total p-values calculated=2,455,600
N=527,367 ,median=0.421, name=FAM
N=445,841 ,median=0.463, name=AMS
N=497,454 ,median=0.444, name=Spretus
N=322,848 ,median=0.432, name=AM
N=196,538 ,median=0.497, name=PWKPhJ
N=193,345 ,median=0.498, name=CASTEiJ
N= 35,413 ,median=0.267, name=CASTEiJ.PWKPhJ
N=190,986 ,median=0.494, name=WSBEiJ
N= 20,845 ,median=0.204, name=WSBEiJ.PWKPhJ
N= 24,963 ,median=0.241, name=WSBEiJ.CASTEiJ
Non repeat
Other
Observed (-log10(p))
Expected (-log10(p))

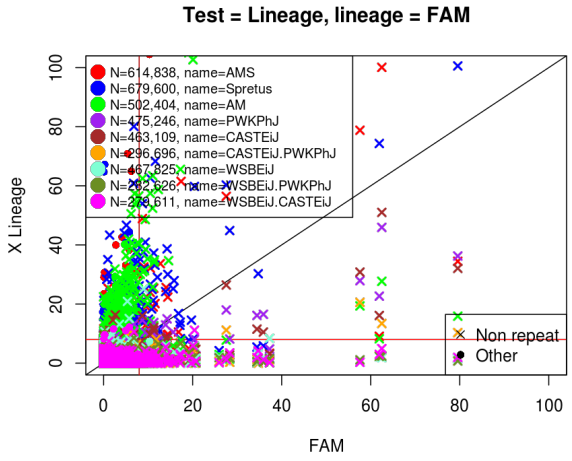


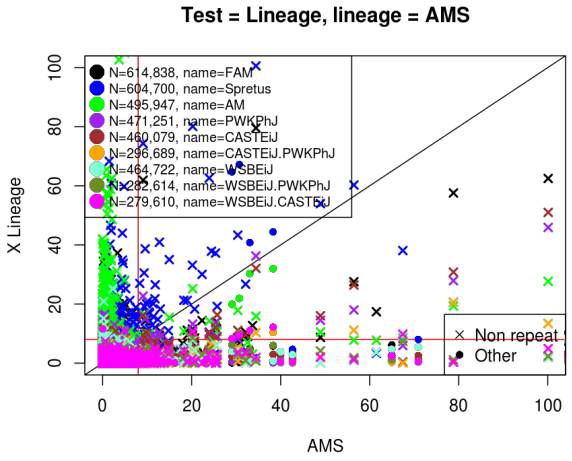
p-values, compare between methods

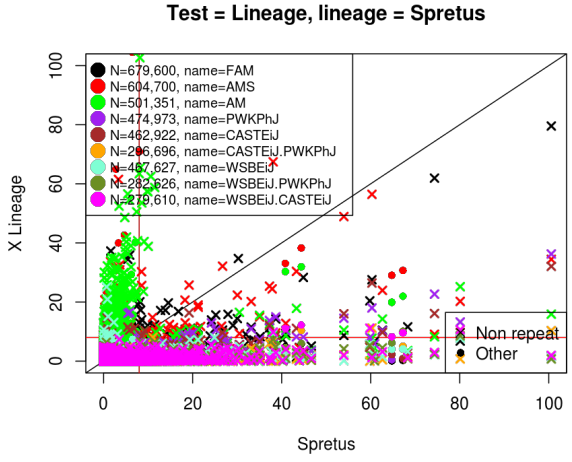


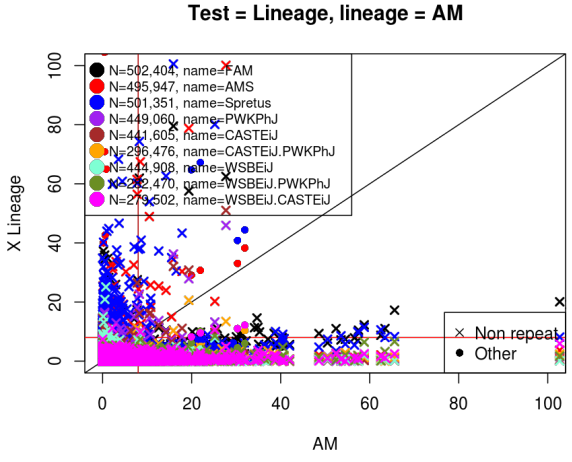


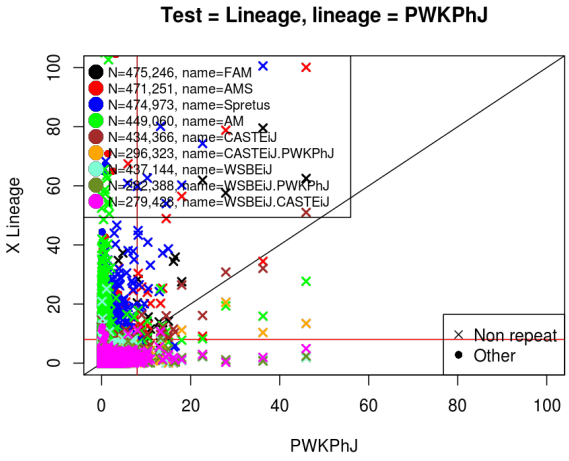
p-values, compare within method, between lineages

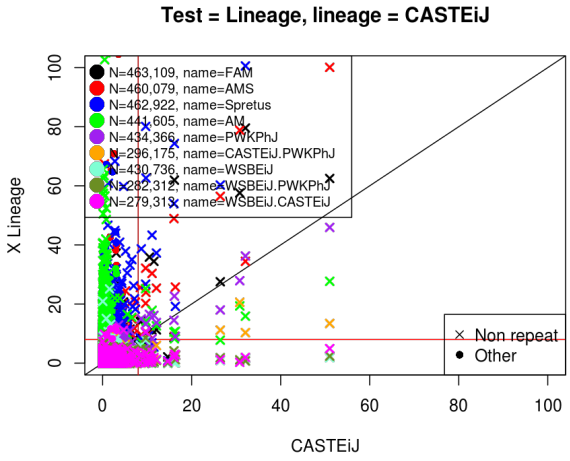

Test = Lineage, lineage = FAM
N=614,838, name=AMS
N=679,600, name=Spretus
N=502,404, name=AM
N=475,246, name=PWKPhJ
N=463,109, name=CASTEiJ
N=296,696, name=CASTEiJ.PWKPhJ
N=467,825, name=WSBEiJ
N=282,626, name=WSBEiJ.PWKPhJ
N=279,611, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
FAM
0
20
40
60
80
100

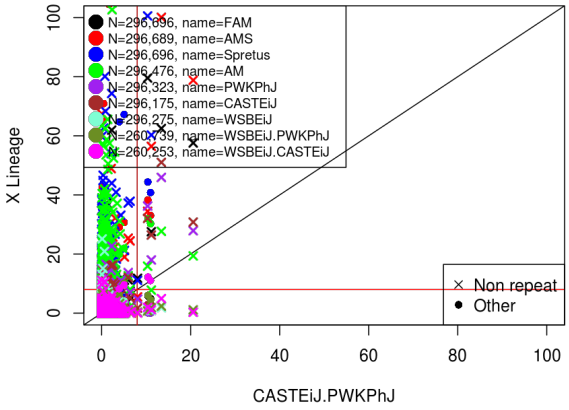

Test = Lineage, lineage = AMS
N=614,838, name=FAM
N=604,700, name=Spretus
N=495,947, name=AM
N=471,251, name=PWKPhJ
N=460,079, name=CASTEiJ
N=296,689, name=CASTEiJ.PWKPhJ
N=464,722, name=WSBEiJ
N=282,614, name=WSBEiJ.PWKPhJ
N=279,610, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
AMS
0
20
40
60
80
100

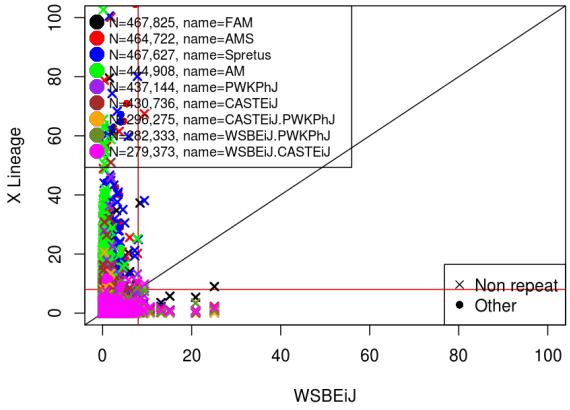

Test = Lineage, lineage = Spretus
N=679,600, name=FAM
N=604,700, name=AMS
N=501,351, name=AM
N=474,973, name=PWKPhJ
N=462,922, name=CASTEiJ
N=296,696, name=CASTEiJ.PWKPhJ
N=467,627, name=WSBEiJ
N=282,626, name=WSBEiJ.PWKPhJ
N=279,610, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
Spretus
0
20
40
60
80
100

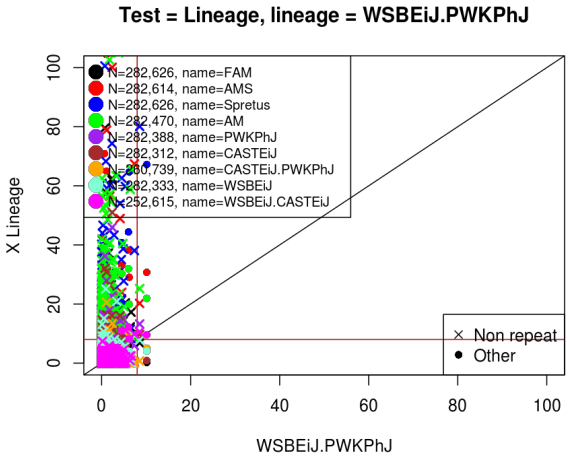

Test = Lineage, lineage = AM
N=502,404, name=FAM
N=495,947, name=AMS
N=501,351, name=Spretus
N=449,060, name=PWKPhJ
N=441,605, name=CASTEiJ
N=296,476, name=CASTEiJ.PWKPhJ
N=444,908, name=WSBEiJ
N=282,470, name=WSBEiJ.PWKPhJ
N=279,502, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
AM
0
20
40
60
80
100

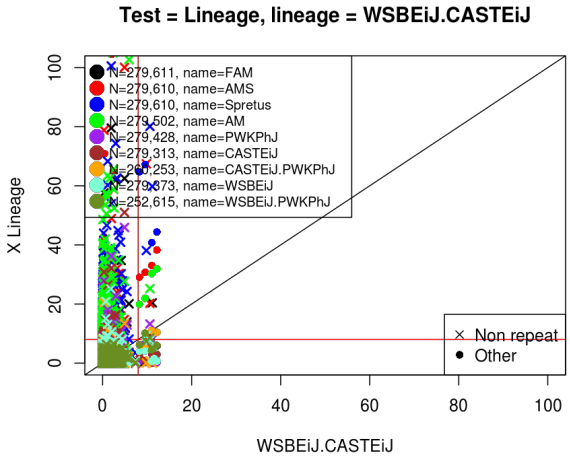

Test = Lineage, lineage = PWKPhJ
N=475,246, name=FAM
N=471,251, name=AMS
N=474,973, name=Spretus
N=449,060, name=AM
N=434,366, name=CASTEiJ
N=296,323, name=CASTEiJ.PWKPhJ
N=437,144, name=WSBEiJ
N=282,388, name=WSBEiJ.PWKPhJ
N=279,438, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
PWKPhJ
0
20
40
60
80
100

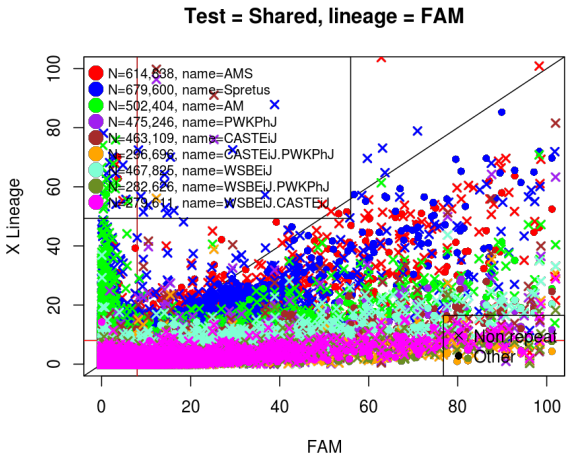

Test = Lineage, lineage = CASTEiJ
N=463,109, name=FAM
N=460,079, name=AMS
N=462,922, name=Spretus
N=441,605, name=AM
N=434,366, name=PWKPhJ
N=296,175, name=CASTEiJ.PWKPhJ
N=430,736, name=WSBEiJ
N=282,312, name=WSBEiJ.PWKPhJ
N=279,31[illegible] name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
CASTEiJ
0
20
40
60
80
100

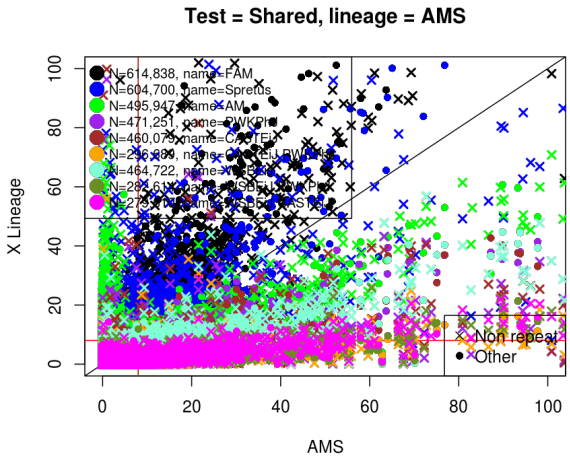

Test = Lineage, lineage = CASTEiJ.PWKPhJ
N=296,696, name=FAM
N=296,689, name=AMS
N=296,696, name=Spretus
N=296,476, name=AM
N=296,323, name=PWKPhJ
N=296,175, name=CASTEiJ
N=296,275, name=WSBEiJ
N=260,739, name=WSBEiJ.PWKPhJ
N=260,253, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
CASTEiJ.PWKPhJ
0
20
40
60
80
100

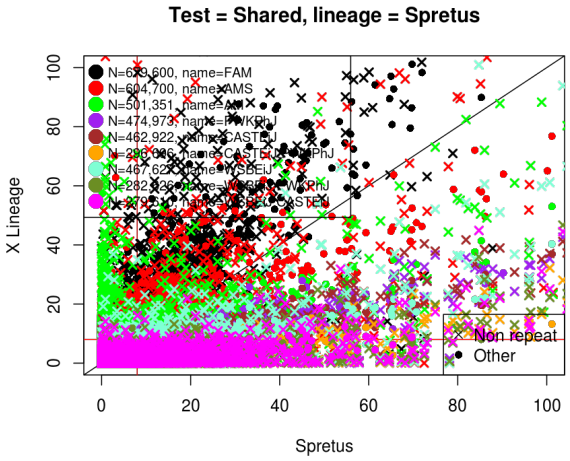

Test = Lineage, lineage = WSBEiJ
N=467,825, name=FAM
N=464,722, name=AMS
N=467,627, name=Spretus
N=444,908, name=AM
N=437,144, name=PWKPhJ
N=430,736, name=CASTEiJ
N=296,275, name=CASTEiJ.PWKPhJ
N=282,333, name=WSBEiJ.PWKPhJ
N=279,373, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
WSBEiJ
0
20
40
60
80
100

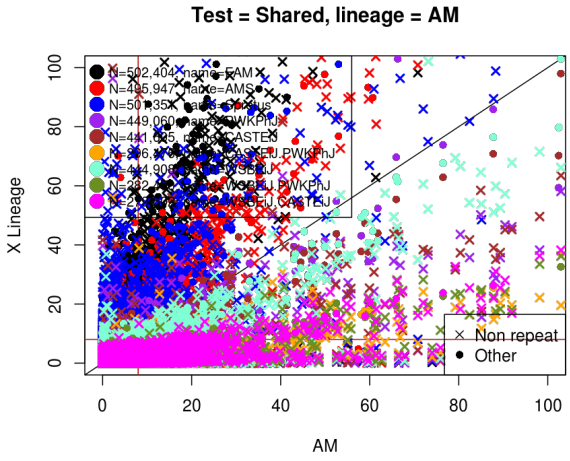

Test = Lineage, lineage = WSBEiJ.PWKPhJ
N=282,626, name=FAM
N=282,614, name=AMS
N=282,626, name=Spretus
N=282,470, name=AM
N=282,388, name=PWKPhJ
N=282,312, name=CASTEiJ
N=280,739, name=CASTEiJ.PWKPhJ
N=282,333, name=WSBEiJ
N=252,615, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
WSBEiJ.PWKPhJ
0
20
40
60
80
100

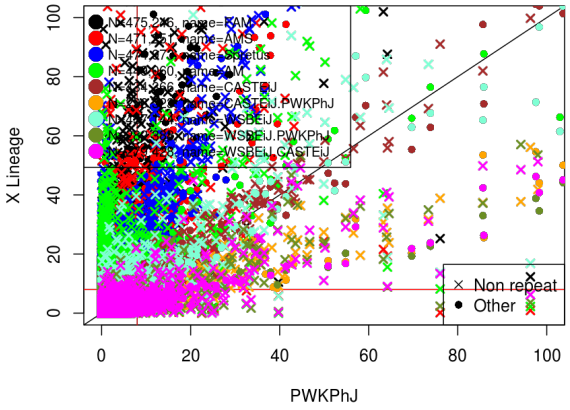

Test = Lineage, lineage = WSBEiJ.CASTEiJ
N=279,611, name=FAM
N=279,610, name=AMS
N=279,610, name=Spretus
N=279,502, name=AM
N=279,428, name=PWKPhJ
N=279,313, name=CASTEiJ
N=280,253, name=CASTEiJ.PWKPhJ
N=279,473, name=WSBEiJ
N=252,615, name=WSBEiJ.PWKPhJ
Non repeat
Other
X Lineage
WSBEiJ.CASTEiJ
0
20
40
60
80
100

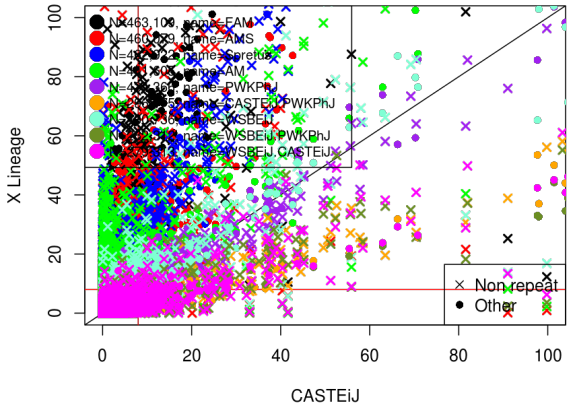

Test = Shared, lineage = FAM
N=614,838, name=AMS
N=679,600, name=Spretus
N=502,404, name=AM
N=475,246, name=PWKPhJ
N=463,109, name=CASTEiJ
N=296,699, name=CASTEiJ.PWKPhJ
N=467,825, name=WSBEiJ
N=282,626, name=WSBEiJ.PWKPhJ
N=279,611, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
FAM
0
20
40
60
80
100

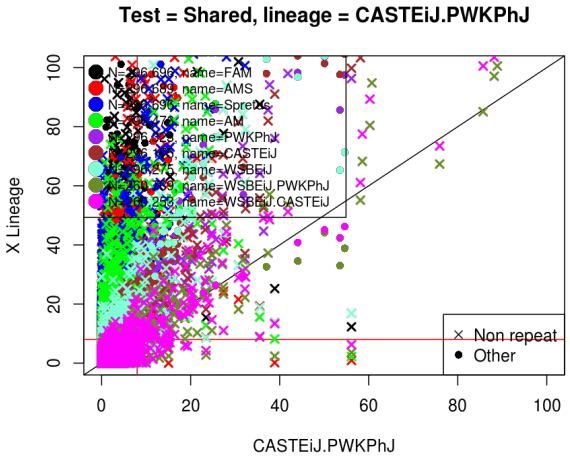

Test = Shared, lineage = AMS
N=614,838, name=FAM
N=604,700, name=Spretus
N=495,947, name=AM
N=471,251, name=PWKPhJ
N=460,075, name=CASTEiJ
N=296,699, name=CASTEiJ.PWKPhJ
N=464,722, name=WSBEiJ
N=282,614, name=WSBEiJ.PWKPhJ
N=279,610, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
AMS
0
20
40
60
80
100

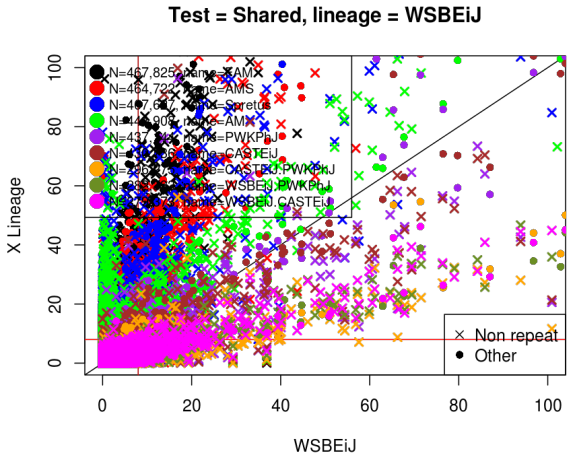

Test = Shared, lineage = Spretus
X Lineage
Spretus
Non repeat
Other

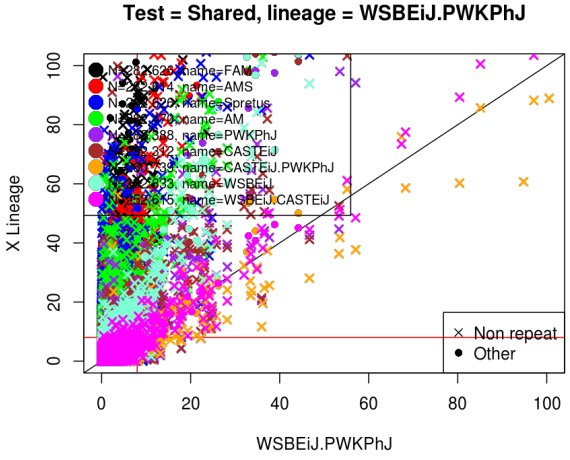

Test = Shared, lineage = AM
X Lineage
AM
Non repeat
Other

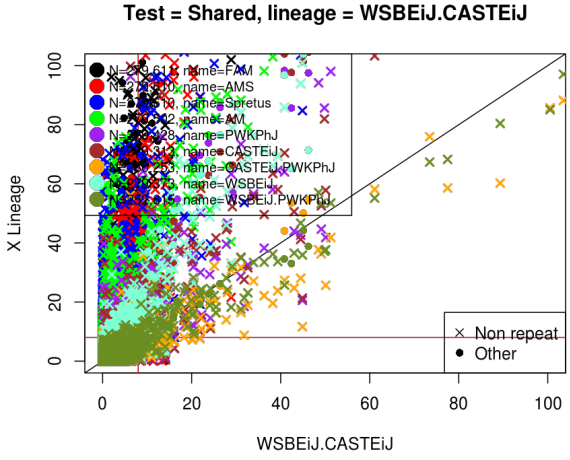

Test = Shared, lineage = PWKPhJ
X Lineage
PWKPhJ
Non repeat
Other

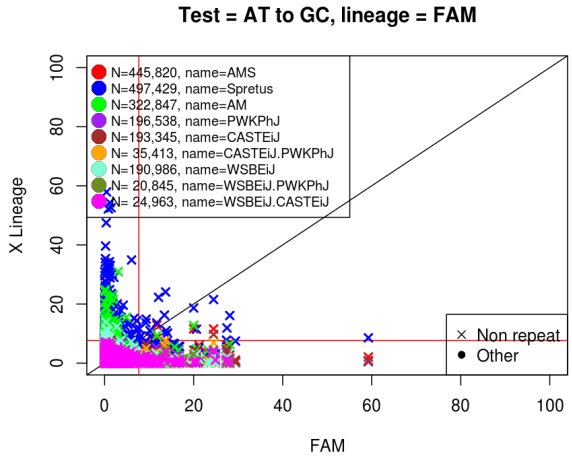

Test = Shared, lineage = CASTEiJ
X Lineage
CASTEiJ
Non repeat
Other

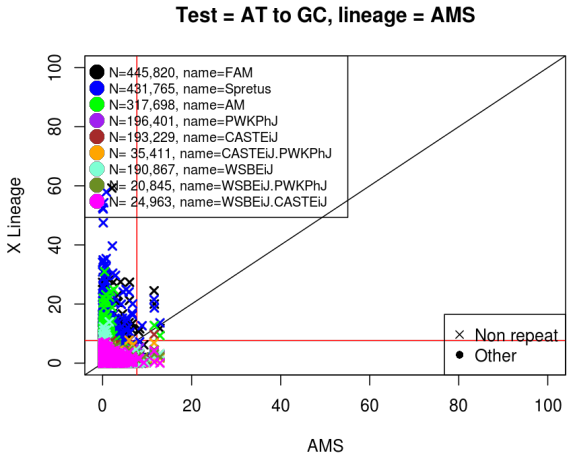

Test = Shared, lineage = CASTEiJ.PWKPhJ
X Lineage
CASTEiJ.PWKPhJ
Non repeat
Other

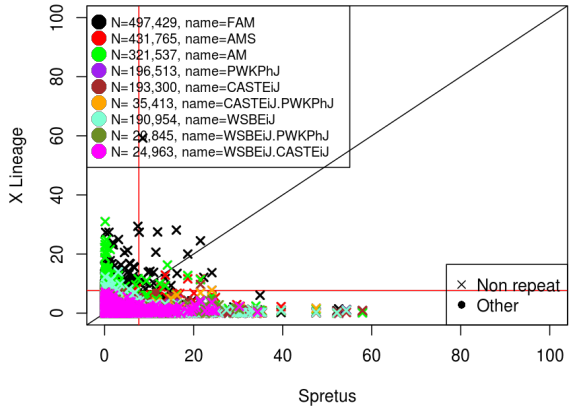

Test = Shared, lineage = WSBEiJ
X Lineage
WSBEiJ
Non repeat
Other

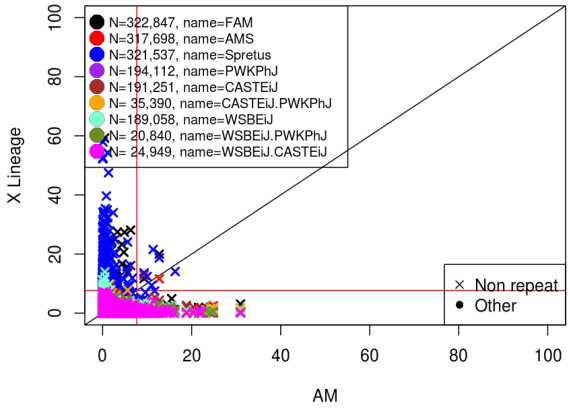

Test = Shared, lineage = WSBEiJ.PWKPhJ
X Lineage
WSBEiJ.PWKPhJ
Non repeat
Other

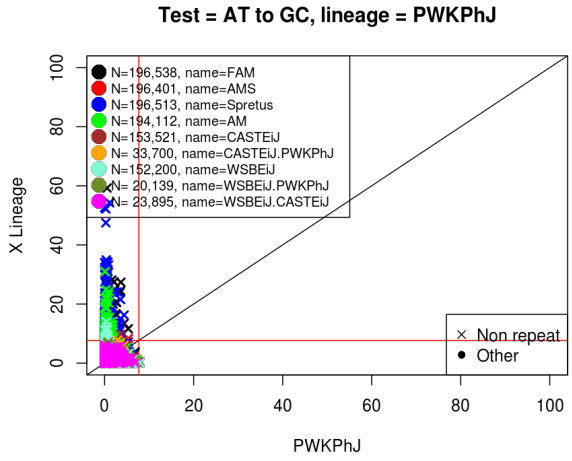

Test = Shared, lineage = WSBEiJ.CASTEiJ
X Lineage
WSBEiJ.CASTEiJ
Non repeat
Other

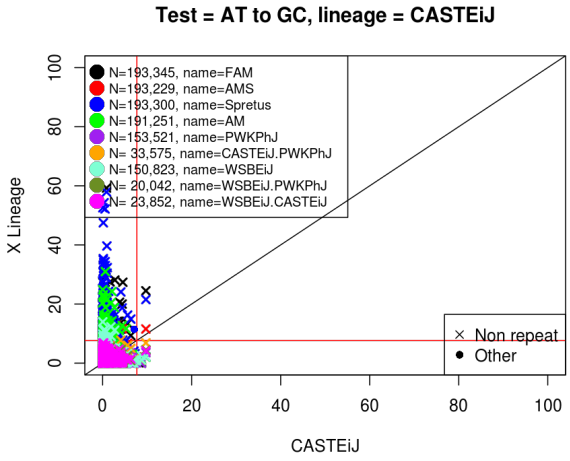

Test = AT to GC, lineage = FAM
N=445,820, name=AMS
N=497,429, name=Spretus
N=322,847, name=AM
N=196,538, name=PWKPhJ
N=193,345, name=CASTEiJ
N= 35,413, name=CASTEiJ.PWKPhJ
N=190,986, name=WSBEiJ
N= 20,845, name=WSBEiJ.PWKPhJ
N= 24,963, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
FAM
0
20
40
60
80
100

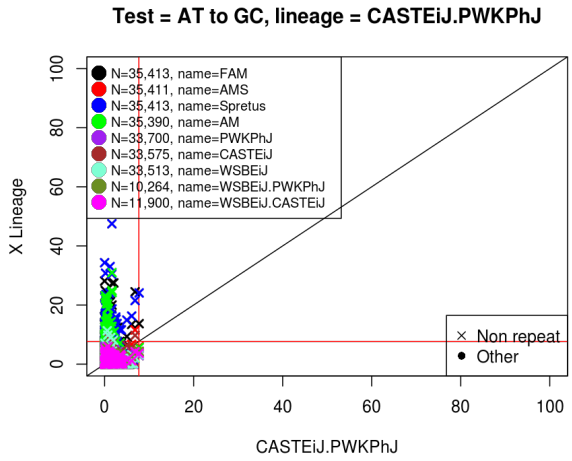

Test = AT to GC, lineage = AMS
N=445,820, name=FAM
N=431,765, name=Spretus
N=317,698, name=AM
N=196,401, name=PWKPhJ
N=193,229, name=CASTEiJ
N= 35,411, name=CASTEiJ.PWKPhJ
N=190,867, name=WSBEiJ
N= 20,845, name=WSBEiJ.PWKPhJ
N= 24,963, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
AMS
0
20
40
60
80
100

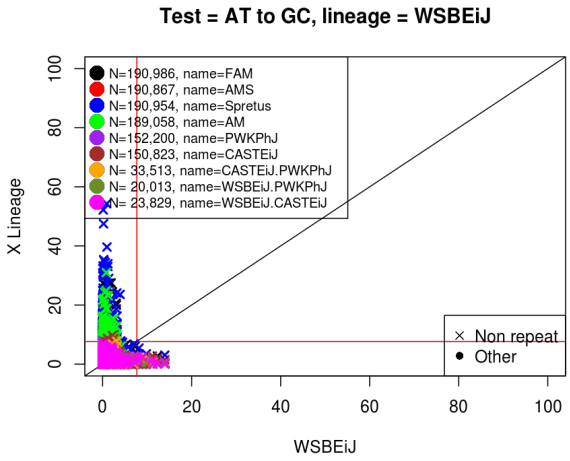

Test = AT to GC, lineage = Spretus
N=497,429, name=FAM
N=431,765, name=AMS
N=321,537, name=AM
N=196,513, name=PWKPhJ
N=193,300, name=CASTEiJ
N= 35,413, name=CASTEiJ.PWKPhJ
N=190,954, name=WSBEiJ
N= 20,845, name=WSBEiJ.PWKPhJ
N= 24,963, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
Spretus
0
20
40
60
80
100

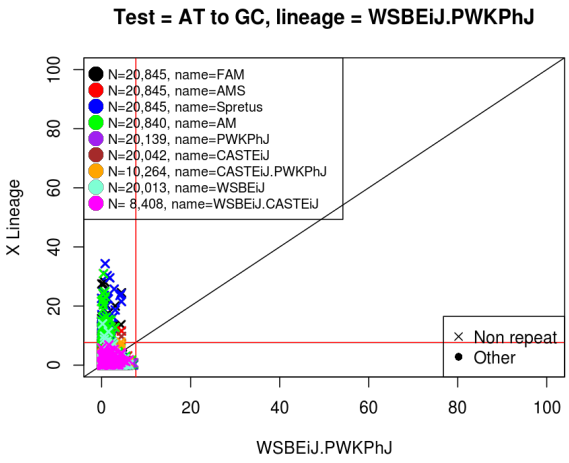

Test = AT to GC, lineage = AM
N=322,847, name=FAM
N=317,698, name=AMS
N=321,537, name=Spretus
N=194,112, name=PWKPhJ
N=191,251, name=CASTEiJ
N= 35,390, name=CASTEiJ.PWKPhJ
N=189,058, name=WSBEiJ
N= 20,840, name=WSBEiJ.PWKPhJ
N= 24,949, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
AM
0
20
40
60
80
100

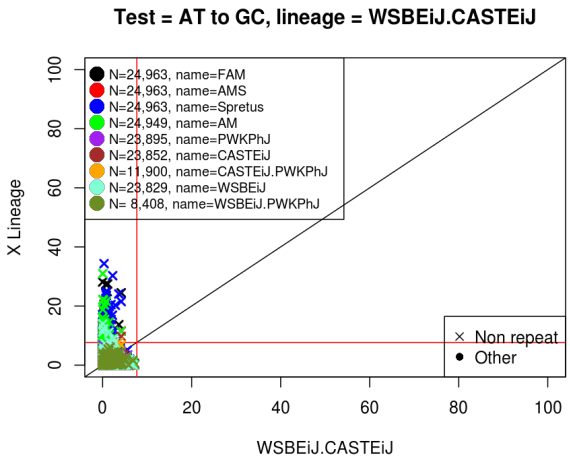

Test = AT to GC, lineage = PWKPhJ
N=196,538, name=FAM
N=196,401, name=AMS
N=196,513, name=Spretus
N=194,112, name=AM
N=153,521, name=CASTEiJ
N= 33,700, name=CASTEiJ.PWKPhJ
N=152,200, name=WSBEiJ
N= 20,139, name=WSBEiJ.PWKPhJ
N= 23,895, name=WSBEiJ.CASTEiJ
× Non repeat
• Other
X Lineage
PWKPhJ
0
20
40
60
80
100


Test = AT to GC, lineage = CASTEiJ
N=193,345, name=FAM
N=193,229, name=AMS
N=193,300, name=Spretus
N=191,251, name=AM
N=153,521, name=PWKPhJ
N= 33,575, name=CASTEiJ.PWKPhJ
N=150,823, name=WSBEiJ
N= 20,042, name=WSBEiJ.PWKPhJ
N= 23,852, name=WSBEiJ.CASTEiJ
× Non repeat
• Other
X Lineage
CASTEiJ
0
20
40
60
80
100


Test = AT to GC, lineage = CASTEiJ.PWKPhJ
N=35,413, name=FAM
N=35,411, name=AMS
N=35,413, name=Spretus
N=35,390, name=AM
N=33,700, name=PWKPhJ
N=33,575, name=CASTEiJ
N=33,513, name=WSBEiJ
N=10,264, name=WSBEiJ.PWKPhJ
N=11,900, name=WSBEiJ.CASTEiJ
× Non repeat
• Other
X Lineage
CASTEiJ.PWKPhJ
0
20
40
60
80
100


Test = AT to GC, lineage = WSBEiJ
N=190,986, name=FAM
N=190,867, name=AMS
N=190,954, name=Spretus
N=189,058, name=AM
N=152,200, name=PWKPhJ
N=150,823, name=CASTEiJ
N= 33,513, name=CASTEiJ.PWKPhJ
N= 20,013, name=WSBEiJ.PWKPhJ
N= 23,829, name=WSBEiJ.CASTEiJ
× Non repeat
• Other
X Lineage
WSBEiJ
0
20
40
60
80
100


Test = AT to GC, lineage = WSBEiJ.PWKPhJ
N=20,845, name=FAM
N=20,845, name=AMS
N=20,845, name=Spretus
N=20,840, name=AM
N=20,139, name=PWKPhJ
N=20,042, name=CASTEiJ
N=10,264, name=CASTEiJ.PWKPhJ
N=20,013, name=WSBEiJ
N= 8,408, name=WSBEiJ.CASTEiJ
Non repeat
Other
X Lineage
WSBEiJ.PWKPhJ
0
20
40
60
80
100


Test = AT to GC, lineage = WSBEiJ.CASTEiJ
N=24,963, name=FAM
N=24,963, name=AMS
N=24,963, name=Spretus
N=24,949, name=AM
N=23,895, name=PWKPhJ
N=23,852, name=CASTEiJ
N=11,900, name=CASTEiJ.PWKPhJ
N=23,829, name=WSBEiJ
N= 8,408, name=WSBEiJ.PWKPhJ
Non repeat
Other
X Lineage
WSBEiJ.CASTEiJ
0
20
40
60
80
100