# DataQuality

*DI Cruz Dávalos*

*7/26/2019*

# Methods

## Samples

The samples included in this study were taken from 24 individuals whose remains belonged to the anthropological collection from the National Museum in Rio de Janeiro, Brazil. Throughout this document we keep the National Museum's identifiers to refer to such individuals. Twenty-two of the individuals were identified from the Museum's archival as belonging to a "Botocudo" group. Similarly, one mummified individual from a cave in Minas Gerais, Brazil, was not associated to any specific group, and one individual was excavated from a shell mound from Santa Catarina, Brazil. Regarding the twenty-two Botocudos, one tooth was sampled for twenty of them, and one piece of skull and petrous bone for the remaining two individuals (MN00019 and MN0008, respectively). One tooth was sampled from the Minas Gerais' mummy and one tooth from the shell mound's individual. Two of the Botocudo individuals presented here (MN00013 and MN00065) have been previously studied by [**?**], but no genomic data was reported at the time.

## DNA extraction, library preparation, and sequencing

```
Twenty-four DNA extracts were prepared: one from petrous bone, one from a piece of skull and twenty-two
```

## Genomic data quality assessment

### Mapping

Remnants of adapters, low-quality bases and nucleotides reported as "N" were trimmed from the reads with AdapterRemoval version 2.1.7 ([**?**]). Reads of 30 bp length and above were mapped to the human genome reference *built 19* with BWA aln version 0.7.15 ([**?**]), disabling the seed to avoid mapping bias due to damage at the 5' termini of the reads ([**?**]). Reads with a mapping quality score equal or greater than 30 were retained. Duplicate reads were identified and removed with picard tools MarkDuplicates version 2.9.0 (http://broadinstitute.github.io/picard/), and indel realignment was performed with GATK version 3.7 with default options ([**?**]). Molecular damage parameters were obtained with mapDamage2 ([**?**]).

### Contamination estimation

```
We estimated contamination using a Bayesian statistical approach on mitochondrial data (\cite{Fu2013}),
```

### Error rate estimation

### Molecular sex determination and uniparental markers

```
Molecular sex was determined by computing the ratio of reads mapping to the Y chromosome with respect t
To call Y-chromosome and mitochondrial haplogroups, we used ANGSD version 0.921 (\cite{Korneliussen2014
```

### Y-chromosome and mitochondrial DNA analyses

# Results

## Processing of genomes and ancient DNA authentication

We shotgun-sequenced 24 samples to an average depth of coverage between 0.001× and 9.2× (Table **??**).

```
<!-- %Between 21,554,888 and 1,124,846,215 reads were sequenced per sample.  -->
```

After trimming adaptors and removing low-quality bases from the reads, between 69.1% and 95.6% of the reads were retained and used as input for mapping.% (that is, between 16,504,949 and 1,030,165,295 reads per sample). The clonality levels (percentage of mapped reads classified as PCR duplicates) within samples ranged from 0.50% to 26.0%. After removing duplicates, we obtained between 70,205 and 533,336,166 reads mapped per sample. Therefore, the percentage of retained reads that were uniquely mapped (i.e., endogenous content) per sample ranged from 0.03% to 51.8%.

The sequenced reads show the common signatures of molecular damage observed in ancient samples, such as short lengths and high rates of deamination (Figure **??**). Retained and mapped reads had similar lengths, with an average of 47.1 - 67.5 bp and 42.0 - 67.1 bp per sample, respectively. Non-USER-treated libraries showed average deamination rates between –% and –% at the termini of the reads.

We estimated contamination based on mitochondrial data. The estimates vary according to the number of reads used in the calculations, which also varies depending on whether we consider transitions as mismatches to the endogenous mitochondrial genome. We notice a larger dispersion in the posterior distribution when using ~2,000 reads or less.

When accounting for all polymorphism types and samples with more than 2,000 mitochondrial reads (n = 19, mitochondrial coverage: 9.9× - 222.4×), the maximum a posteriori estimate for contamination is between 0.69% and 8.41%. If we remove transitions from the estimation, we have 14 samples with more than 2,000 (mitochondrial coverage: 35.6× - 222.4×) for which we estimated between 0.03% and 3.10% of contaminant reads.

Regarding samples with mitochondrial coverage above 10×, we estimated less than 6% of contaminant reads (Table **??**).

## Data quality assessment

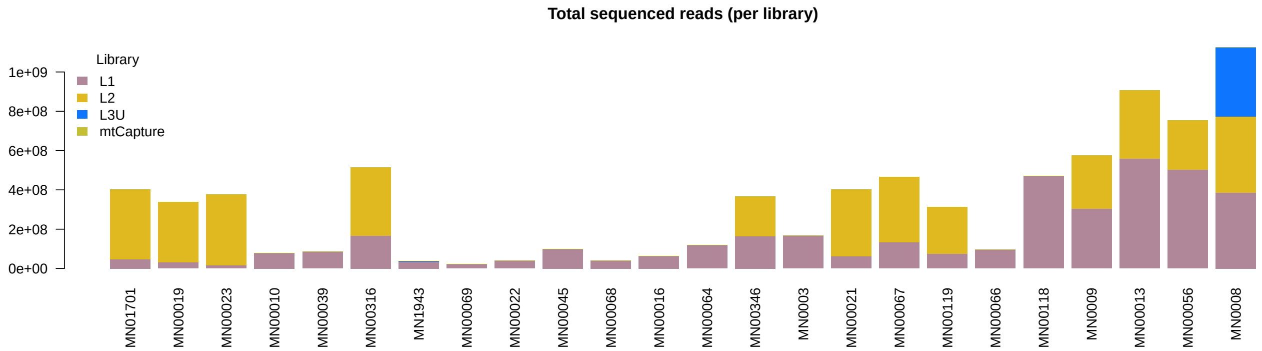
## Molecular damage patterns

## Contamination estimates

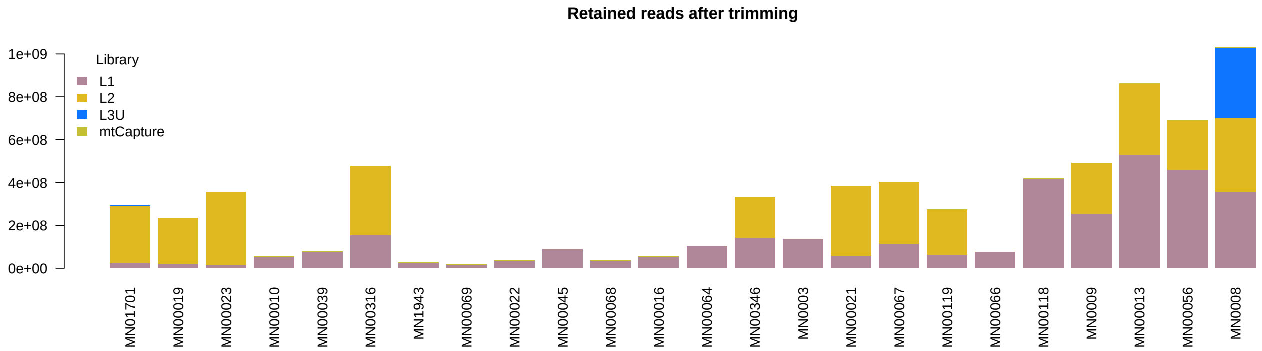
## Molecular sex determination

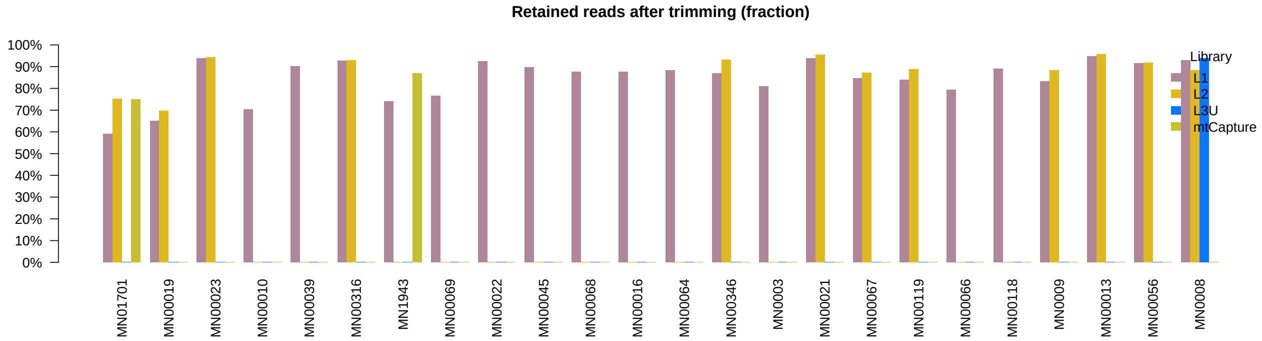
# Sequenced reads

## Total number of reads


Total sequenced reads (per library)

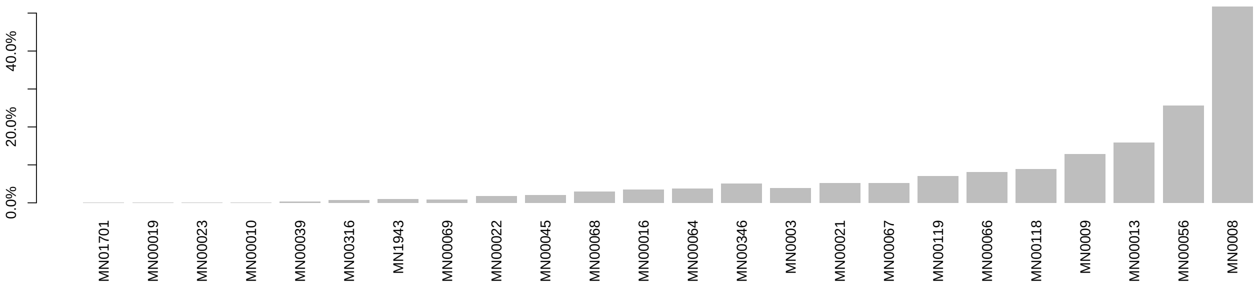
## Reads retained after trimming


Retained reads after trimming

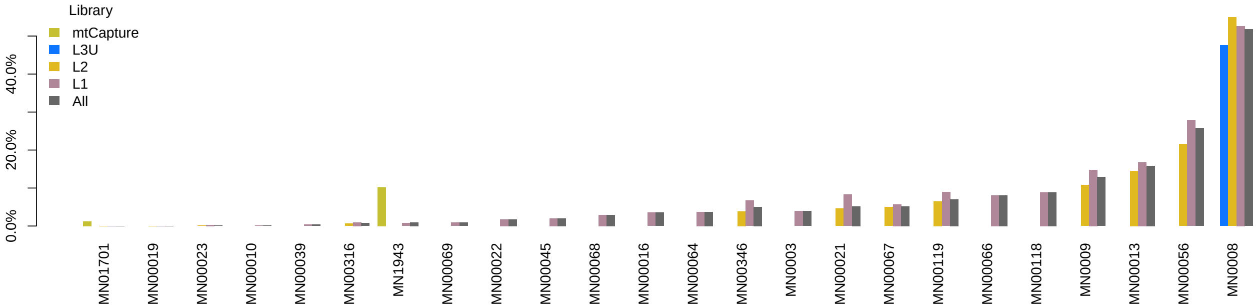

Retained reads after trimming (fraction)

# Endogenous content

Endogenous content per sample

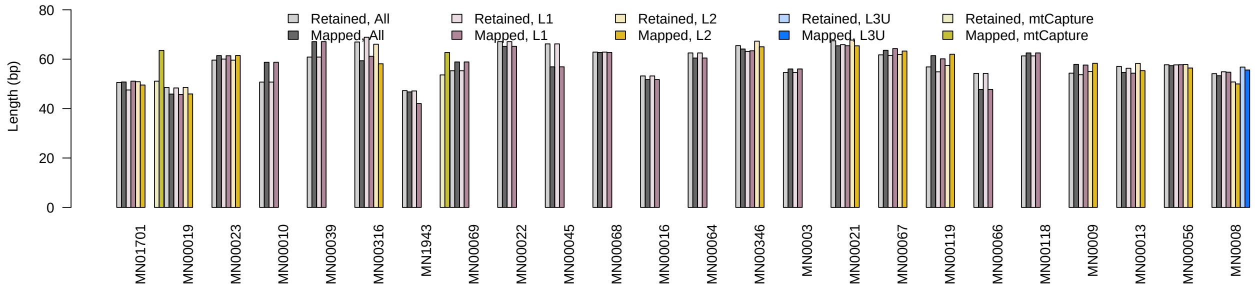


Endogenous content per library

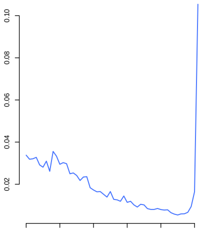


# Read length

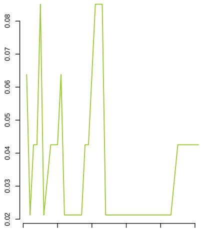
Average read length (retained and mapped)



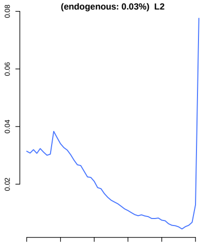
nucMN01701
(endogenous: 0.06%) L1



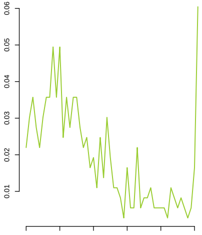
mitoMN01701
(endogenous: 0.06%) L1



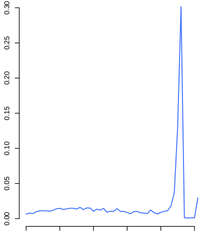
nucMN01701
(endogenous: 0.03%) L2



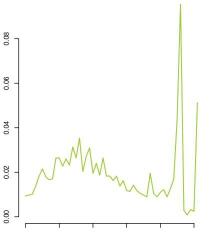
mitoMN01701
(endogenous: 0.03%) L2



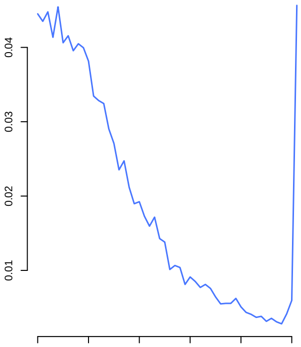
nucMN01701
(endogenous: 1.24%) mtCapture



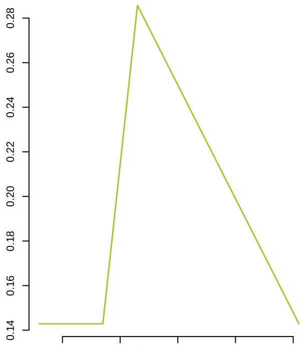
mitoMN01701
(endogenous: 1.24%) mtCapture



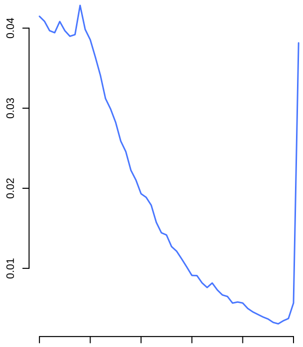
**nucMN00019**
**(endogenous: 0.07%) L1**



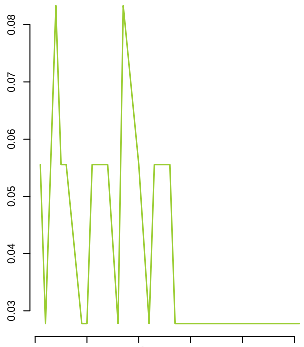
**mitoMN00019**
**(endogenous: 0.07%) L1**



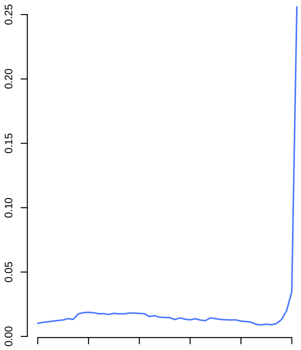
**nucMN00019**
**(endogenous: 0.06%) L2**



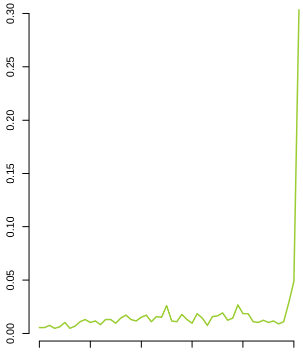
**mitoMN00019**
**(endogenous: 0.06%) L2**



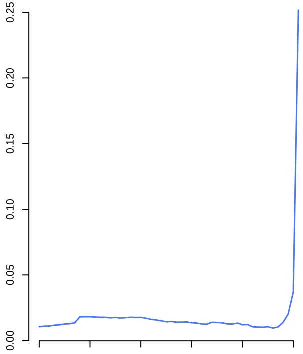
**nucMN00023**
**(endogenous: 0.31%) L1**



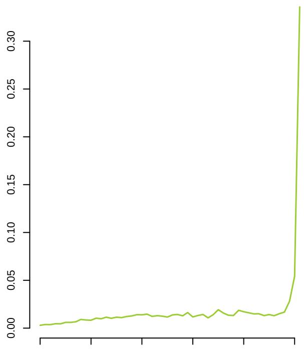
**mitoMN00023**
**(endogenous: 0.31%) L1**



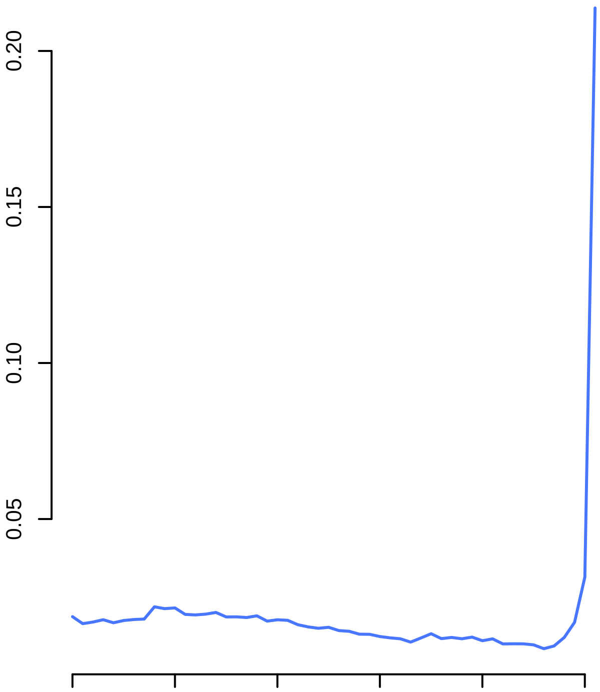
**nucMN00023**
**(endogenous: 0.11%) L2**



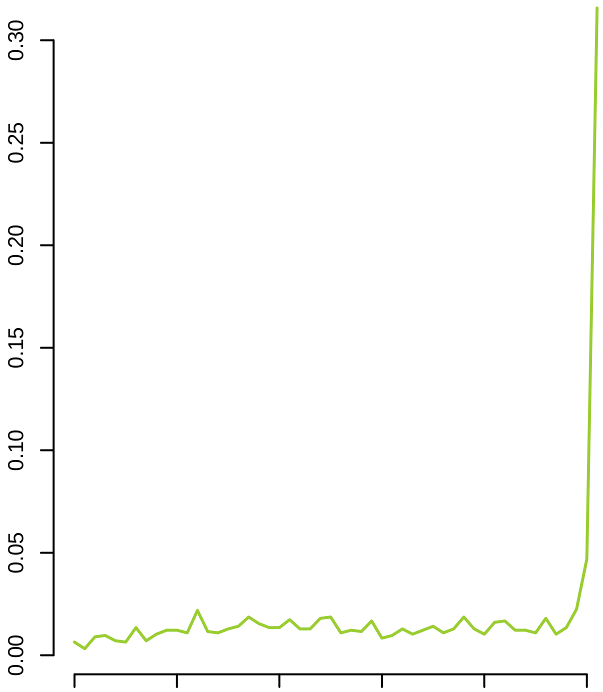
**mitoMN00023**
**(endogenous: 0.11%) L2**

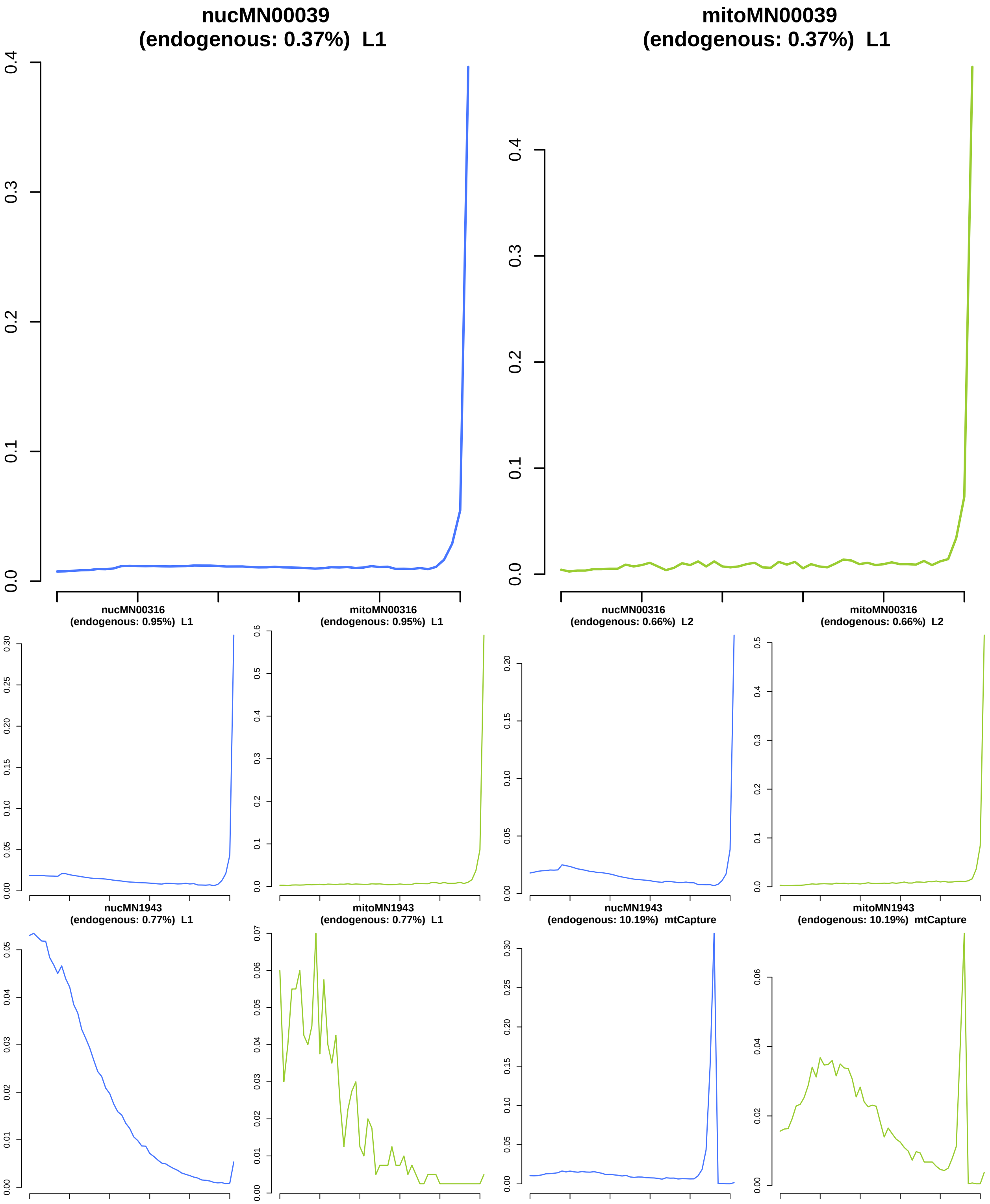


**nucMN00010**
**(endogenous: 0.13%) L1**

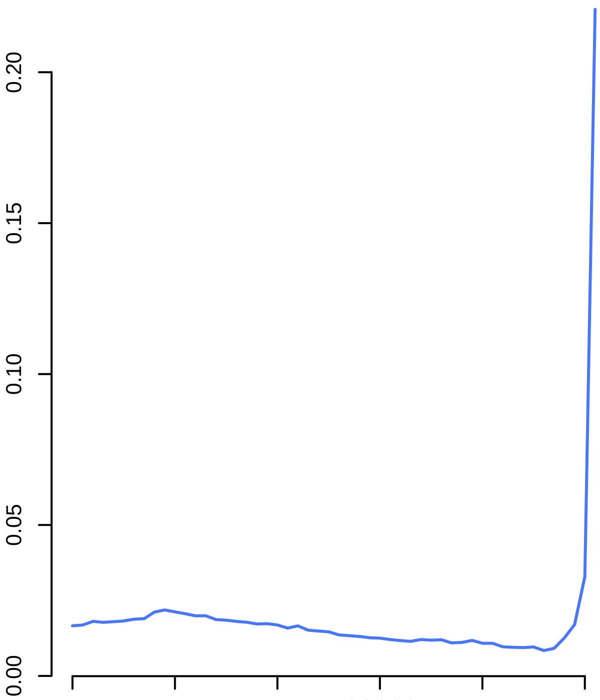


**mitoMN00010**
**(endogenous: 0.13%) L1**

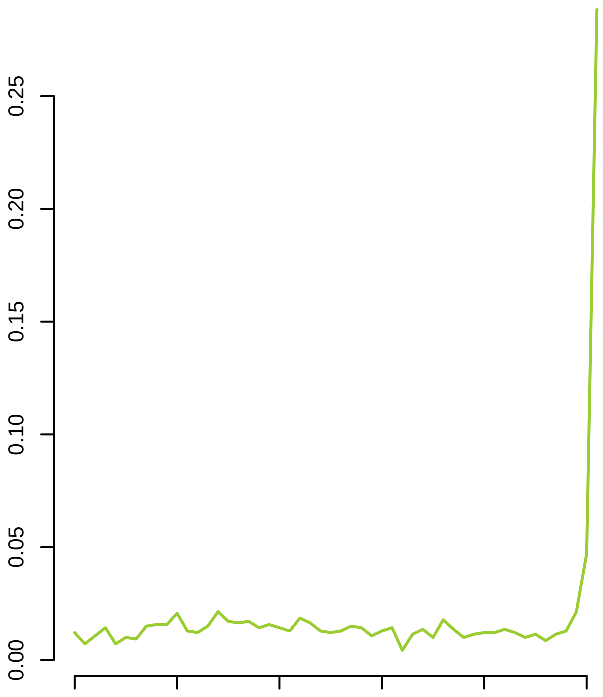



nucMN00039
(endogenous: 0.37%) L1
0.0
0.1
0.2
0.3
0.4
mitoMN00039
(endogenous: 0.37%) L1
0.0
0.1
0.2
0.3
0.4
nucMN00316
(endogenous: 0.95%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
mitoMN00316
(endogenous: 0.95%) L1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
nucMN00316
(endogenous: 0.66%) L2
0.00
0.05
0.10
0.15
0.20
mitoMN00316
(endogenous: 0.66%) L2
0.0
0.1
0.2
0.3
0.4
0.5
nucMN1943
(endogenous: 0.77%) L1
0.00
0.01
0.02
0.03
0.04
0.05
mitoMN1943
(endogenous: 0.77%) L1
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
nucMN1943
(endogenous: 10.19%) mtCapture
0.00
0.05
0.10
0.15
0.20
0.25
0.30
mitoMN1943
(endogenous: 10.19%) mtCapture
0.00
0.02
0.04
0.06

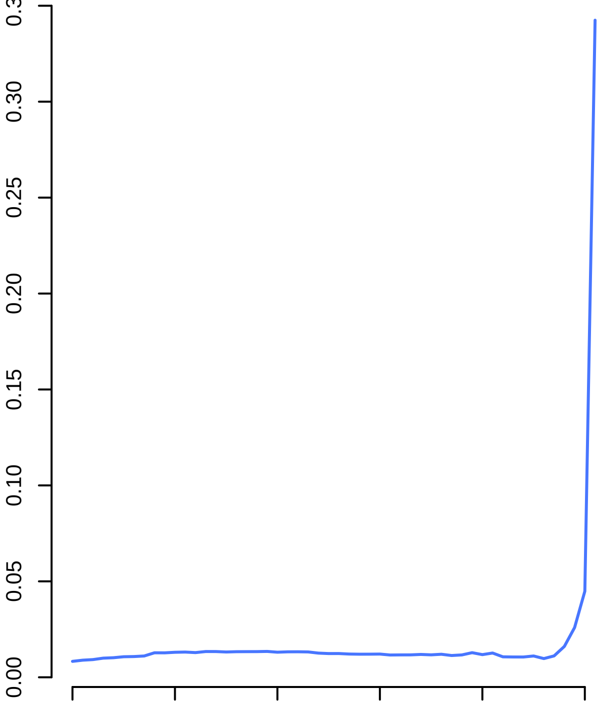
nucMN00069
(endogenous: 0.89%)  L1



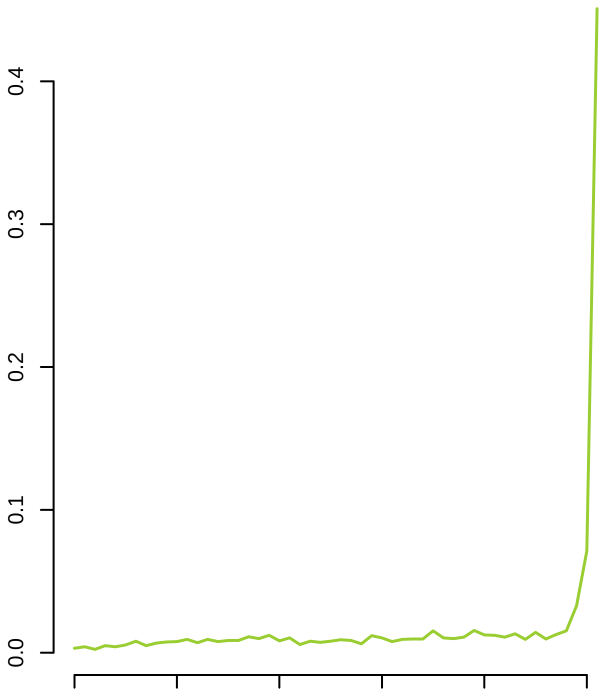
mitoMN00069
(endogenous: 0.89%)  L1

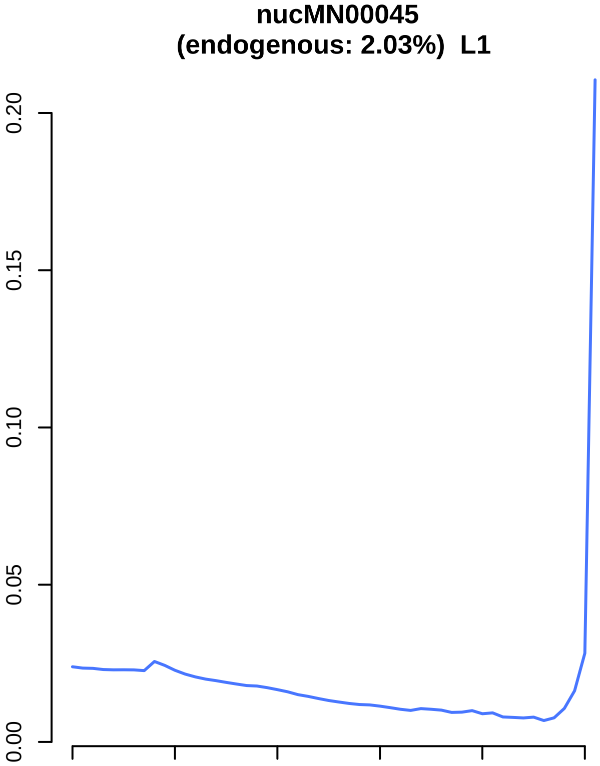


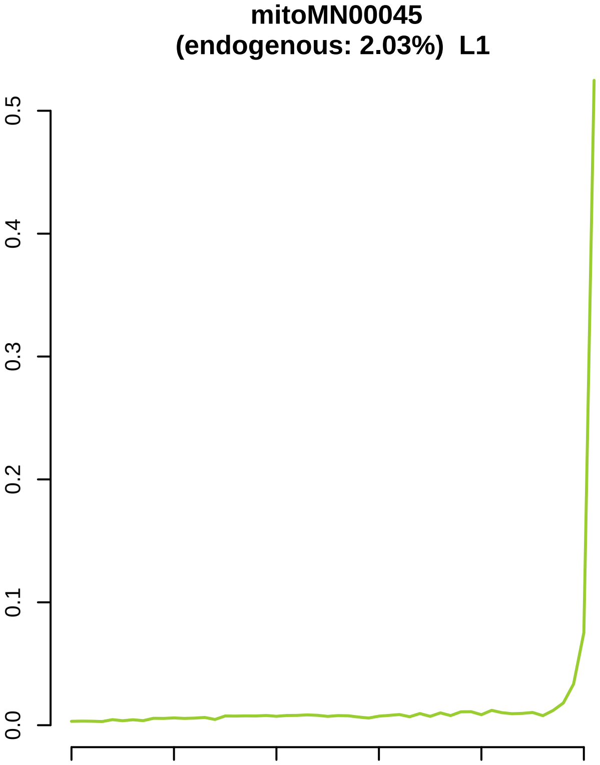
nucMN00022
(endogenous: 1.74%)  L1

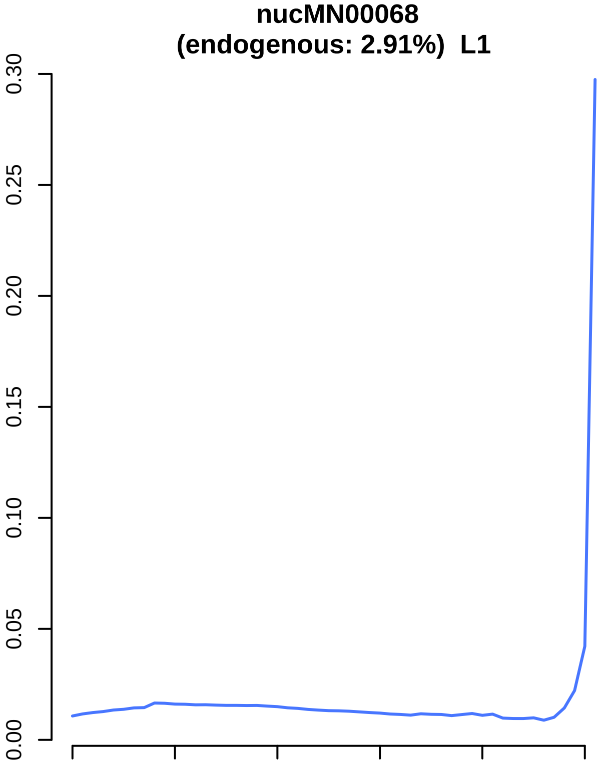


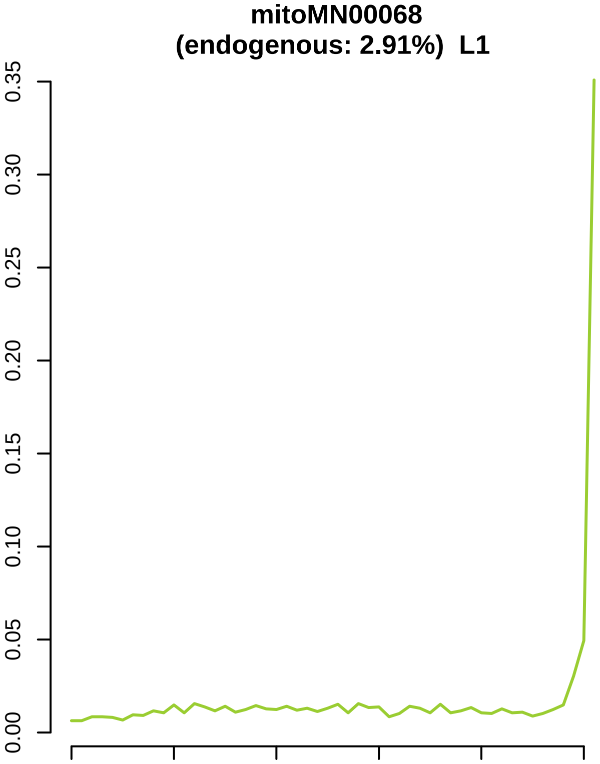
mitoMN00022
(endogenous: 1.74%)  L1

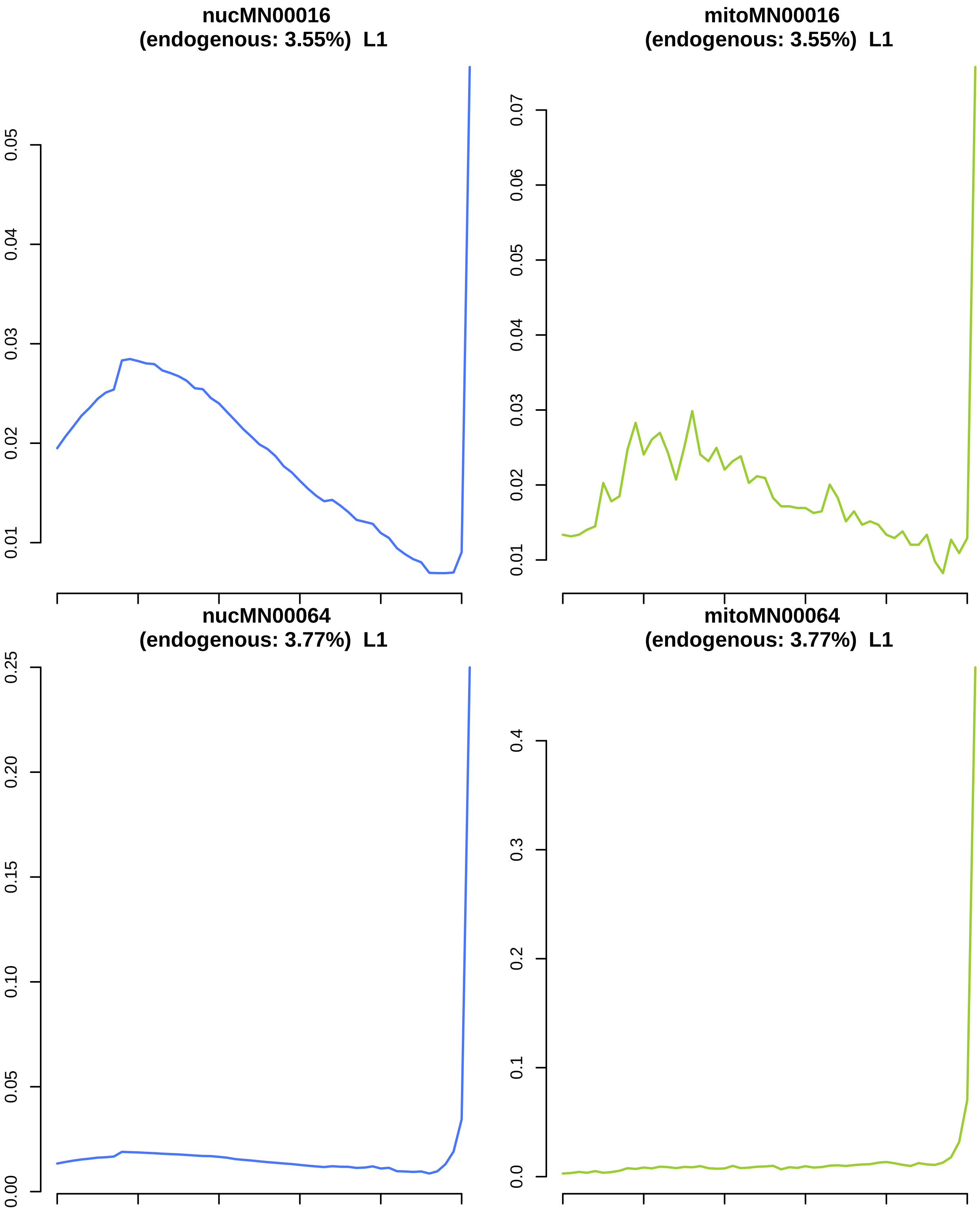



nucMN00045
(endogenous: 2.03%) L1
0.00
0.05
0.10
0.15
0.20


mitoMN00045
(endogenous: 2.03%) L1
0.0
0.1
0.2
0.3
0.4
0.5


nucMN00068
(endogenous: 2.91%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30


mitoMN00068
(endogenous: 2.91%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35

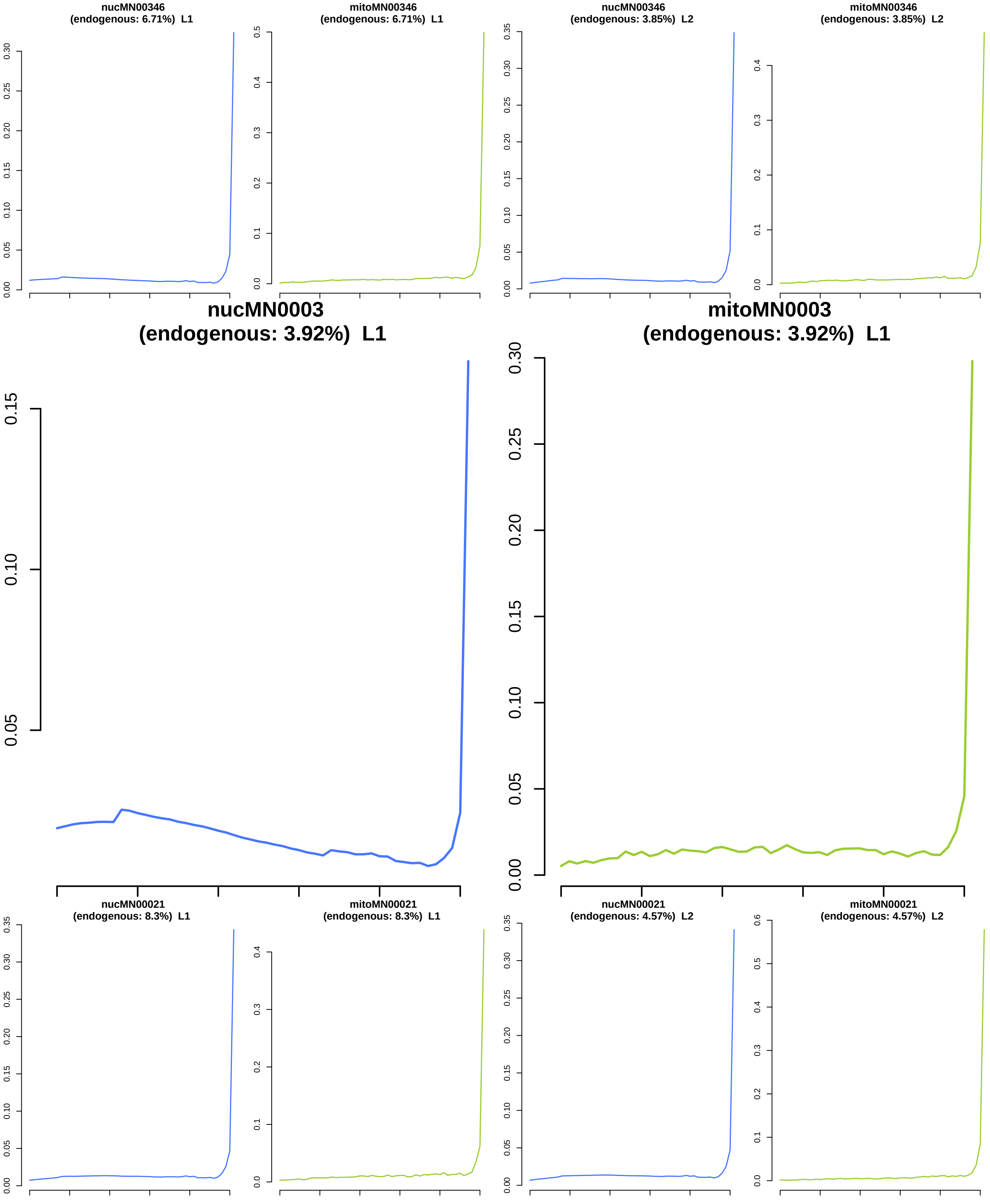


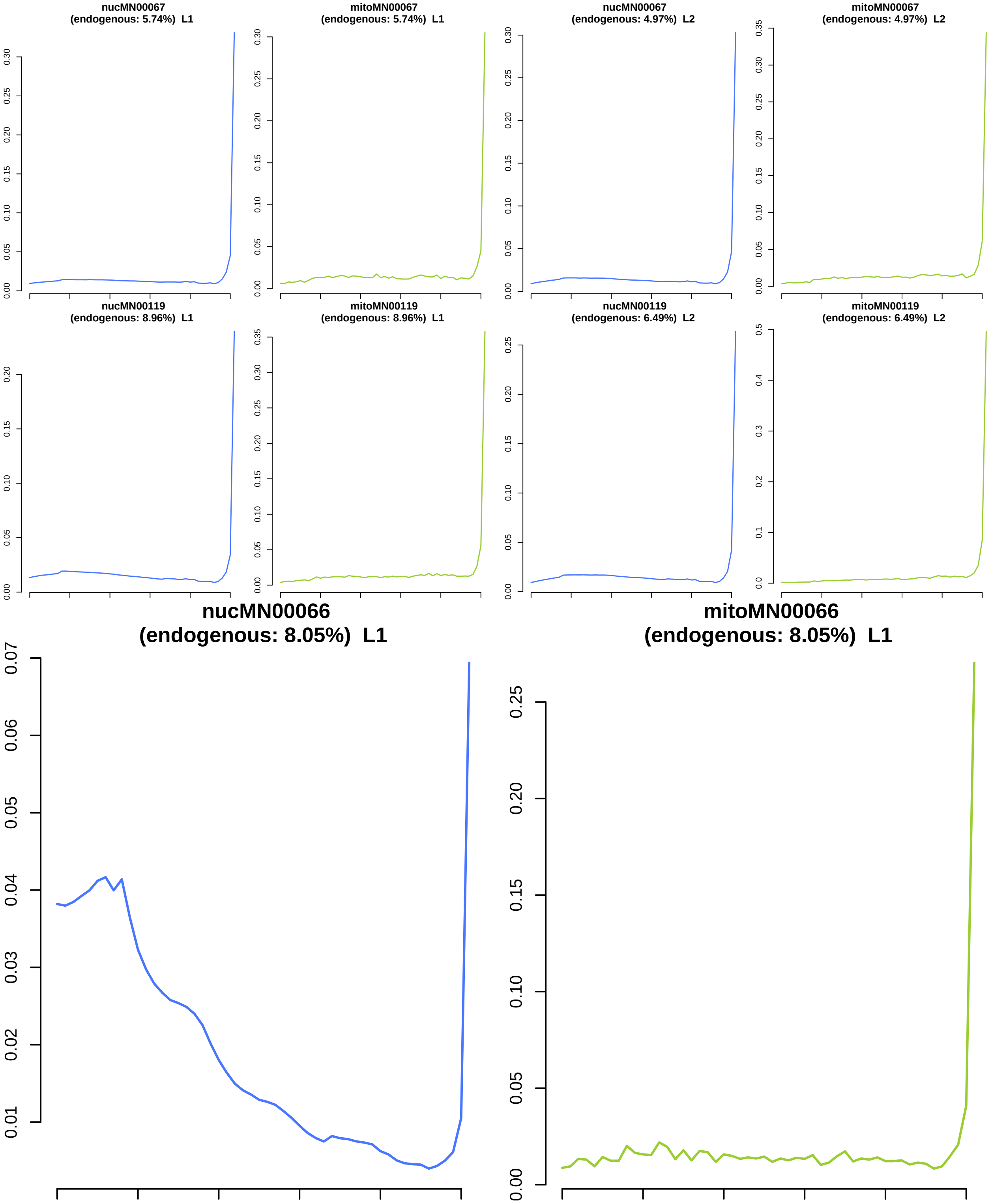
nucMN00016
(endogenous: 3.55%) L1

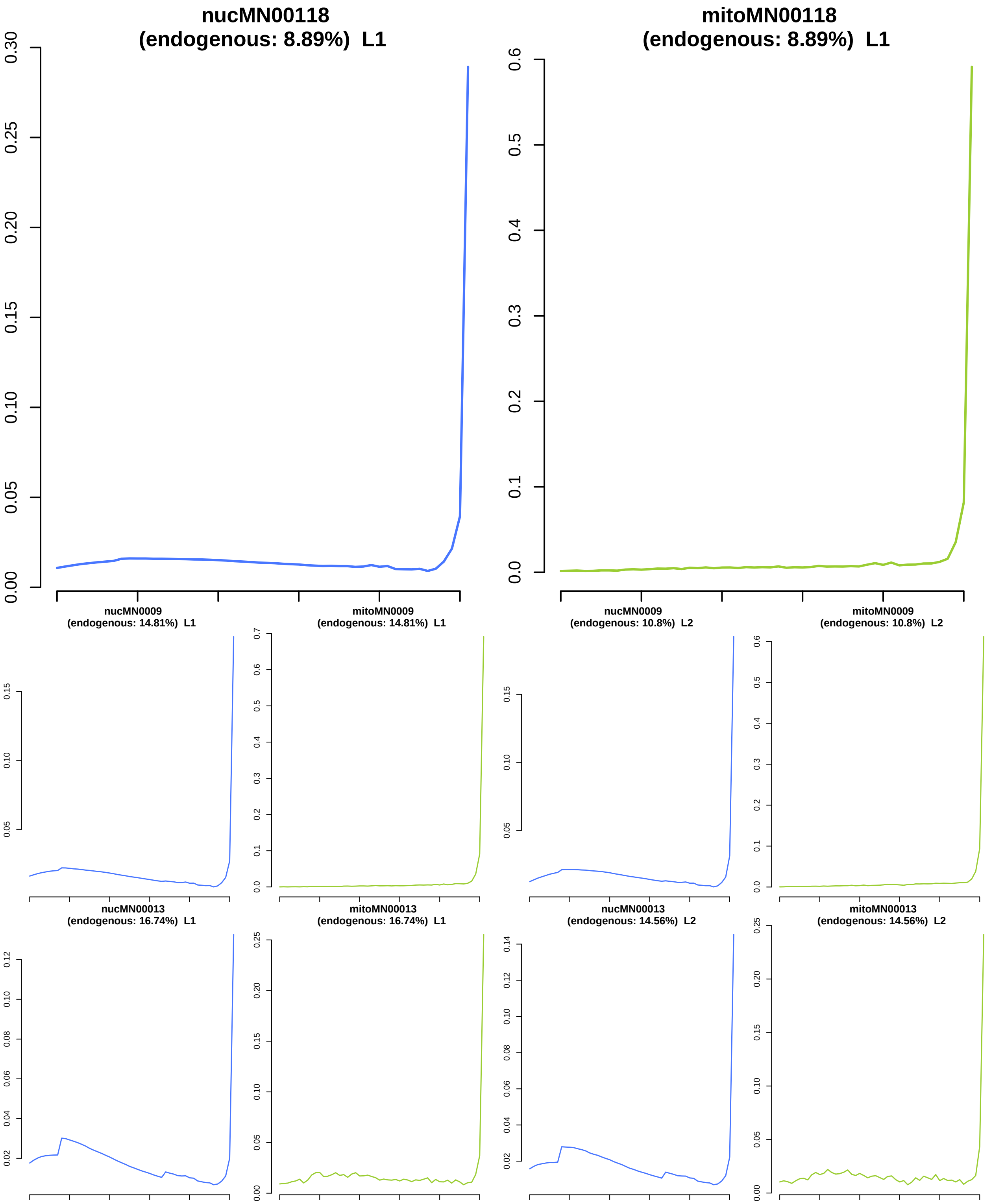
mitoMN00016
(endogenous: 3.55%) L1

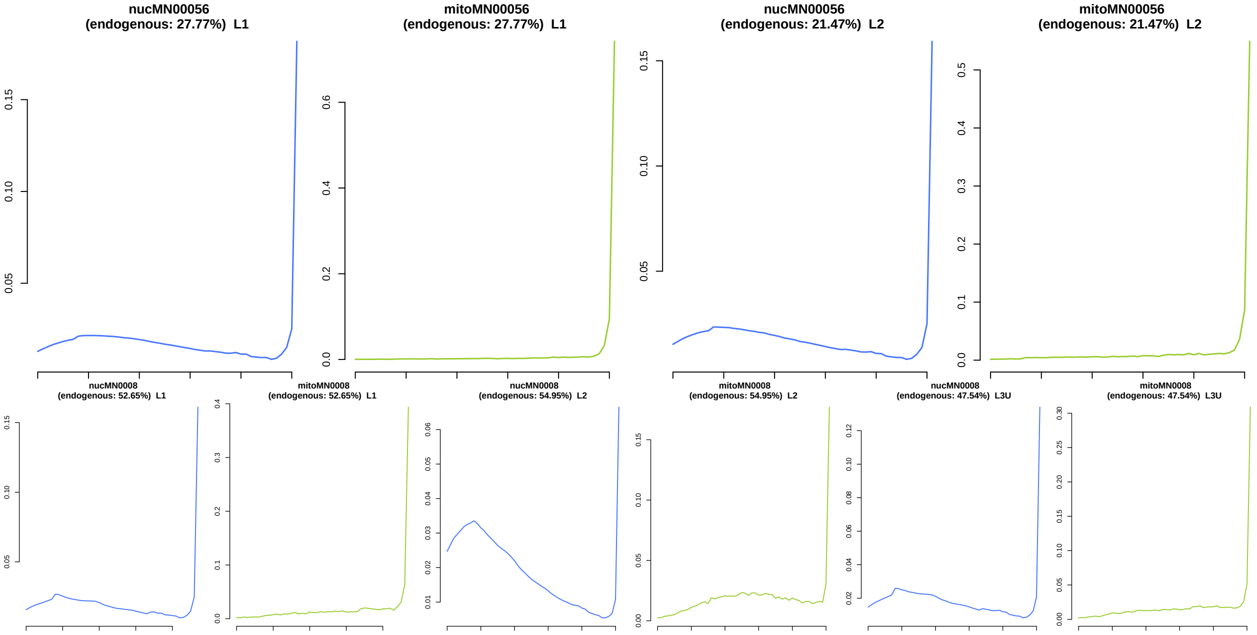
nucMN00064
(endogenous: 3.77%) L1

mitoMN00064
(endogenous: 3.77%) L1

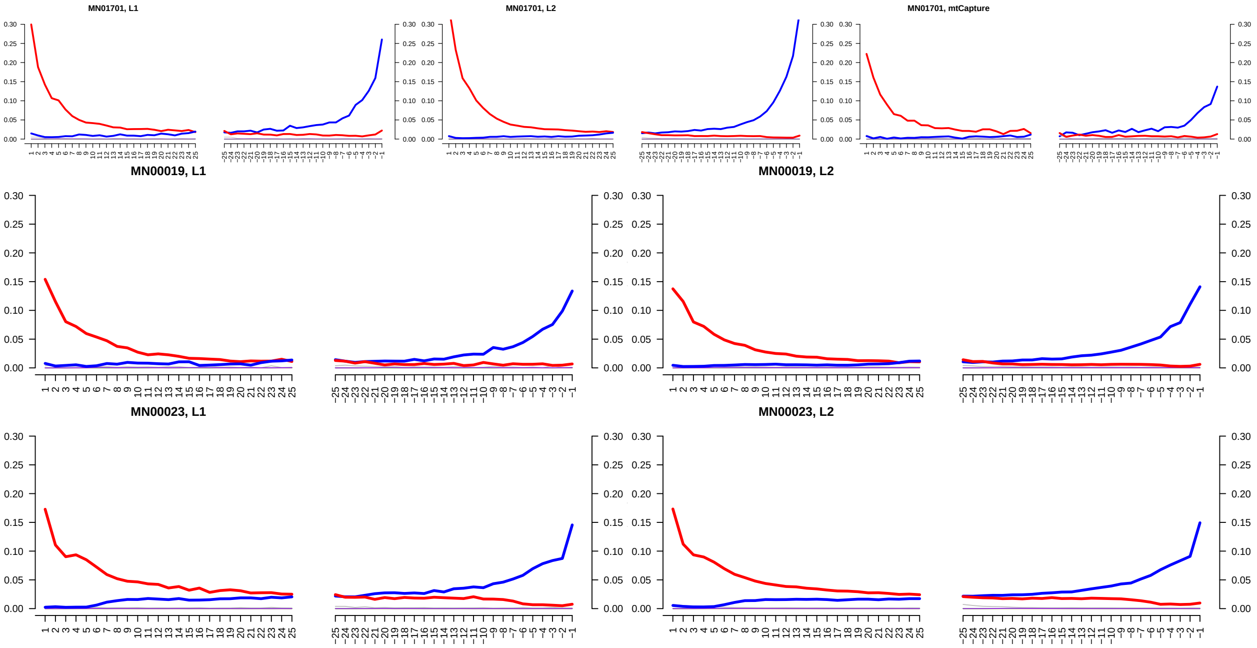

nucMN00346
(endogenous: 6.71%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
mitoMN00346
(endogenous: 6.71%) L1
0.0
0.1
0.2
0.3
0.4
0.5
nucMN00346
(endogenous: 3.85%) L2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mitoMN00346
(endogenous: 3.85%) L2
0.0
0.1
0.2
0.3
0.4
nucMN0003
(endogenous: 3.92%) L1
0.05
0.10
0.15
mitoMN0003
(endogenous: 3.92%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
nucMN00021
(endogenous: 8.3%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mitoMN00021
(endogenous: 8.3%) L1
0.0
0.1
0.2
0.3
0.4
nucMN00021
(endogenous: 4.57%) L2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mitoMN00021
(endogenous: 4.57%) L2
0.0
0.1
0.2
0.3
0.4
0.5
0.6


nucMN00067
(endogenous: 5.74%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
mitoMN00067
(endogenous: 5.74%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
nucMN00067
(endogenous: 4.97%) L2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
mitoMN00067
(endogenous: 4.97%) L2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
nucMN00119
(endogenous: 8.96%) L1
0.00
0.05
0.10
0.15
0.20
mitoMN00119
(endogenous: 8.96%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
nucMN00119
(endogenous: 6.49%) L2
0.00
0.05
0.10
0.15
0.20
0.25
mitoMN00119
(endogenous: 6.49%) L2
0.0
0.1
0.2
0.3
0.4
0.5
nucMN00066
(endogenous: 8.05%) L1
0.01
0.02
0.03
0.04
0.05
0.06
0.07
mitoMN00066
(endogenous: 8.05%) L1
0.00
0.05
0.10
0.15
0.20
0.25

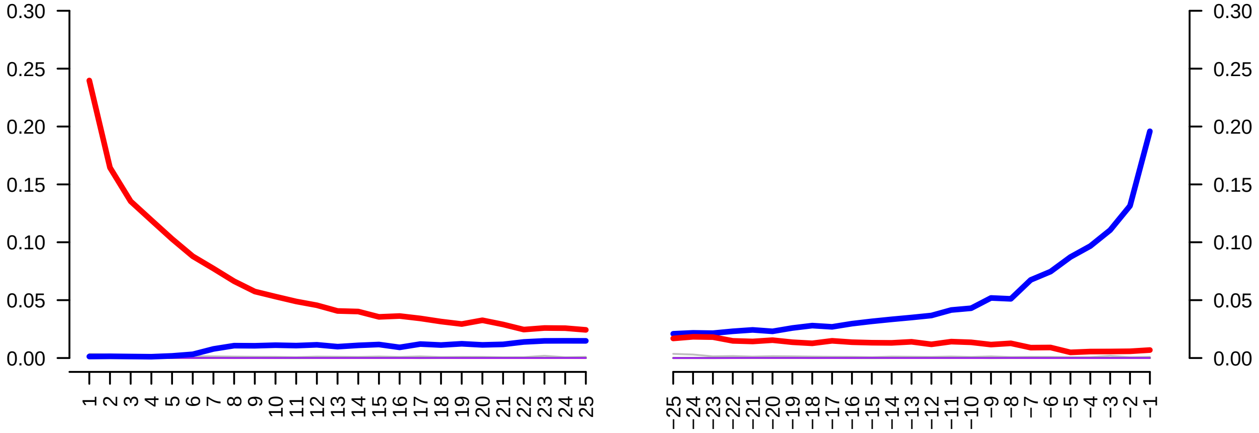

nucMN00118
(endogenous: 8.89%) L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
mitoMN00118
(endogenous: 8.89%) L1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
nucMN0009
(endogenous: 14.81%) L1
0.05
0.10
0.15
mitoMN0009
(endogenous: 14.81%) L1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
nucMN0009
(endogenous: 10.8%) L2
0.05
0.10
0.15
mitoMN0009
(endogenous: 10.8%) L2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
nucMN00013
(endogenous: 16.74%) L1
0.02
0.04
0.06
0.08
0.10
0.12
mitoMN00013
(endogenous: 16.74%) L1
0.00
0.05
0.10
0.15
0.20
0.25
nucMN00013
(endogenous: 14.56%) L2
0.02
0.04
0.06
0.08
0.10
0.12
0.14
mitoMN00013
(endogenous: 14.56%) L2
0.00
0.05
0.10
0.15
0.20
0.25



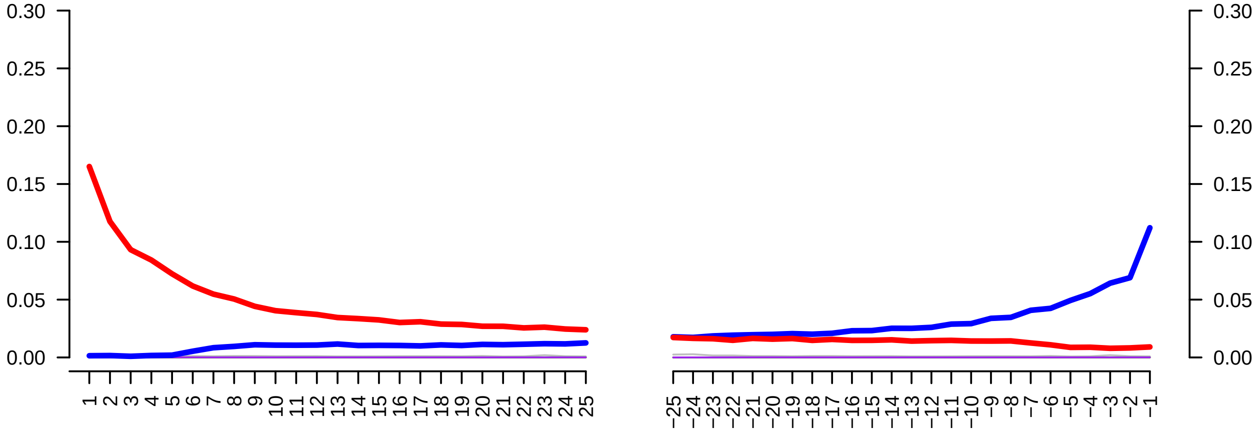
# Molecular damage



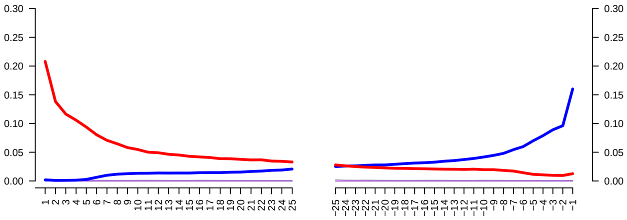
MN00010, L1



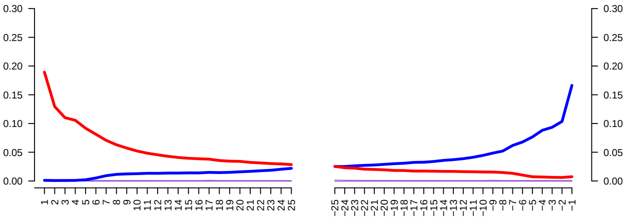
MN00039, L1



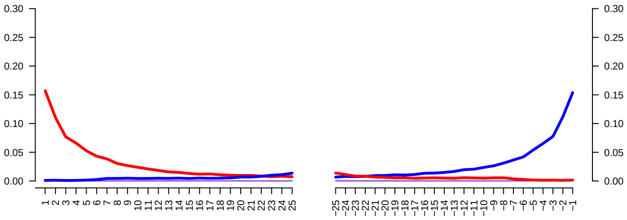
MN00316, L1



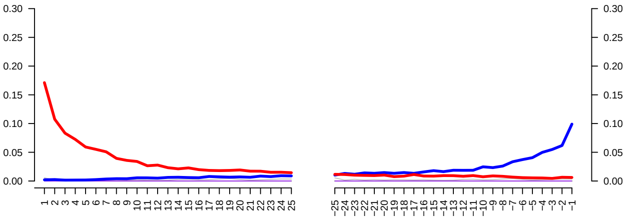
MN00316, L2



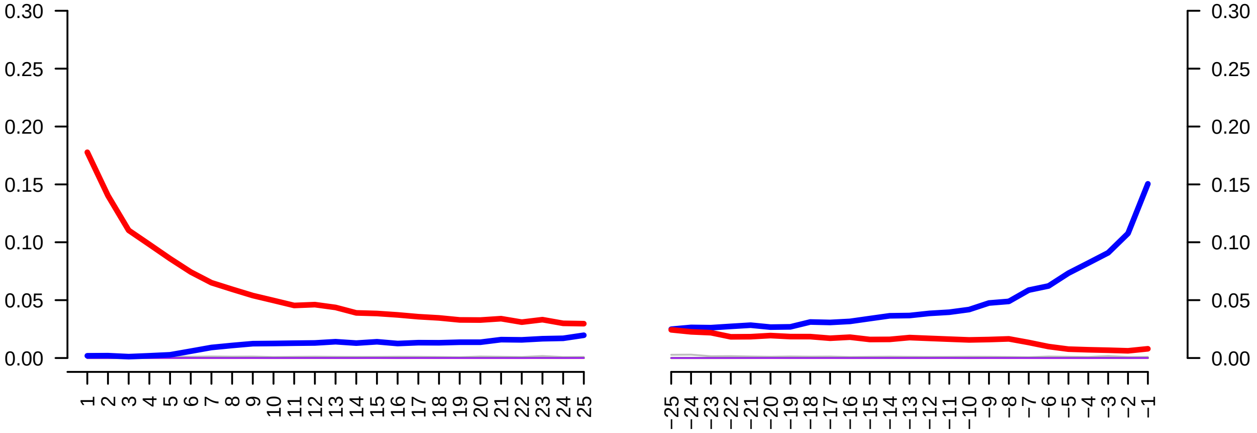
MN1943, L1



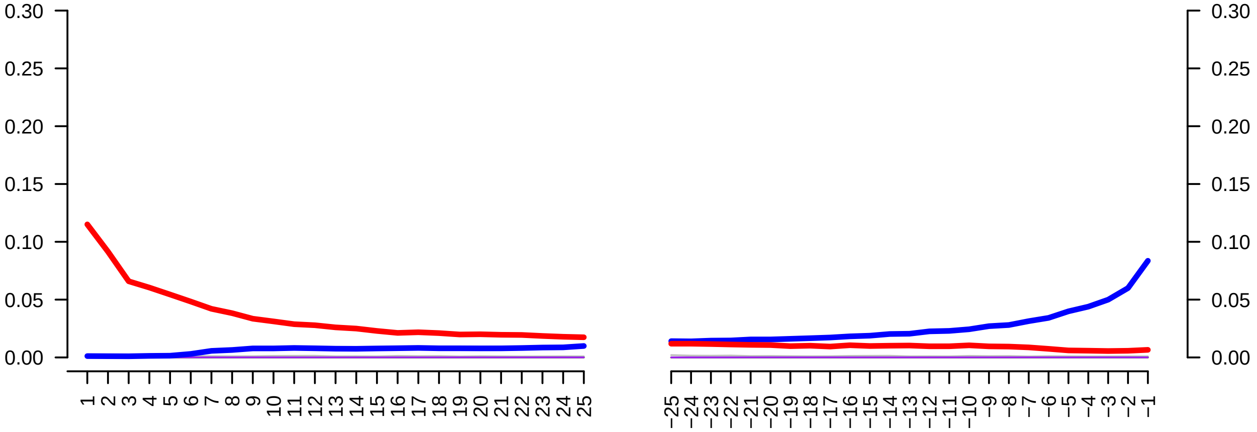
MN1943, mtCapture



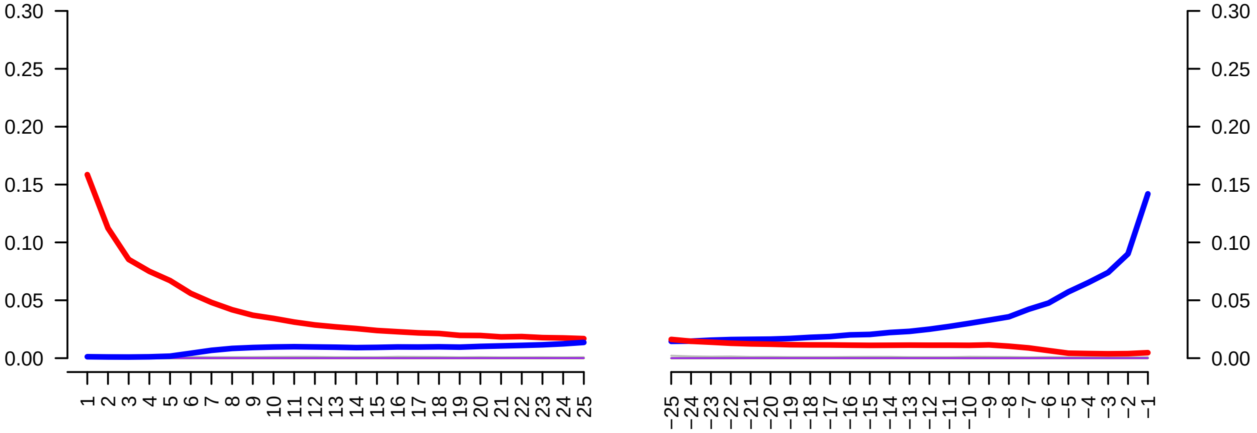
**MN00069, L1**

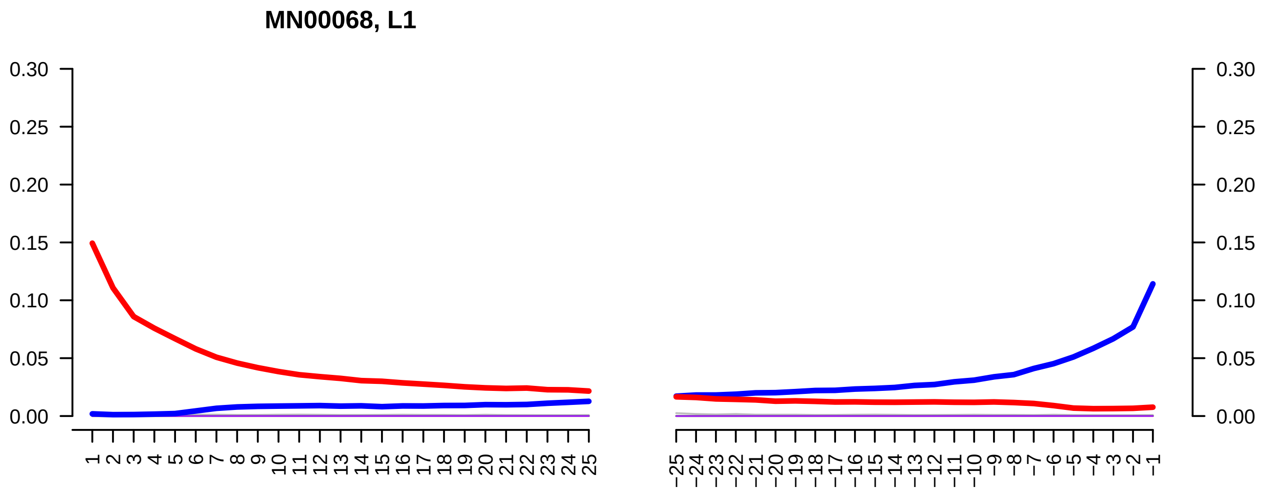


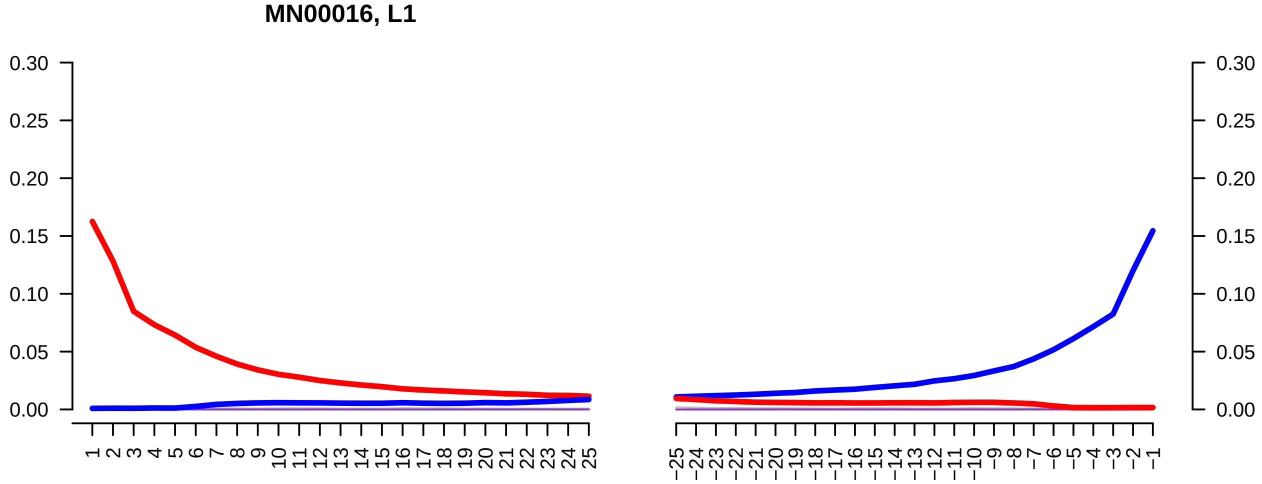
**MN00022, L1**

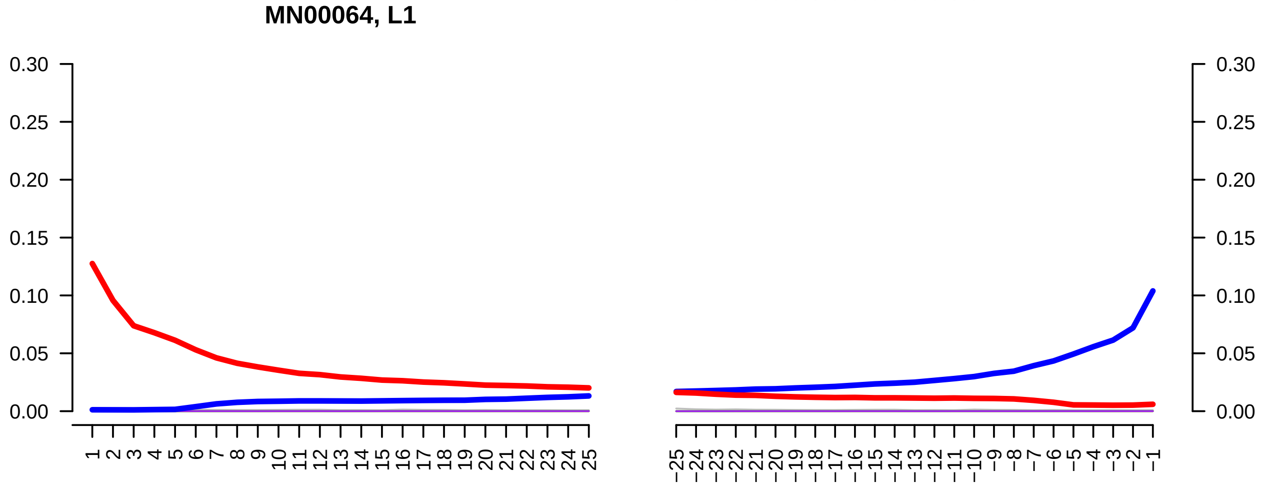


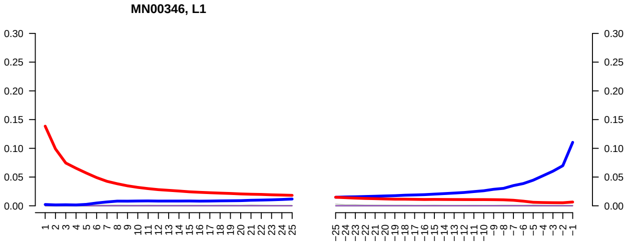
**MN00045, L1**

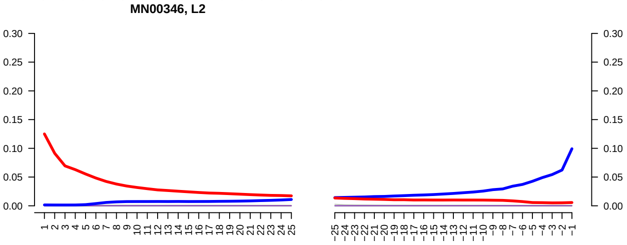



MN00068, L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
−25 −24 −23 −22 −21 −20 −19 −18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1

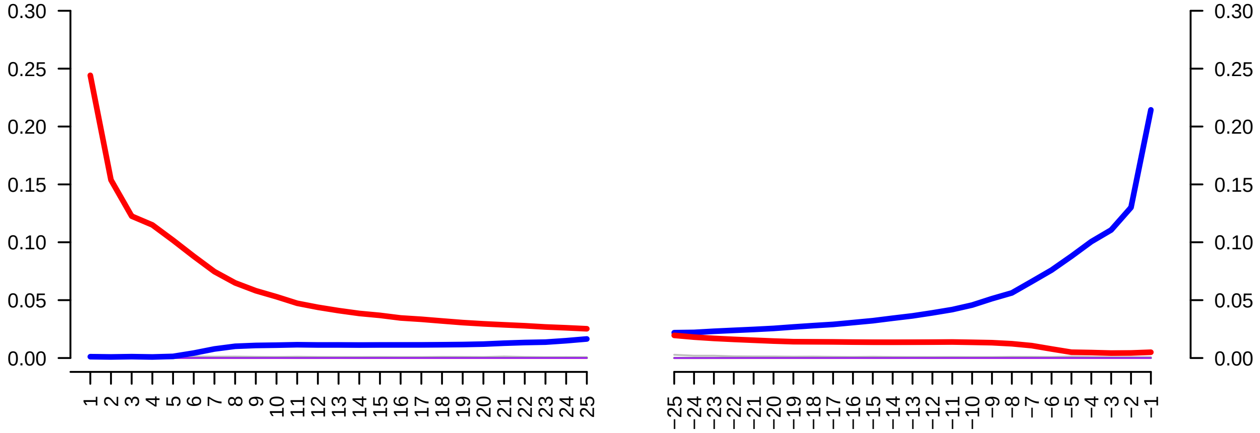

MN00016, L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
−25 −24 −23 −22 −21 −20 −19 −18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1


MN00064, L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
−25 −24 −23 −22 −21 −20 −19 −18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1

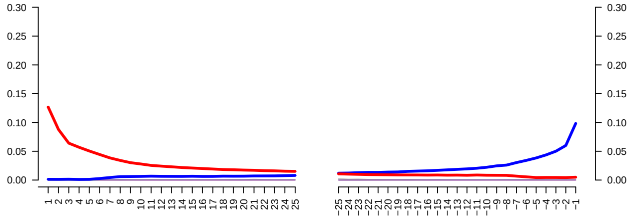

MN00346, L1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
−25 −24 −23 −22 −21 −20 −19 −18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1


MN00346, L2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
−25 −24 −23 −22 −21 −20 −19 −18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1

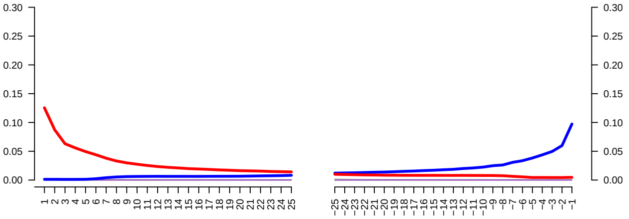
MN0003, L1



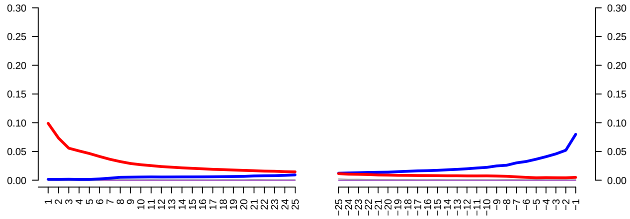
MN00021, L1



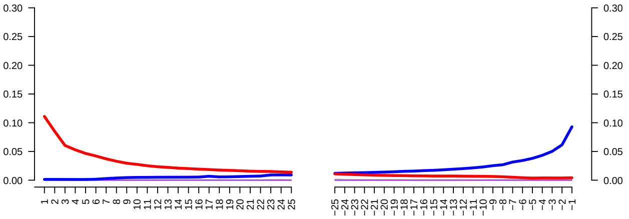
MN00021, L2



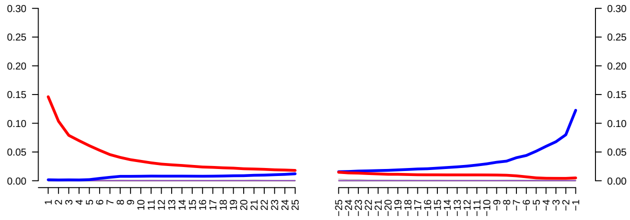
MN00067, L1



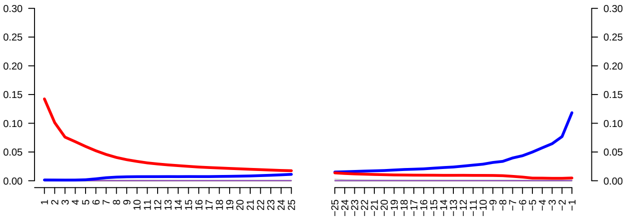
MN00067, L2



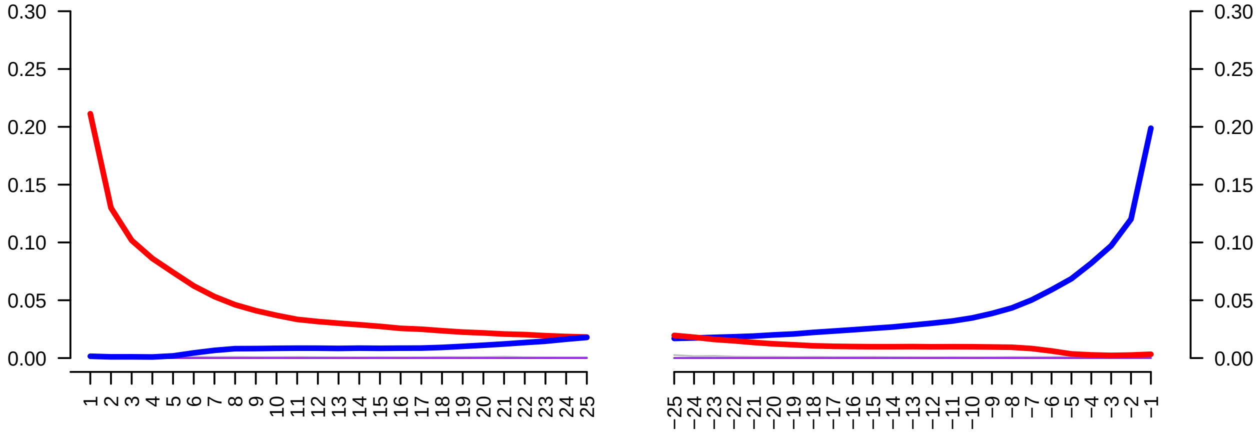
MN00119, L1



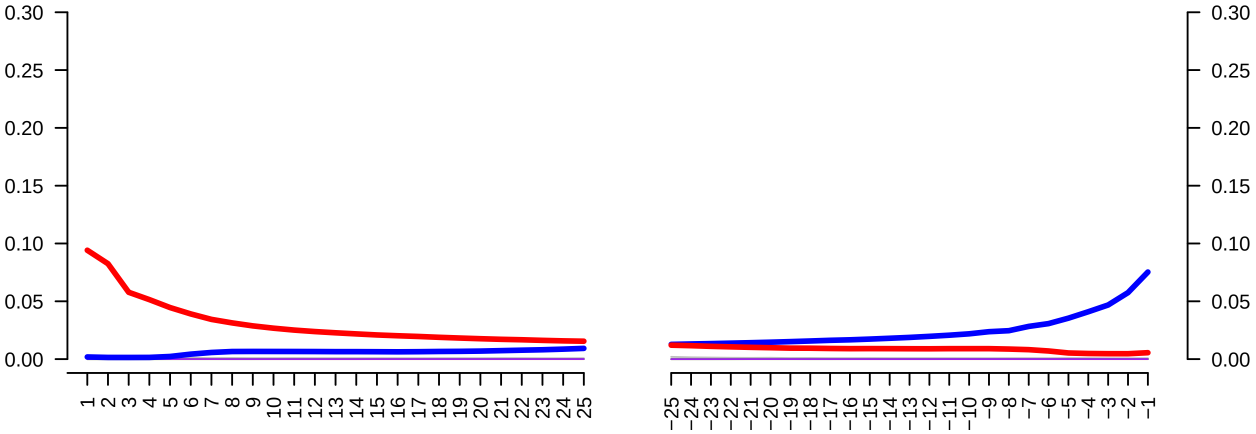
MN00119, L2



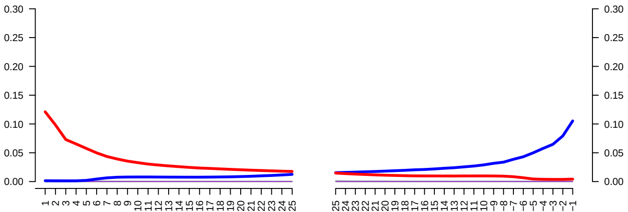
MN00066, L1



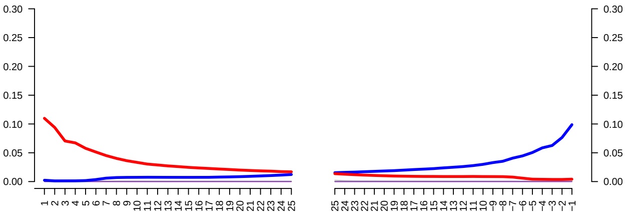
MN00118, L1



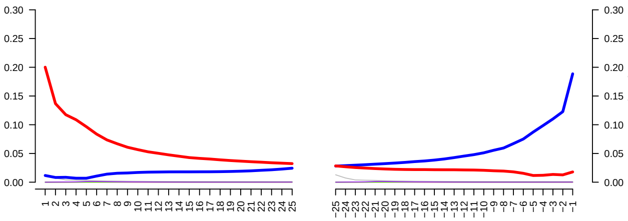
MN0009, L1



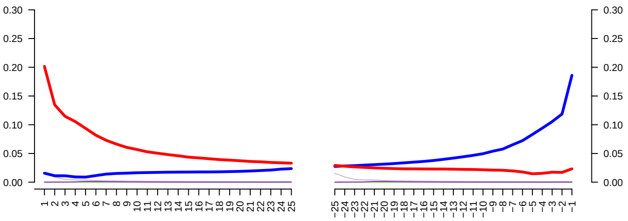
MN0009, L2



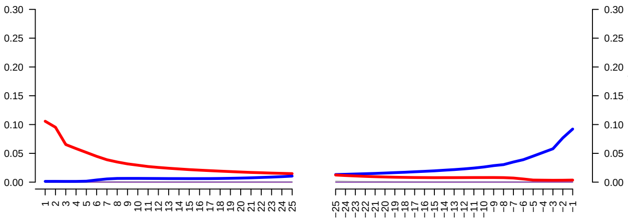
MN00013, L1



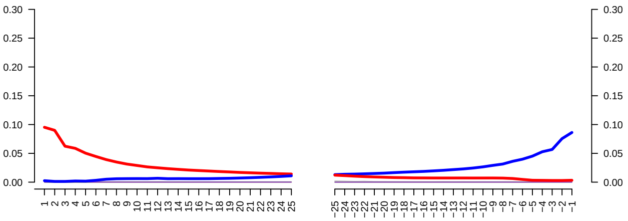
MN00013, L2



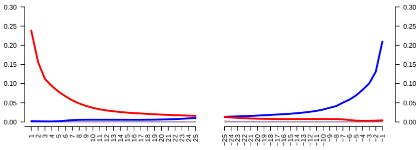
MN00056, L1



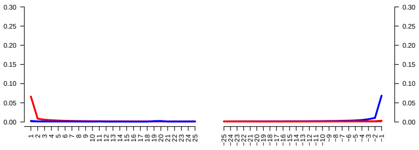
MN00056, L2



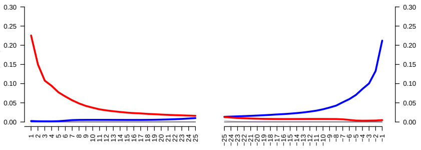
MN0008, L1



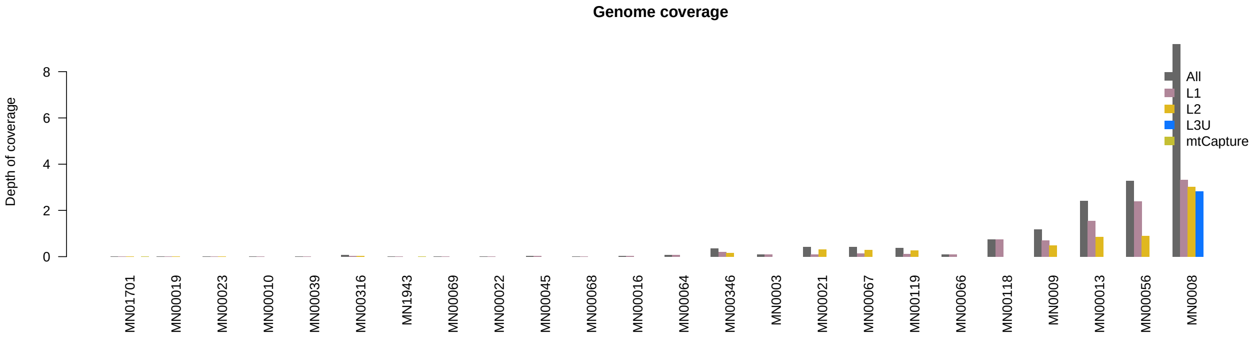
MN0008, L2

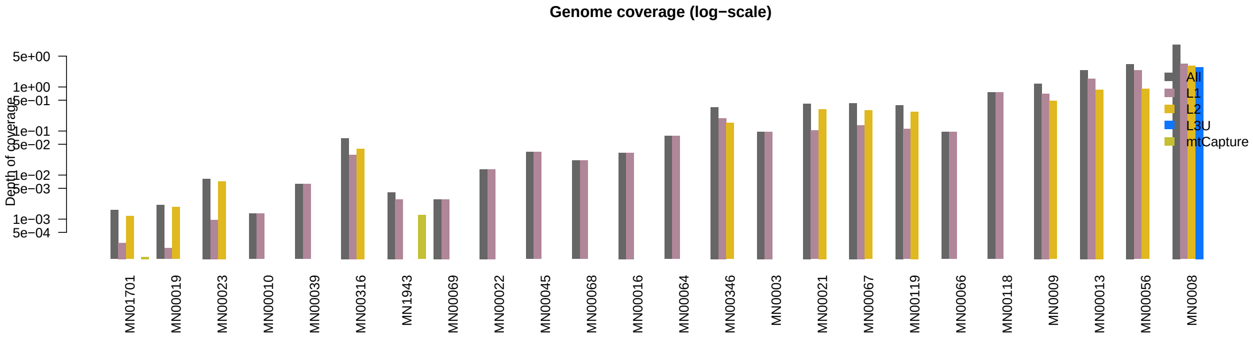


MN0008, L3U

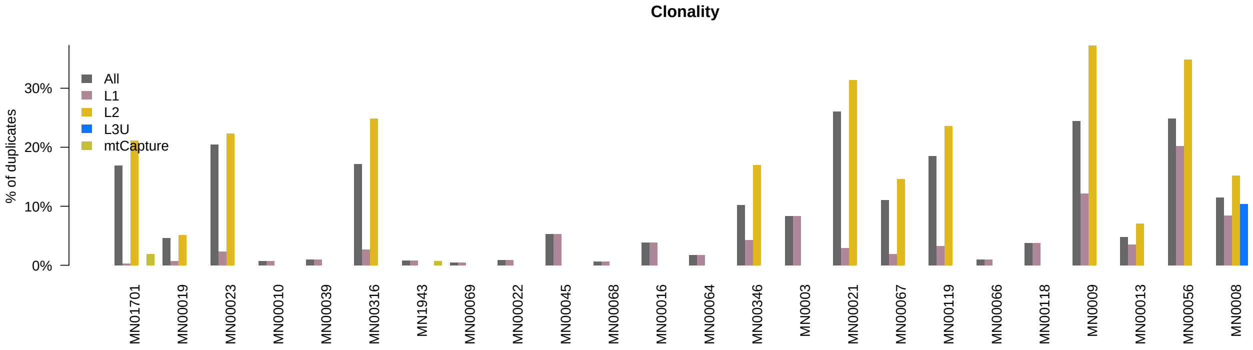


# Genome coverage



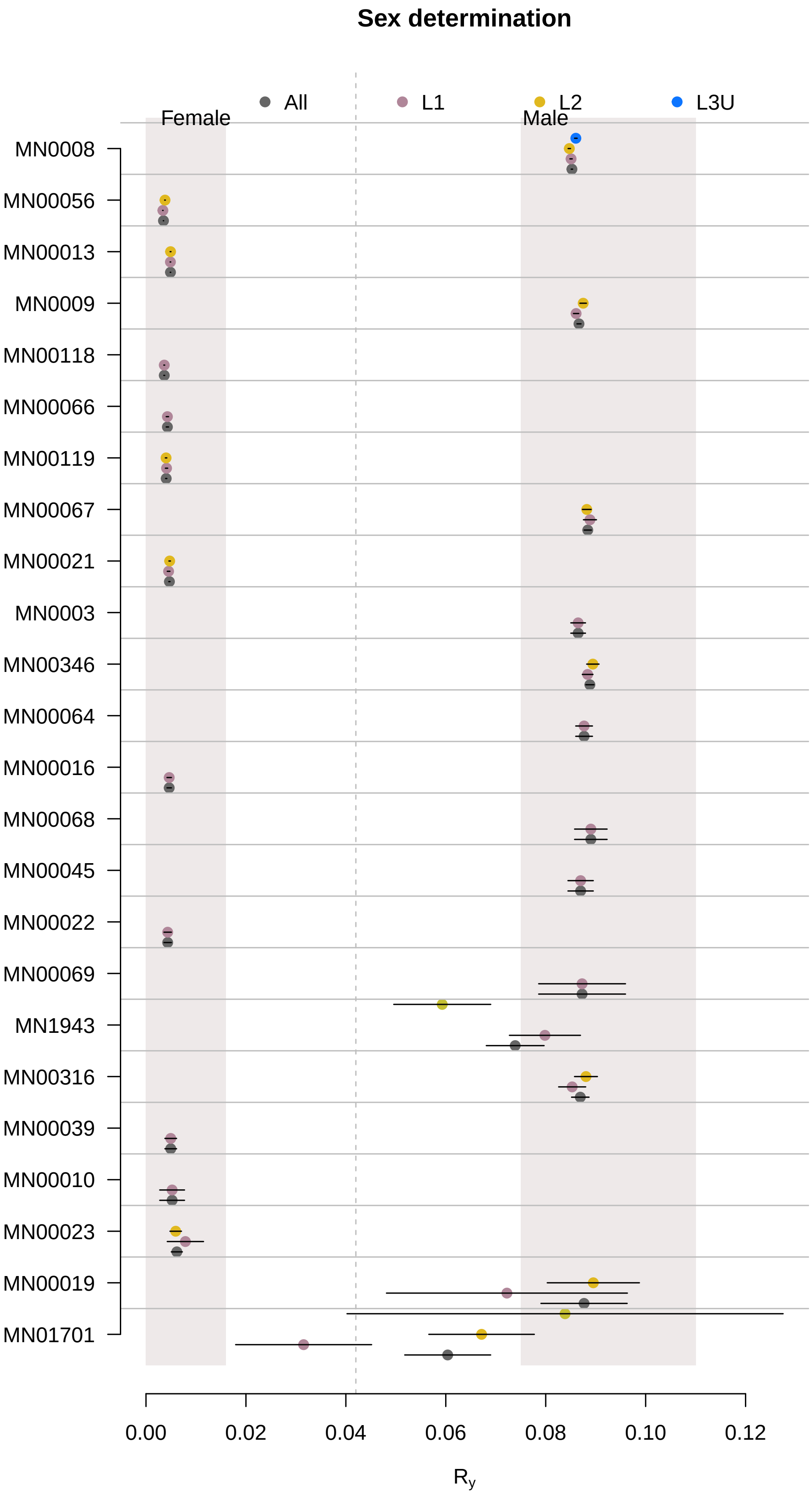


# Clonality



# Sex determination

```
## [1] 120
```


Sex determination
All
L1
L2
L3U
Female
Male
MN0008
MN00056
MN00013
MN0009
MN00118
MN00066
MN00119
MN00067
MN00021
MN0003
MN00346
MN00064
MN00016
MN00068
MN00045
MN00022
MN00069
MN1943
MN00316
MN00039
MN00010
MN00023
MN00019
MN01701
0.00
0.02
0.04
0.06
0.08
0.10
0.12
$R_y$

# Contamination estimates

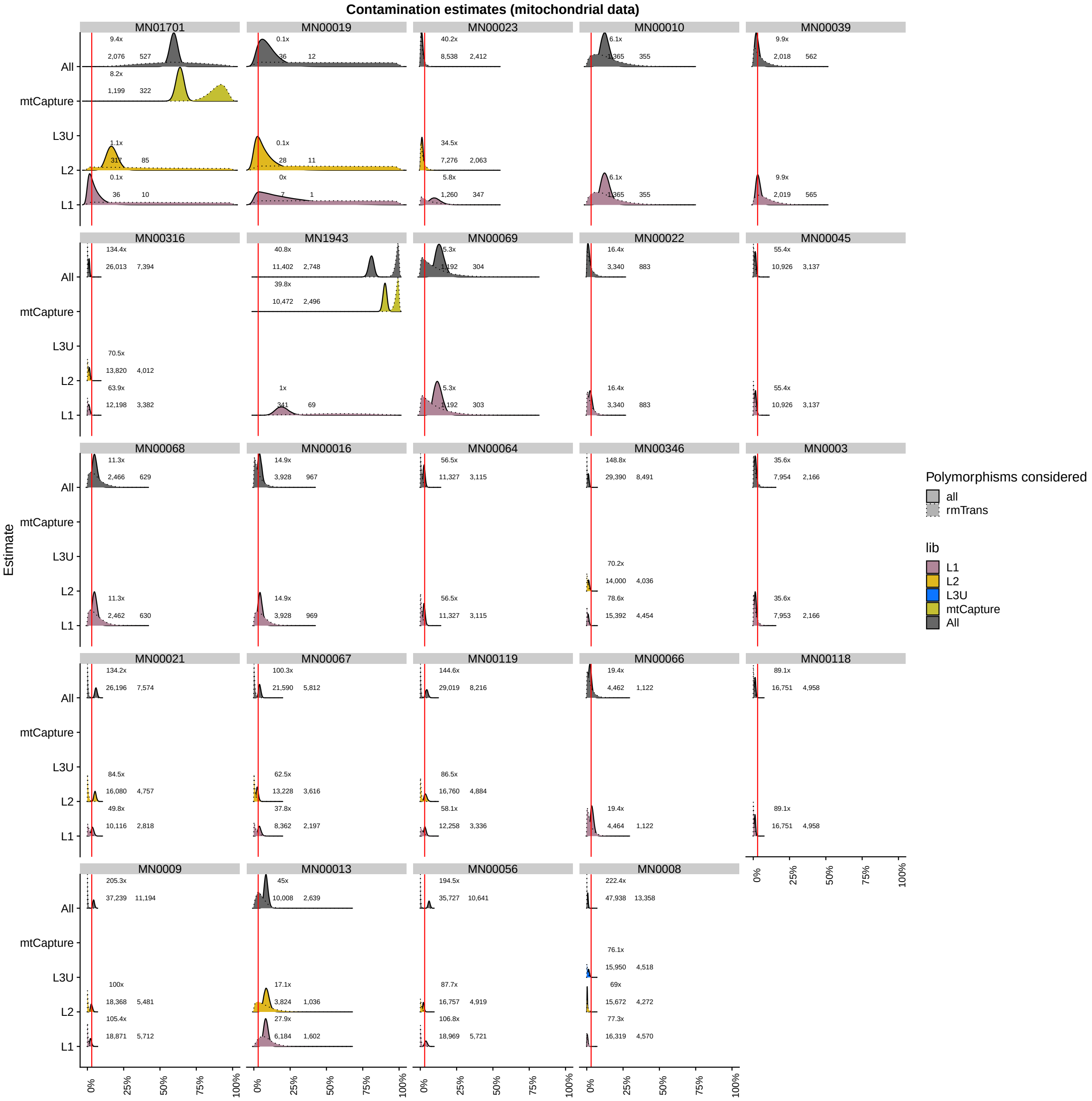
Work in progress

## Mitochondrial

Text on plots:

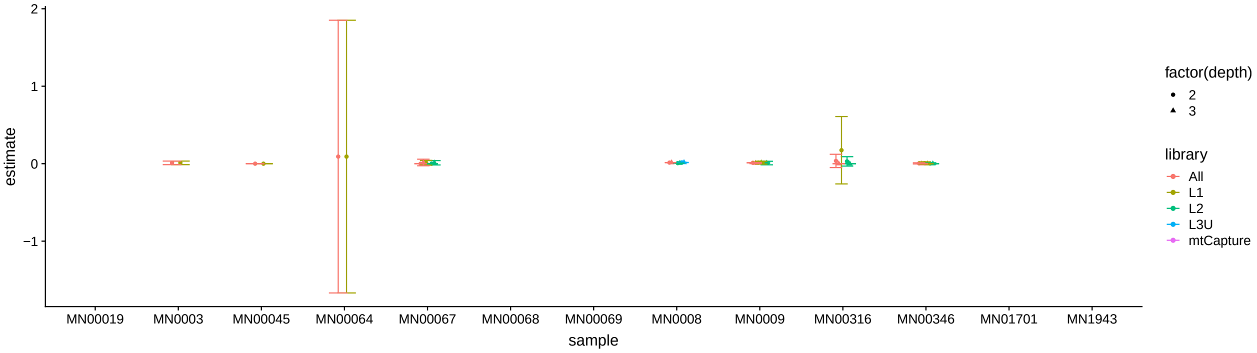
Mitochondrial coverage,

number of reads (all polymorphisms), number of reads (removing transitions)



# X

Work in progress



| sample | library | mito_estimate | mito_est_low | mito_est_upper | X_estimate | X_est_low | X_est_upper | mito_numReads | mito_coverage | X_num_X_sites | min_depth | sex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MN00019 | All | 6% | 2% | 26% | NA% | NA% | NA% | 36 | 0.1240 | NA | 3 | XY |
| MN00019 | L1 | 4% | 0% | 66% | NA% | NA% | NA% | 7 | 0.0213 | NA | 3 | consistent with XY but not XX |
| MN00019 | L2 | 2% | 0% | 26% | NA% | NA% | NA% | 28 | 0.1030 | NA | 3 | XY |
| MN0003 | All | 2% | 0% | 4% | NA% | NA% | NA% | 7954 | 35.6000 | NA | 3 | XY |
| MN0003 | L1 | 2% | 0% | 4% | NA% | NA% | NA% | 7953 | 35.6000 | NA | 3 | XY |
| MN00045 | All | 2% | 0% | 2% | NA% | NA% | NA% | 10926 | 55.4000 | NA | 3 | XY |
| MN00045 | L1 | 2% | 0% | 2% | NA% | NA% | NA% | 10926 | 55.4000 | NA | 3 | XY |
| MN00064 | All | 2% | 2% | 4% | NA% | NA% | NA% | 11327 | 56.5000 | NA | 3 | XY |
| MN00064 | L1 | 2% | 2% | 4% | NA% | NA% | NA% | 11327 | 56.5000 | NA | 3 | XY |
| MN00067 | All | 4% | 2% | 6% | 2% | -2% | 6% | 21590 | 100.0000 | 82 | 3 | XY |
| MN00067 | L1 | 4% | 2% | 6% | NA% | NA% | NA% | 8362 | 37.8000 | NA | 3 | XY |
| MN00067 | L2 | 2% | 2% | 4% | 2% | -2% | 4% | 13228 | 62.5000 | 29 | 3 | XY |
| MN00068 | All | 4% | 2% | 10% | NA% | NA% | NA% | 2466 | 11.3000 | NA | 3 | XY |
| MN00068 | L1 | 4% | 2% | 8% | NA% | NA% | NA% | 2462 | 11.3000 | NA | 3 | XY |
| MN00069 | All | 12% | 8% | 20% | NA% | NA% | NA% | 1192 | 5.3400 | NA | 3 | XY |
| MN00069 | L1 | 12% | 8% | 18% | NA% | NA% | NA% | 1192 | 5.3400 | NA | 3 | XY |
| MN0008 | All | 0% | 0% | 2% | 2% | 2% | 2% | 47938 | 222.0000 | 47991 | 3 | XY |
| MN0008 | L1 | 0% | 0% | 2% | NA% | NA% | NA% | 16319 | 77.3000 | NA | 3 | XY |
| MN0008 | L2 | 0% | 0% | 0% | 0% | 0% | 0% | 15672 | 69.0000 | 12347 | 3 | XY |
| MN0008 | L3U | 2% | 0% | 2% | 2% | 2% | 2% | 15950 | 76.1000 | 10001 | 3 | XY |
| MN0009 | All | 4% | 4% | 6% | 0% | 0% | 2% | 37239 | 205.0000 | 1417 | 3 | XY |
| MN0009 | L1 | 2% | 2% | 4% | 2% | 0% | 4% | 18871 | 105.0000 | 314 | 3 | XY |
| MN0009 | L2 | 2% | 2% | 4% | 0% | -2% | 4% | 18368 | 100.0000 | 162 | 3 | XY |
| MN00316 | All | 2% | 0% | 2% | 0% | 0% | 0% | 26013 | 134.0000 | 1 | 3 | XY |
| MN00316 | L1 | 0% | 0% | 2% | NA% | NA% | NA% | 12198 | 63.9000 | NA | 3 | XY |
| MN00316 | L2 | 2% | 0% | 2% | 0% | 0% | 0% | 13820 | 70.5000 | 1 | 3 | XY |
| MN00346 | All | 2% | 0% | 2% | 0% | 0% | 0% | 29390 | 149.0000 | 97 | 3 | XY |
| MN00346 | L1 | 0% | 0% | 2% | 0% | 0% | 0% | 15392 | 78.6000 | 10 | 3 | XY |
| MN00346 | L2 | 2% | 0% | 2% | 0% | 0% | 0% | 14000 | 70.2000 | 20 | 3 | XY |
| MN01701 | L2 | 20% | 14% | 28% | NA% | NA% | NA% | 317 | 1.0700 | NA | 3 | consistent with XY but not XX |
| MN01701 | mtCapture | NA% | NA% | NA% | NA% | NA% | NA% | 1199 | 8.2400 | NA | 3 | consistent with XY but not XX |

| sample | library | mito_estimate | mito_est_lower | mito_est_upper | X_estimate | X_est_lower | X_est_upper | mito_num_Reads | mito_coverage | X_num_sites | min_depth | sex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MN1943 | All | 20% | 14% | 30% | NA% | NA% | NA% | 11402 | 40.8000 | NA | 3 | consistent with XY but not XX |
| MN1943 | L1 | 22% | 14% | 32% | NA% | NA% | NA% | 341 | 1.0100 | NA | 3 | consistent with XY but not XX |

**Supplementary table 1**