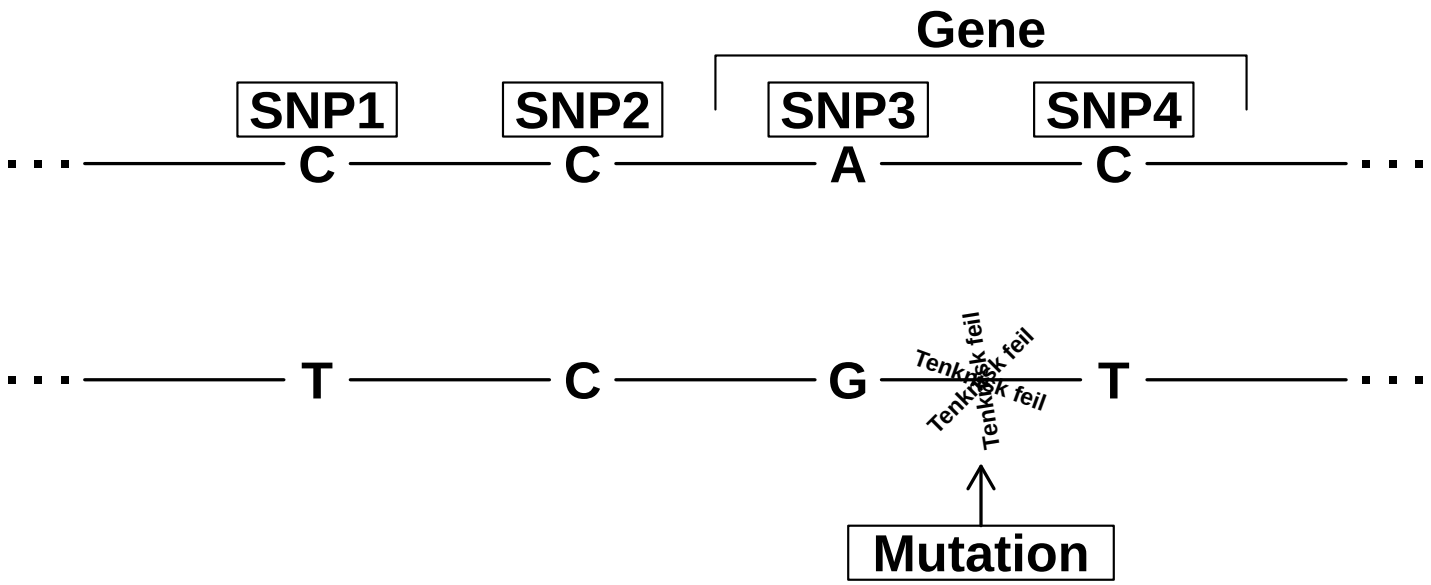
Run `haplin`

with haplotypes

SNPs, HAPLOTYPES, MUTATIONS

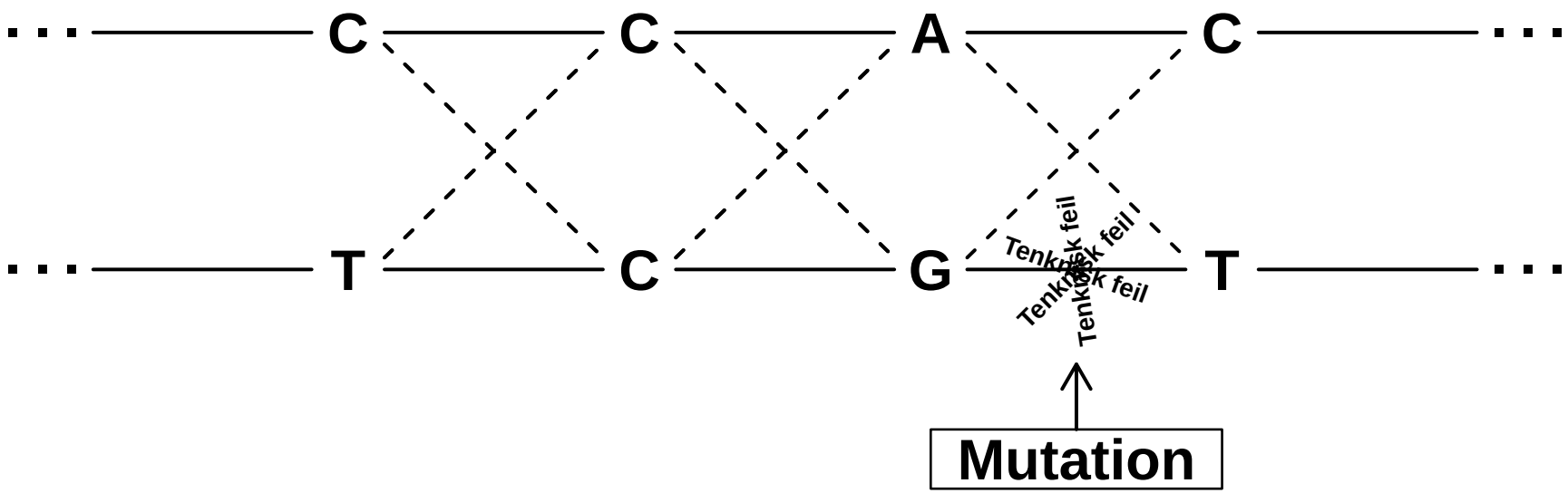
- Two haplotypes in one individual
- One from the mother, one from the father



Haplotypes = C-C-A-C and T-C-G-T

SNP = “Single Nucleotide Polymorphism”

PROBLEM: HAPLOTYPES UNKNOWN!



- Can construct $2^4 = 16$ different haplotypes (here, only 8)
- In general: 2^L haplotypes with L SNPs
- In real life only a limited number of haplotypes at a locus (but don't know which ones)

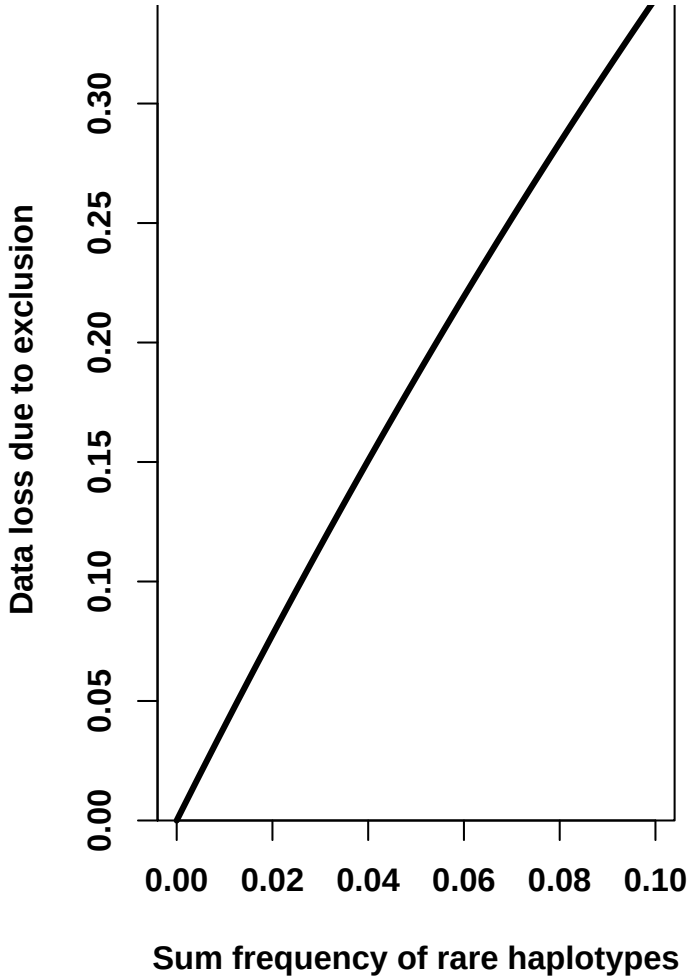
THE NUMBER OF POSSIBLE TRIADS

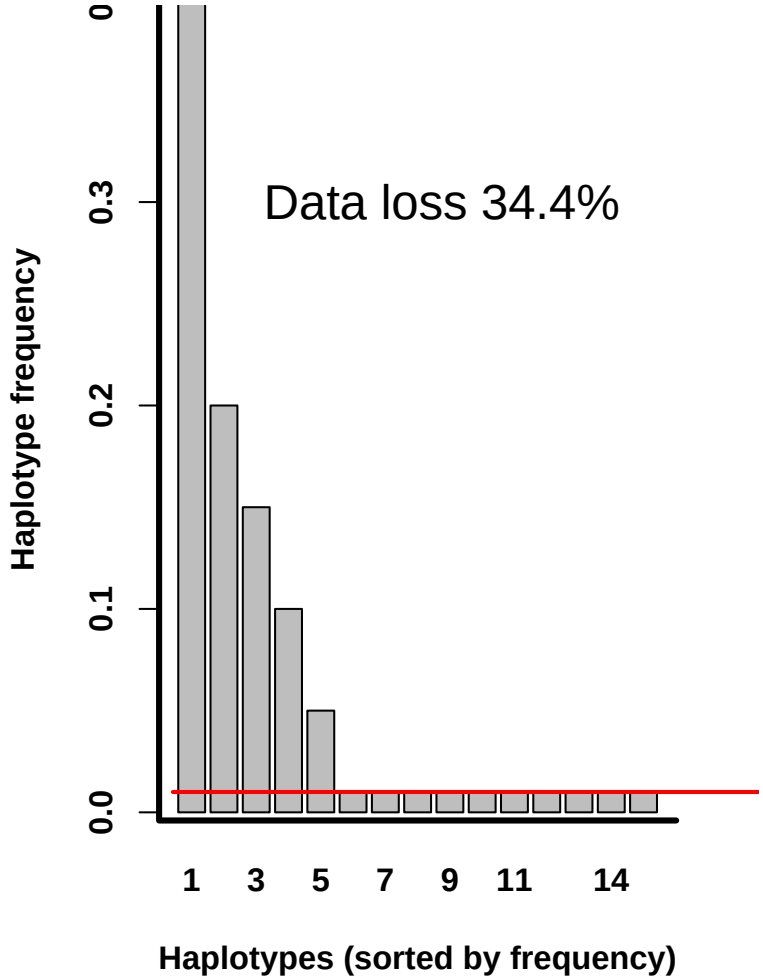
SNPs	Haplotypes	Comb. 2 haplotypes	Triads
L	$K = 2^L$	$K(K+1)/2$	$K^4 - K(K-1)/2$
1	2	3	15
2	4	10	250
3	8	36	4 068
4	16	136	65 416
5	32	528	1 048 080
6	64	2080	16 775 200

Problems:

- Missing data can make the number of possibilites much larger
- Too many parameters to estimate
  (risk estimates for all haplotype combinations)
- Computationally extremely demanding

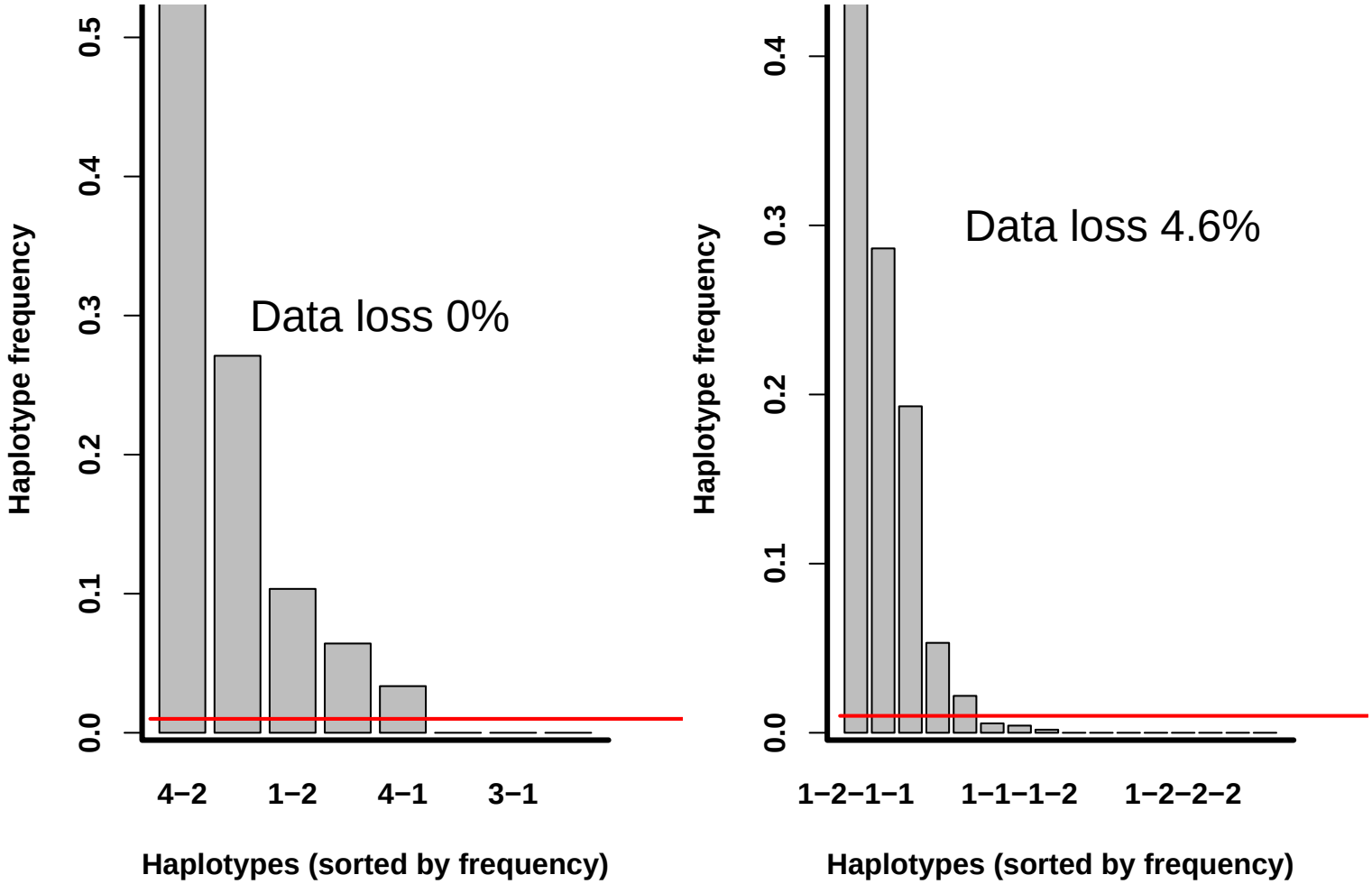
THERE IS A POTENTIAL DATA LOSS DUE TO RARE HAPLOTYPES





Haplin applies a threshold to haplotype frequencies (default = 0.01)

ACTUAL DATA LOSS DUE TO RARE HAPLOTYPES IS MODERATE



Haplin applies a threshold to haplotype frequencies (default = 0.01)

HAPLIN RUN, WITH HAPLOTYPES

Load data:

```
pres.data <- genDataLoad(filename = "data_preprocessed",
 dir.in = "data")
```

Haplotype run:

- Build haplotypes from three first SNPs (`markers = 1:3`)
- Impute missing (`use.missing = T`)
- We use `reference = "ref.cat"` to ease comparison with later results

```
haplin(data = pres.data, markers = 1:3,
use.missing = T, reference = "ref.cat")
```

HAPLIN OUTPUT: CONVERGENCE

```
Using EM to estimate full model:
EM iter: 1    |GLM deviance: 6.66134e-15  |Coefficients: 6.24336e-1
EM iter: 2    |GLM deviance: 469.778      |Coefficients: -1.92129
EM iter: 3    |GLM deviance: 451.692      |Coefficients: -2.39197
EM iter: 4    |GLM deviance: 453.761      |Coefficients: -2.44295
EM iter: 5    |GLM deviance: 453.983      |Coefficients: -2.44777
EM iter: 6    |GLM deviance: 453.999      |Coefficients: -2.4482
EM iter: 7    |GLM deviance: 453.998      |Coefficients: -2.4482
EM iter: 8    |GLM deviance: 453.997      |Coefficients: -2.44817
EM iter: 9    |GLM deviance: 453.996      |Coefficients: -2.44815
EM iter: 10   |GLM deviance: 453.996      |Coefficients: -2.44815
EM iter: 11   |GLM deviance: 453.996      |Coefficients: -2.44814
EM iter: 12   |GLM deviance: 453.996      |Coefficients: -2.44814
EM iter: 13   |GLM deviance: 453.996      |Coefficients: -2.44814
EM iter: 14   |GLM deviance: 453.996      |Coefficients: -2.44814
EM iter: 15   |GLM deviance: 453.996      |Coefficients: -2.44814
EM iter: 16   |GLM deviance: 453.996      |Coefficients: -2.44814
```

Convergence looks OK (it usually does), but suggests low LD(?)

HAPLIN OUTPUT:

```
Accounting for possible loss of triads:

 Cause of loss      Triads removed Triads remaining
 Missing data                    0               559
 Mendelian incons.               0               559
 Unused haplotypes               0               559
```

No loss of data.

HAPLIN OUTPUT: MARKER SUMMARY INFO

```
Marker rs1:
Missing alleles:  168
 Allele Frequency Percent
 c       146        4.6
 G      3040       95.4
 total  3186      100.0
Chi-squared test for HWE, p-value:  0.7075

Marker rs3:
Missing alleles:  188
 Allele Frequency Percent
 A      2460       77.7
 t       706       22.3
 total  3166      100.0
Chi-squared test for HWE, p-value:  0.3621

Marker rs5:
Missing alleles:  180
 Allele Frequency Percent
 a      1196       37.7
 T      1978       62.3
 total  3174      100.0
Chi-squared test for HWE, p-value:  0.3014
```

All HWE tests OK. Note marker names.

HAPLIN OUTPUT: HAPLOTYPE FREQUENCIES

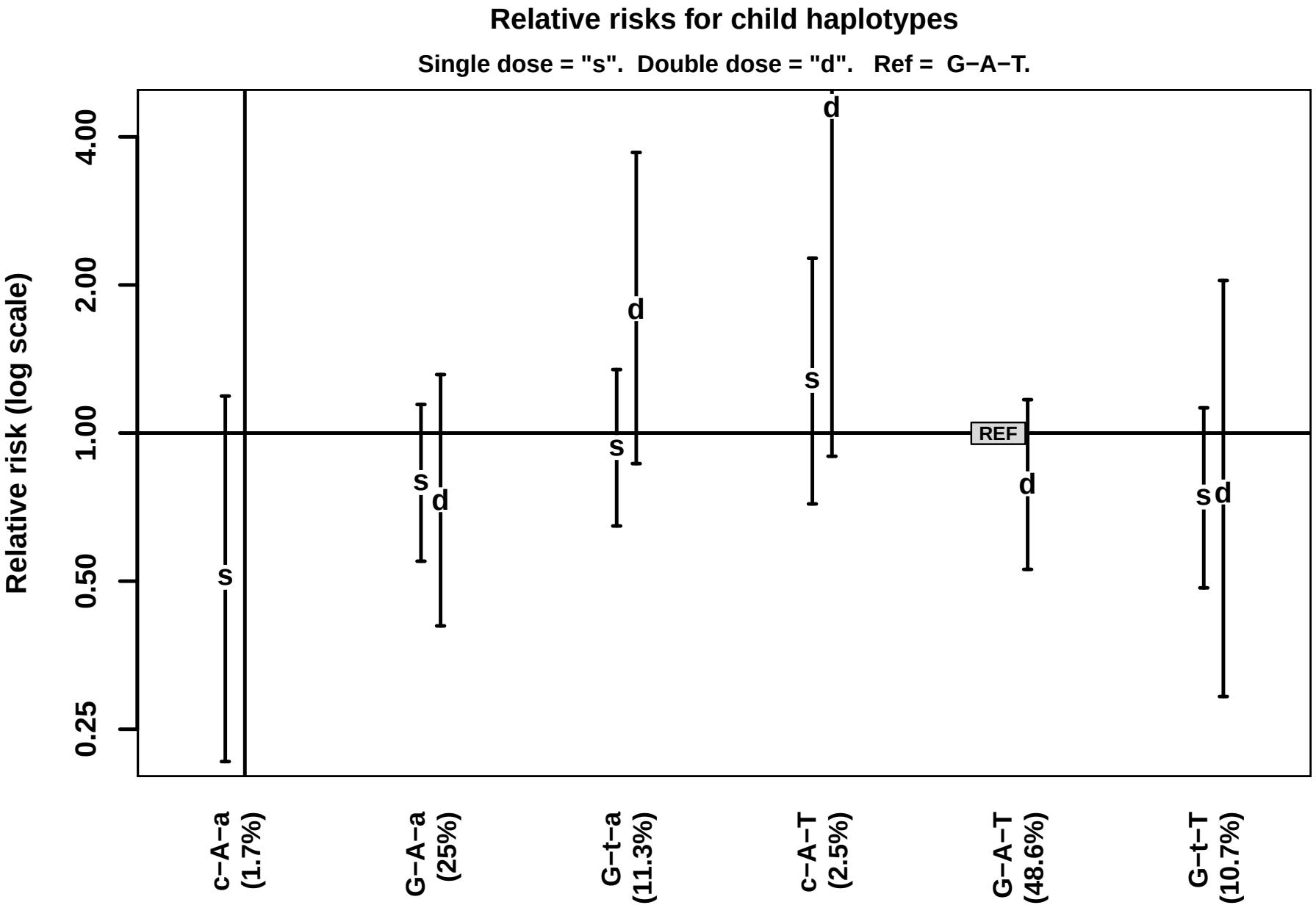
```
Haplotypes removed because of low frequencies:
c-t-a c-t-T

Haplotypes used in the analysis, with coding:
c-A-a G-A-a G-t-a c-A-T G-A-T G-t-T
    1     2     3     4     5     6

Number of haplotypes: 6

Haplotype frequencies with 95% confidence intervals:
 Haplotype Frequency(%) lower upper
 c-A-a      1.68         1.04  2.70
 G-A-a     24.99        22.34 27.84
 G-t-a     11.31         9.43 13.56
 c-A-T      2.50         1.69  3.66
 G-A-T     48.65        45.49 51.78
 G-t-T     10.67         8.81 12.82
```

HAPLIN OUTPUT: EFFECT ESTIMATES

```
Single- and double dose effects (Relative Risk) with 95% confidence intervals:
Reference method: ref.cat
Reference category: 5
Response model: free

----Child haplotypes----
 Haplotype Dose     Relative Risk Lower CI   Upper CI   P-value
 c-A-a     Single   0.509         0.215      1.19       0.124
 c-A-a     Double   2.68e-07      0          Inf        0.989

 G-A-a     Single   0.793         0.549      1.14       0.213
 G-A-a     Double   0.732         0.406      1.31       0.306

 G-t-a     Single   0.935         0.647      1.35       0.711
 G-t-a     Double   1.79          0.867      3.72       0.122

 c-A-T     Single   1.28          0.718      2.27       0.397
 c-A-T     Double   4.61          0.897      24.3       0.0688

 G-A-T     Single   REF
 G-A-T     Double   0.789         0.528      1.17       0.239

 G-t-T     Single   0.741         0.485      1.13       0.159
 G-t-T     Double   0.759         0.291      2.04       0.574
```

May suggest double dose (recessive) effects of haplotype `c-A-T`.
But not extremely convincing. `c-A-T` is rare, and CI is thus wide.

HAPLIN OUTPUT: EFFECT ESTIMATES



HAPLIN OUTPUT: LIKELIHOOD RATIO TEST

```
Overall test for difference between null model (no effects)
and full model:
------------
LIKELIHOOD RATIO TEST:
 Loglike null model:        -2825.4134
 Loglike full model:        -2817.9155
 df:                           11.0000
 Likelihood ratio p-value:      0.1827
```

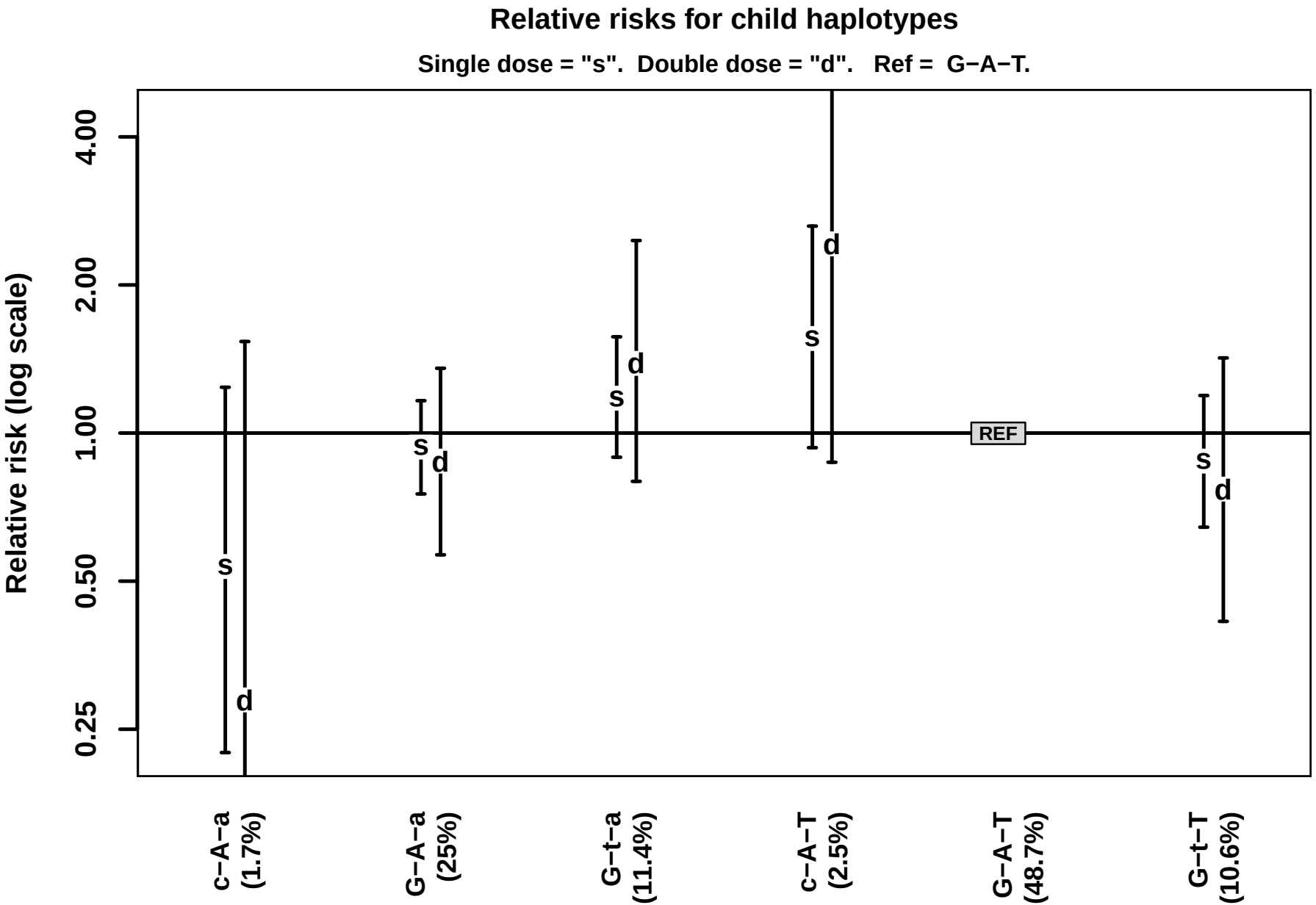
The likelihood test shows non-significant overall.
(But is less sensitive to details since number of df's is large.)

SAME HAPLIN RUN, WITH MULTIPLICATIVE RESPONSE

- Haplotypes of three first SNPs
- Impute missing
- Assume (“force”) a multiplicative dose-response model

```
result <- haplin(data = pres.data, markers = 1:3,
 use.missing = T, response = "mult",
   reference = "ref.cat")
haptable(result)
```

HAPLIN OUTPUT: EFFECT ESTIMATES



A NOTE ON REFERENCE CATEGORIES IN HAPLIN

Reference category:

- We have used `reference = "ref.cat"`
- Haplin uses the most frequent haplotype as reference
- All relative risks are against this reference

Reciprocal reference:

- `reference = "reciprocal"` is the default in Haplin
- Haplin compares each haplotype with a weighted average of the rest
- Natural in many settings, *but:*
  - Not available in the diallelic setting
  - Not available when `response = "mult"`
  - Not (yet) available when `poo = T`

Population reference:

- `reference = "population"`
- Haplin compares each haplotype with the overall population risk
- Similar to `reciprocal`
  - Not available in the diallelic setting
  - Not available when `response = "mult"`
  - But available when `poo = T`

HAPLOTYPES IN HAPLINSLIDE

- `haplinSlide` can be used with haplotypes
- Set `winlength` argument to, say, 3 or 4
- BUT time-consuming!

Example:

```
result <- haplinSlide(data = pres.data, markers = 1:50,
 winlength = 2, cpus = 4)
result <- haptable(result)
head(result)
```

	window	row.win	marker	alleles	counts	HWE.pv	Original	After
1	rs1-rs3	1	rs1	c/G	114/2682	0.8246339	559	
2	rs1-rs3	2	rs3	A/t	2191/605	0.2340155	559	
3	rs1-rs3	3	<NA>	<NA>	<NA>	NA	559	
4	rs3-rs5	1	rs3	A/t	2191/605	0.2340155	559	
5	rs3-rs5	2	rs5	a/T	1048/1748	0.2266205	559	
6	rs3-rs5	3	<NA>	<NA>	<NA>	NA	559	