Analysis by Dr. Susanne Weller

These analyses is still preliminary!

Graphs should not be used for presentations and publications!

Dr. Susanne Weller, 05/06/2014
HICF1_Objectives for publication:

(1) Present all relevant genetic data in an association chart
(−> will be done once we decide which are the important ones)

(2)Built a tree model with only genetic data

− 2a) assess Missclassification error

− 2b) assess quality (variance) via bootstrapping

− 2c) built a couple of trees

(3) Built a tree with only clinical data

− 3a) assess Missclassification error

− 3b) assess quality (variance) via bootstrapping

− 3c) built a couple of trees

(4) Built a tree with clinical data and genetic data

− 4a) built model

− 4b) assess quality via bootstrapping

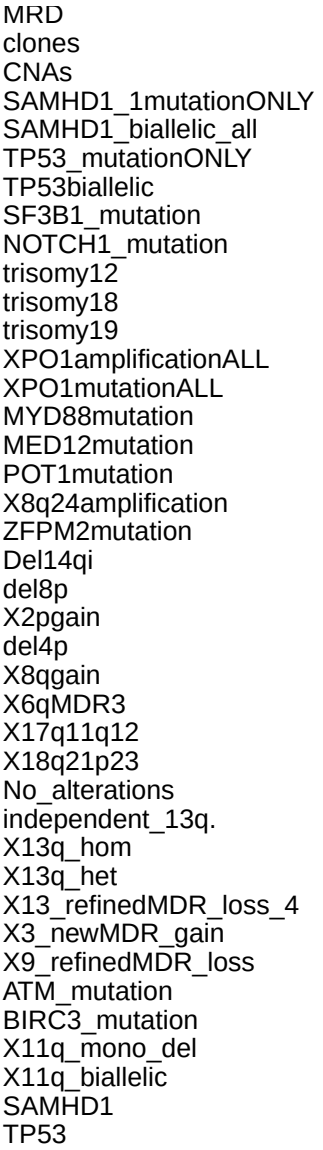
(5) Final comparison

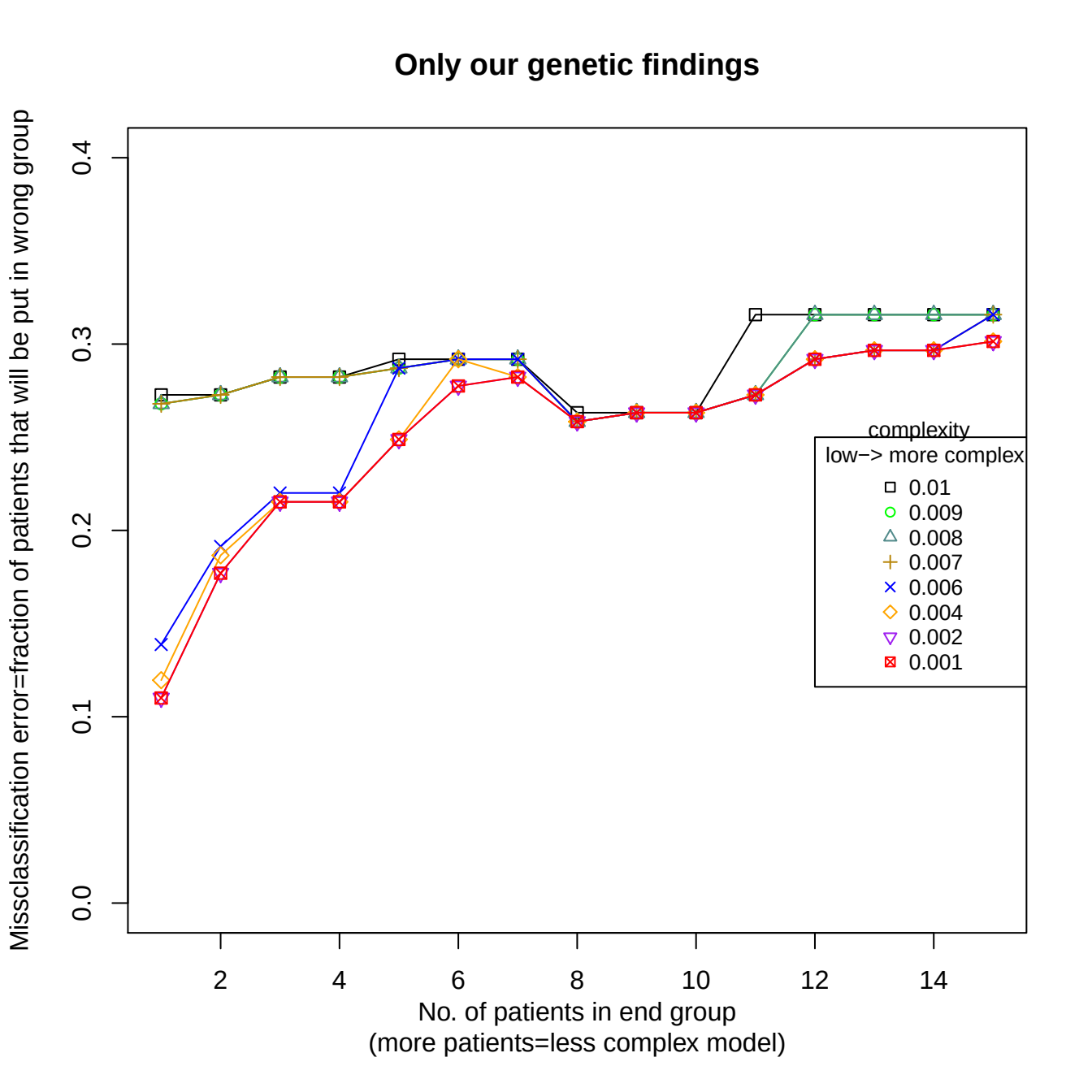
− How much did we improve forecasting of MRD+/− with genetics?

− How well do both models perform seperately?

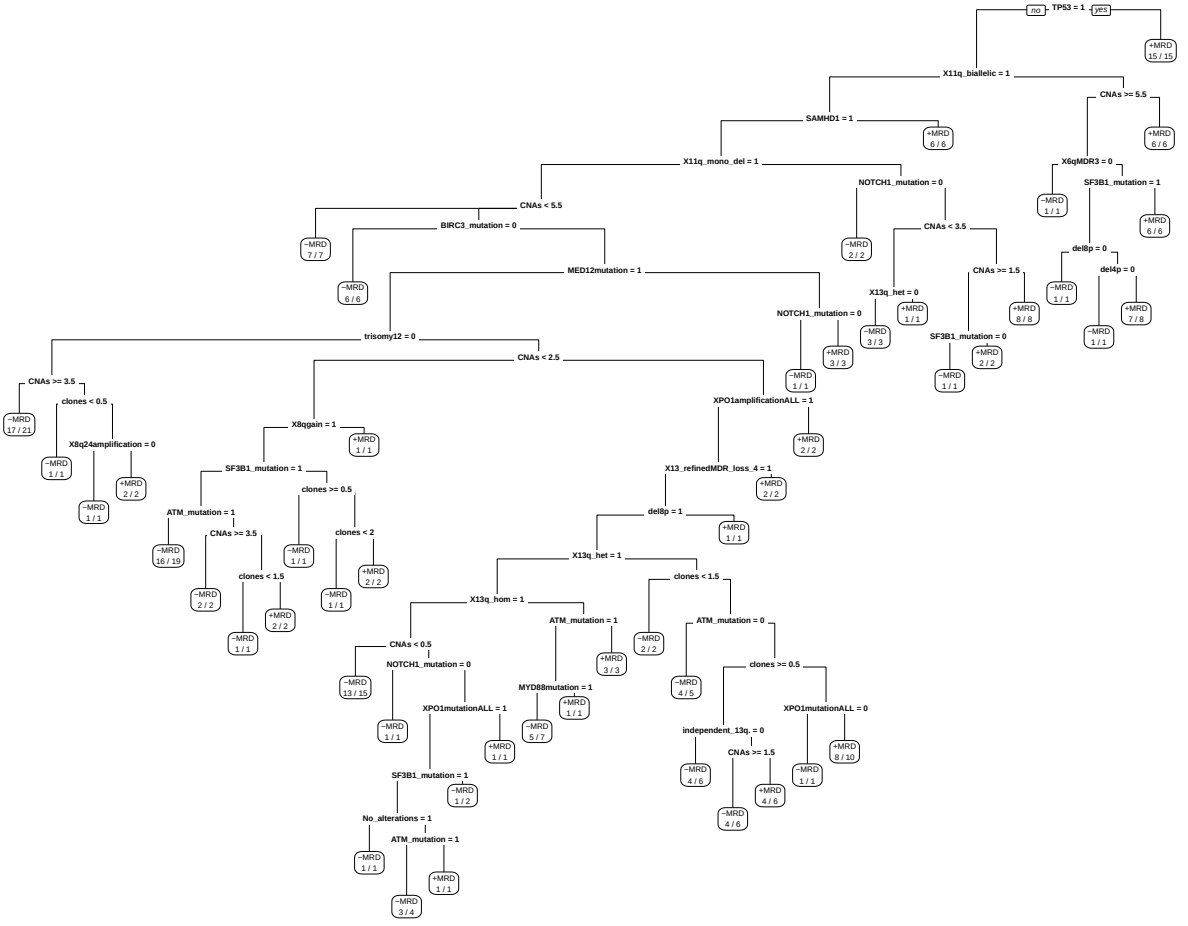
− Are there certain subgroups of patients that benefit particularly from genetic testing?

(2) Built a tree with only our genetic data



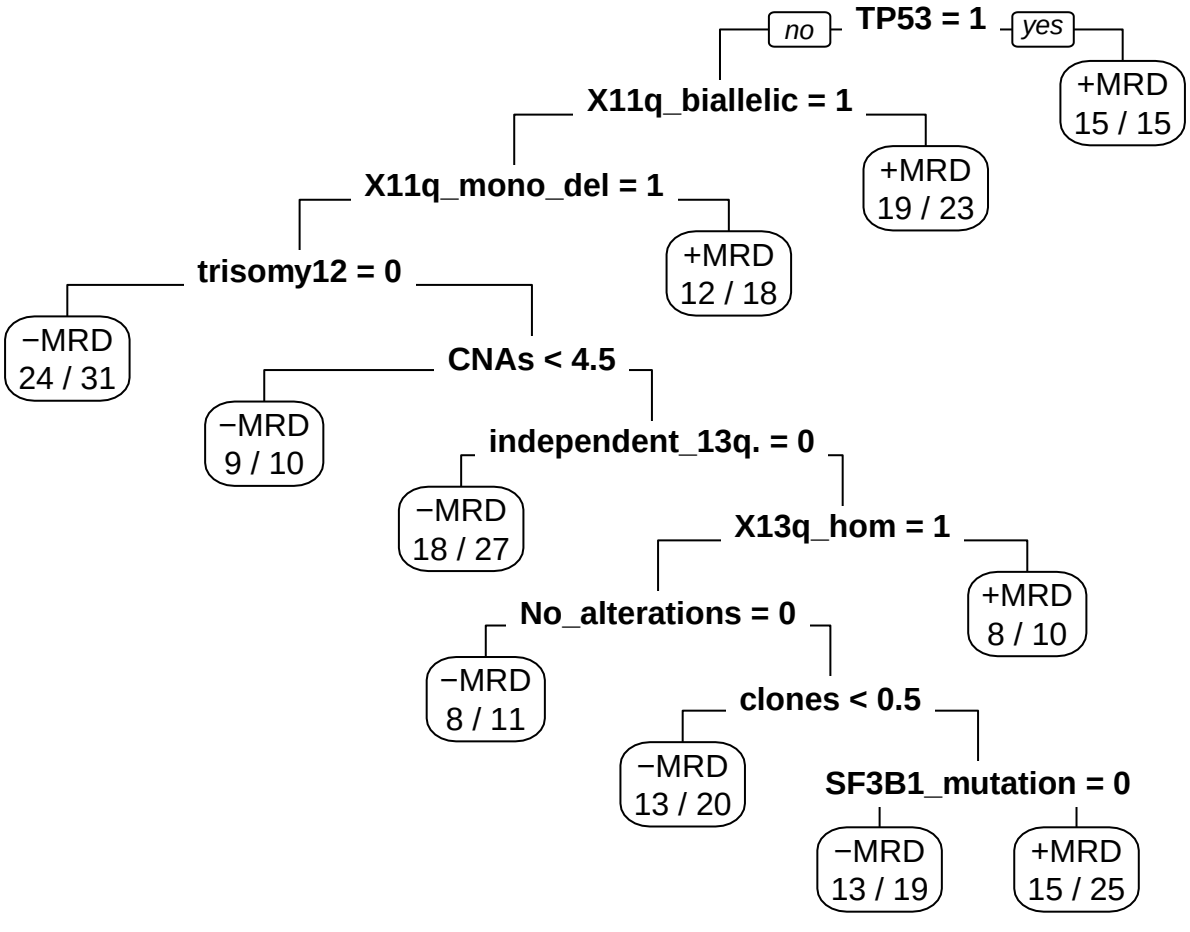

Only our genetic findings
Missclassification error=fraction of patients that will be put in wrong group
0.0
0.1
0.2
0.3
0.4
2
4
6
8
10
12
14
No. of patients in end group
(more patients=less complex model)
complexity
low−> more complex
0.01
0.009
0.008
0.007
0.006
0.004
0.002
0.001

2c) tree1: Only our genetic findings
cp=0.001, endgroup=1, MissClassErr~11%



This is the best tree with only genetics, but highly specific to our data!

2c) tree2:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%

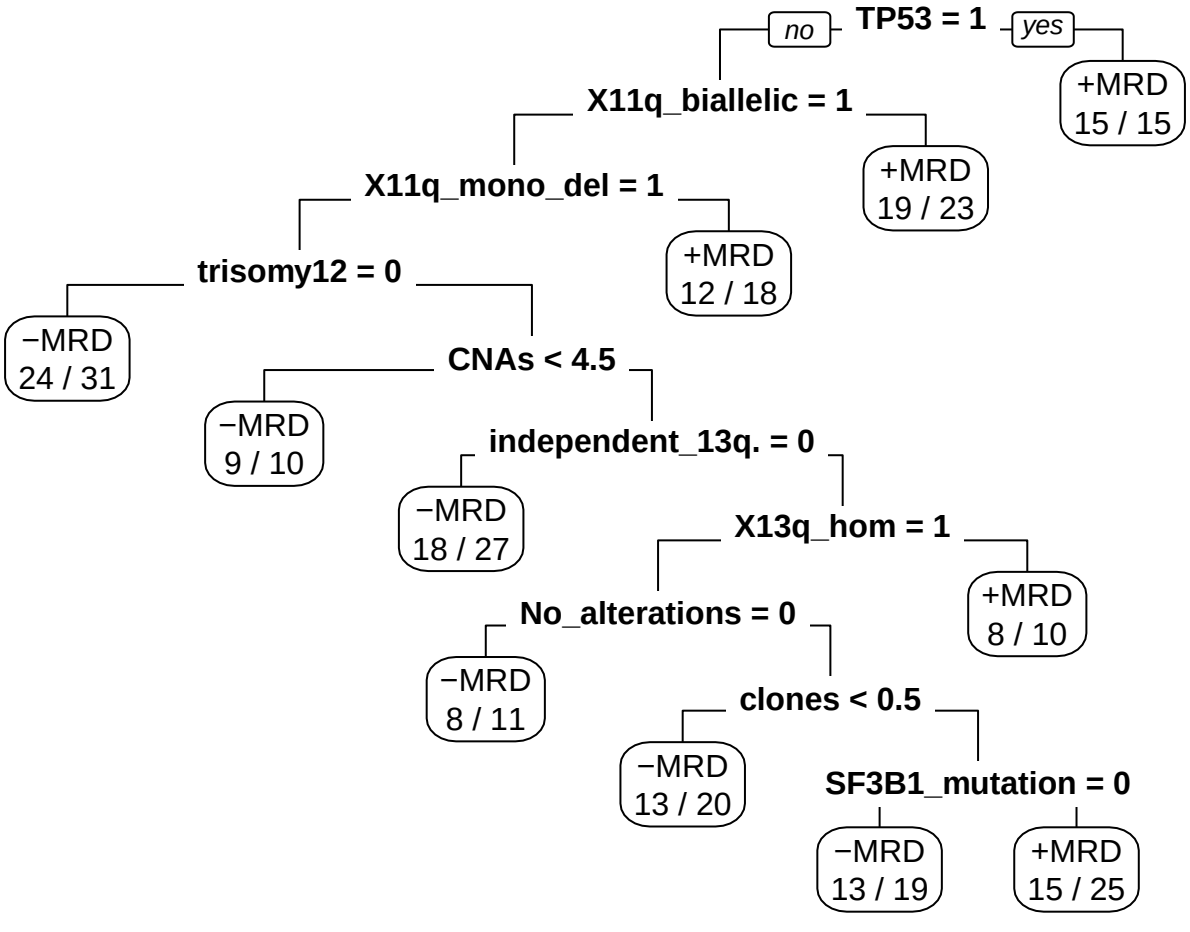


This is a fairly general reasonable tree,
parameters deducted from where all error curves coerce into one point.
Please check if it makes sense in a clinical setting

This gives an overview about tree aesthetics:

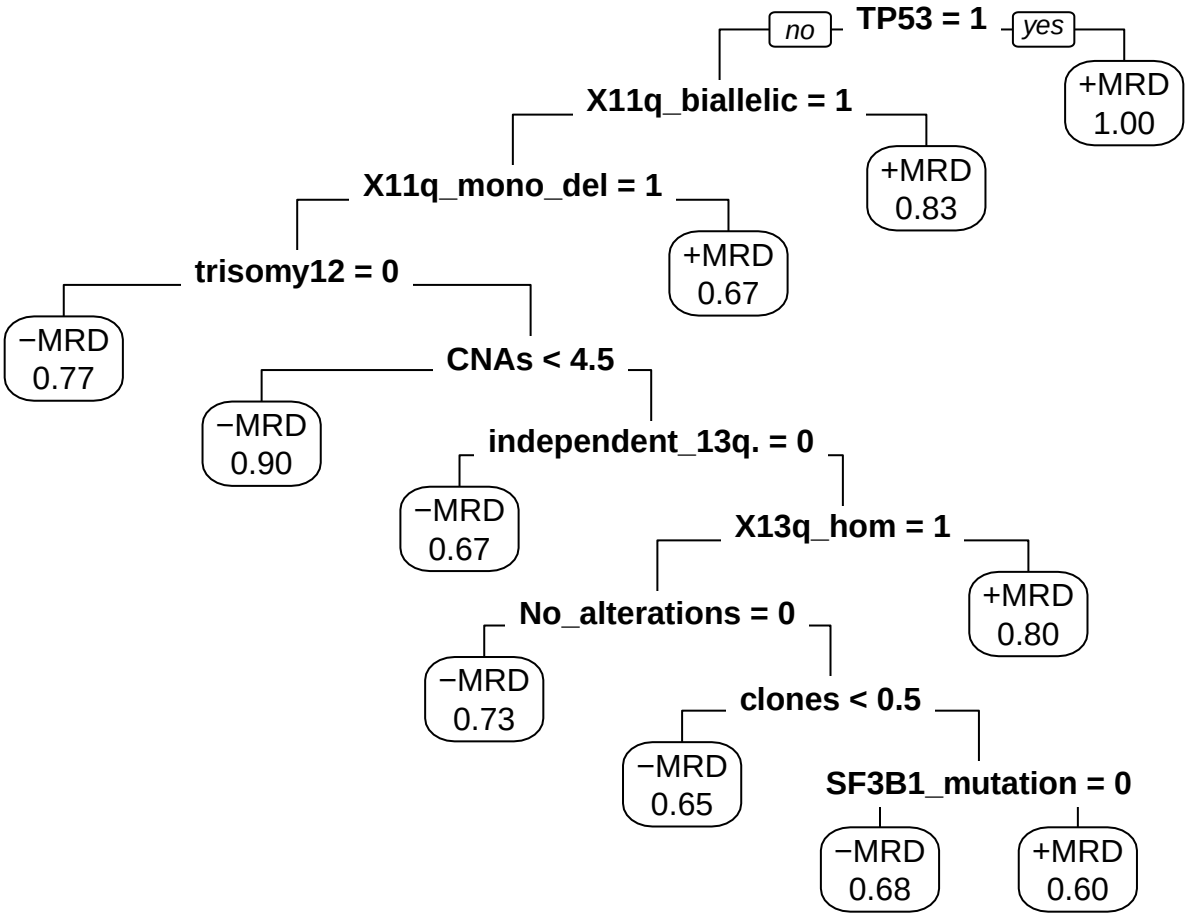
There are around 100 more options to costumise the tree,
just give me any ideas you have, and I'll implement them!

2c) tree3: Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



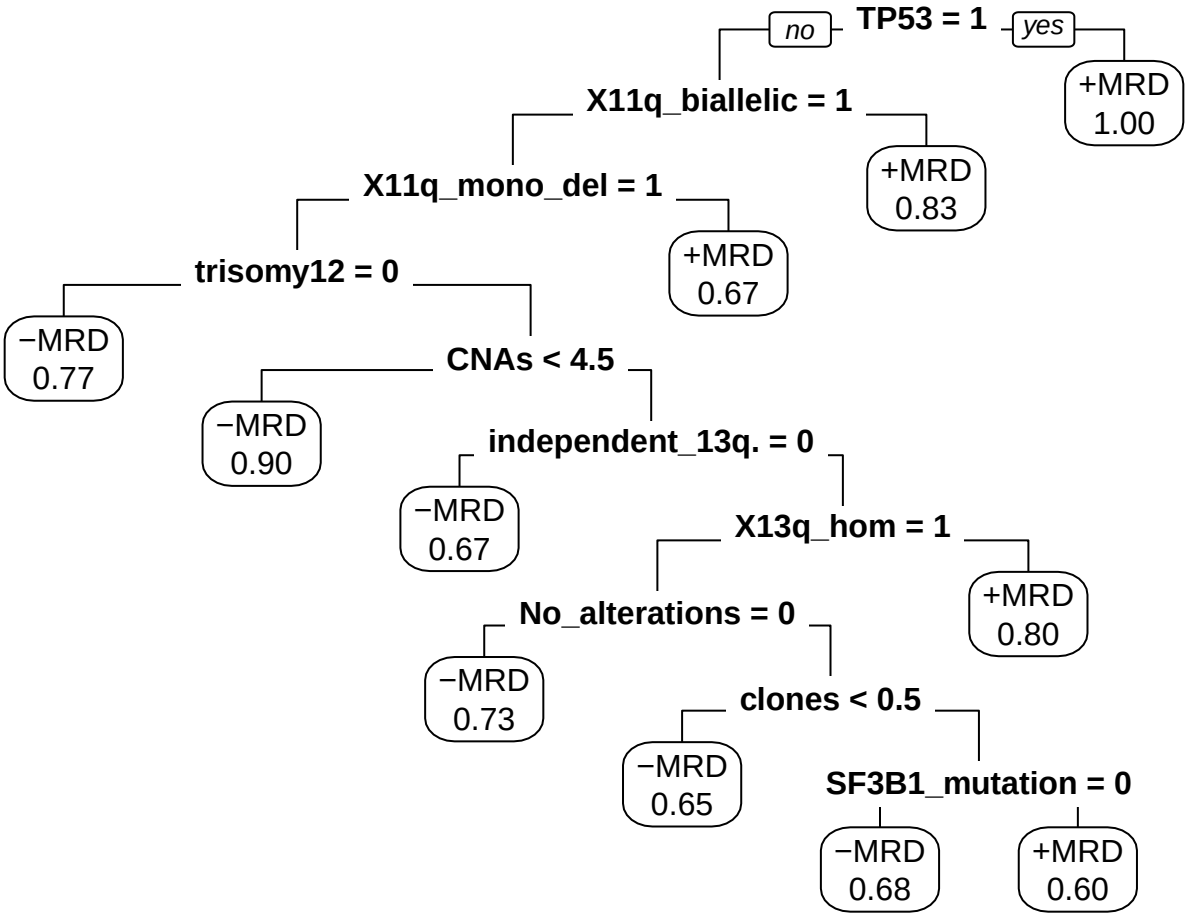
All branches same length, total patient numbers displayed

2c) tree4:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



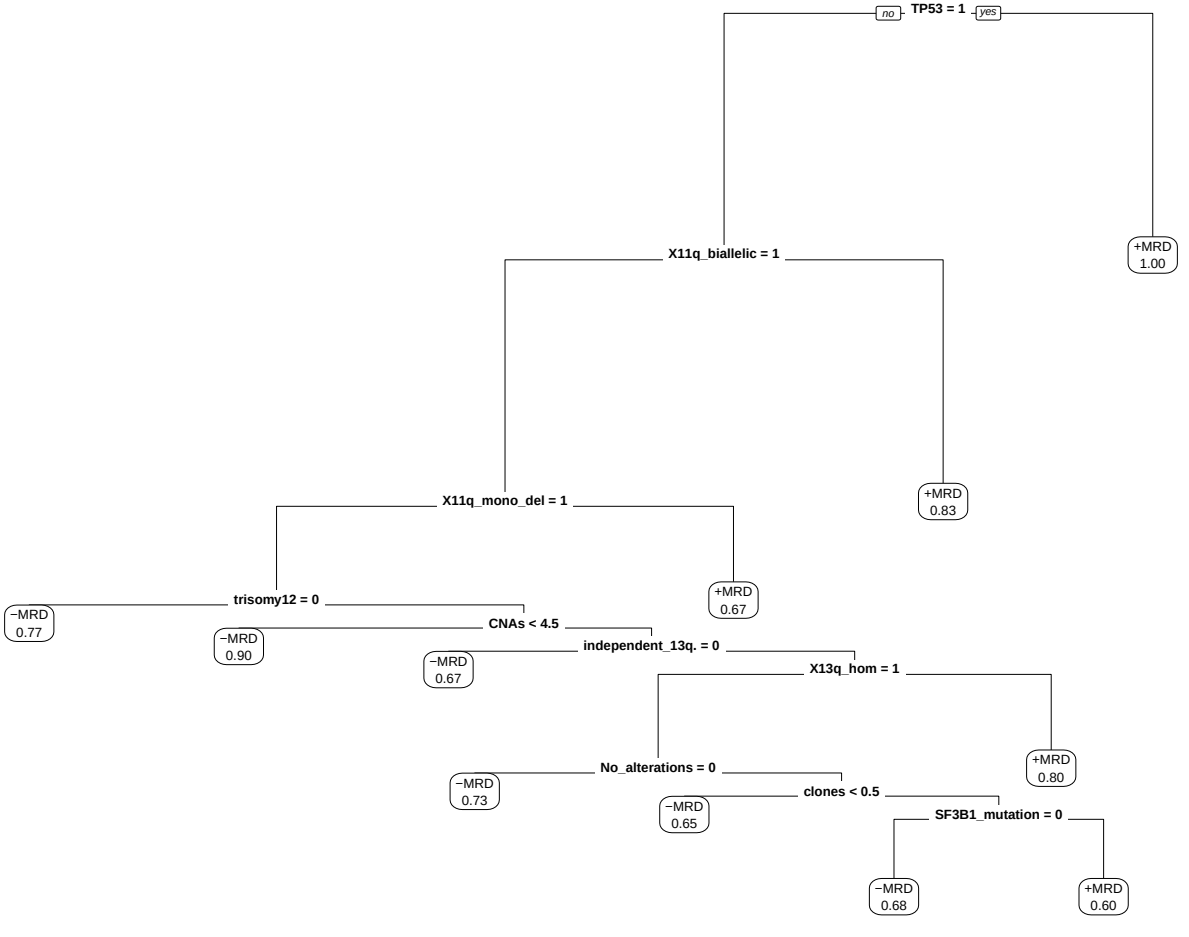
All branches same length, group correct classification rate displayed

2c) tree5:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



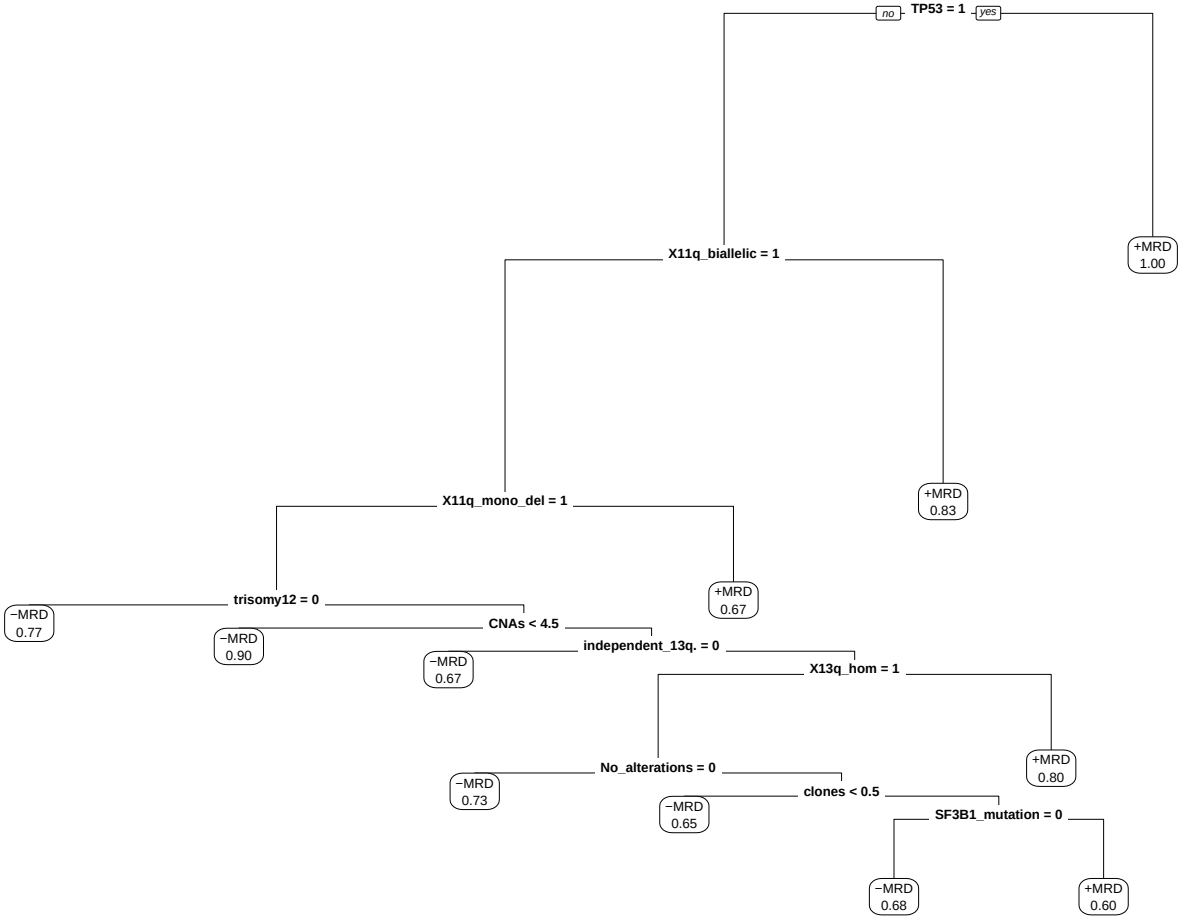
All branches same length, group correct classification rate displayed

2c) tree6:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



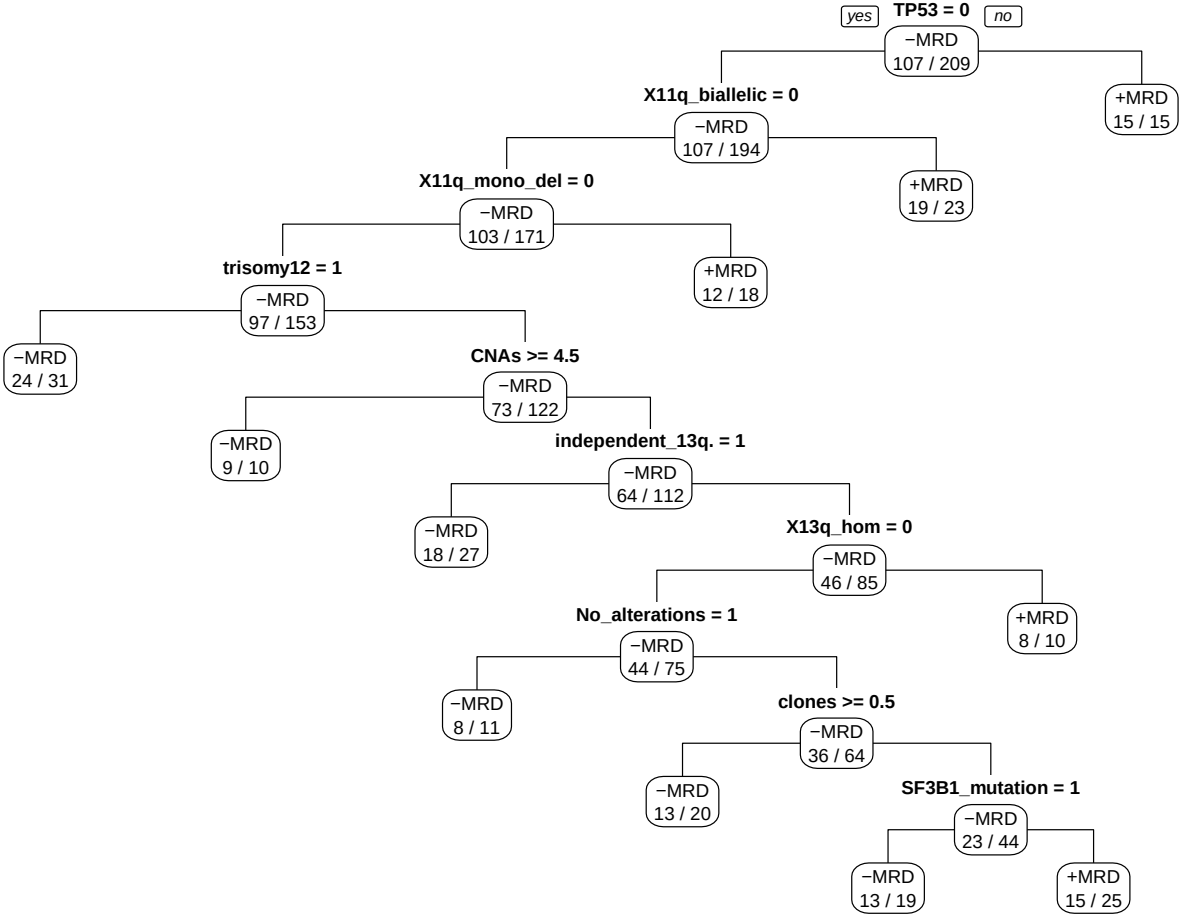
Branch length proportional to fit of model, group correct classification rate displayed

2c) tree7:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



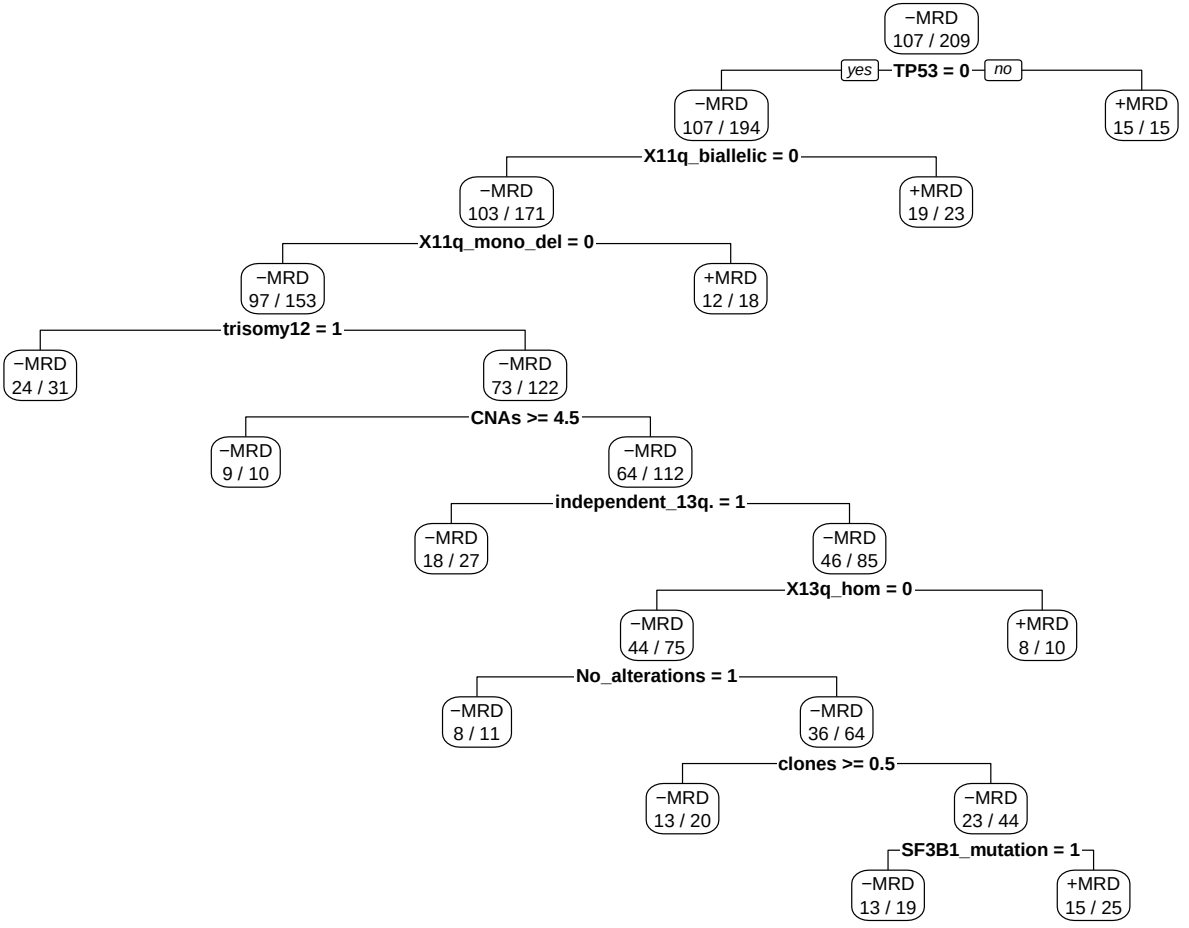
Branch length proportional to fit of model, group correct classification rate displayed

2c) tree8:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



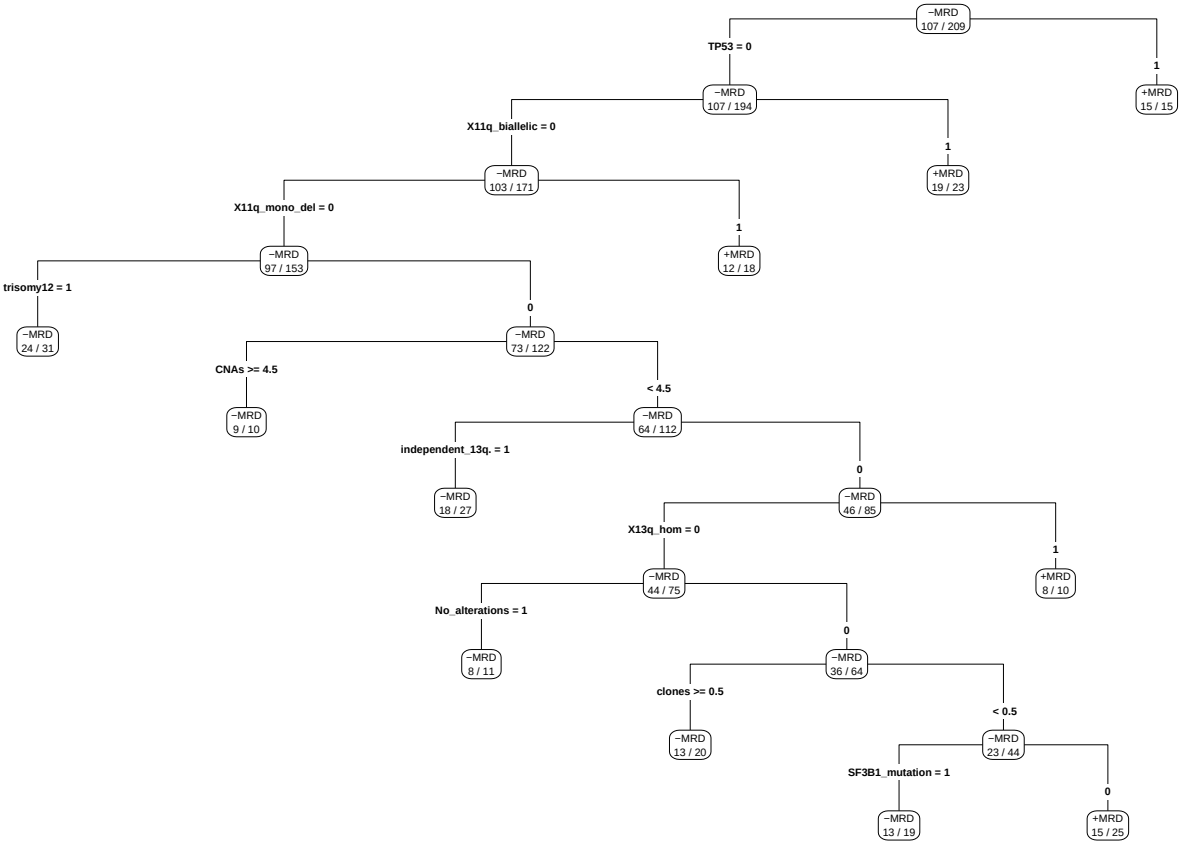
All branches same length, total patient numbers displayed, individual node labelled (type1)

2c) tree9:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



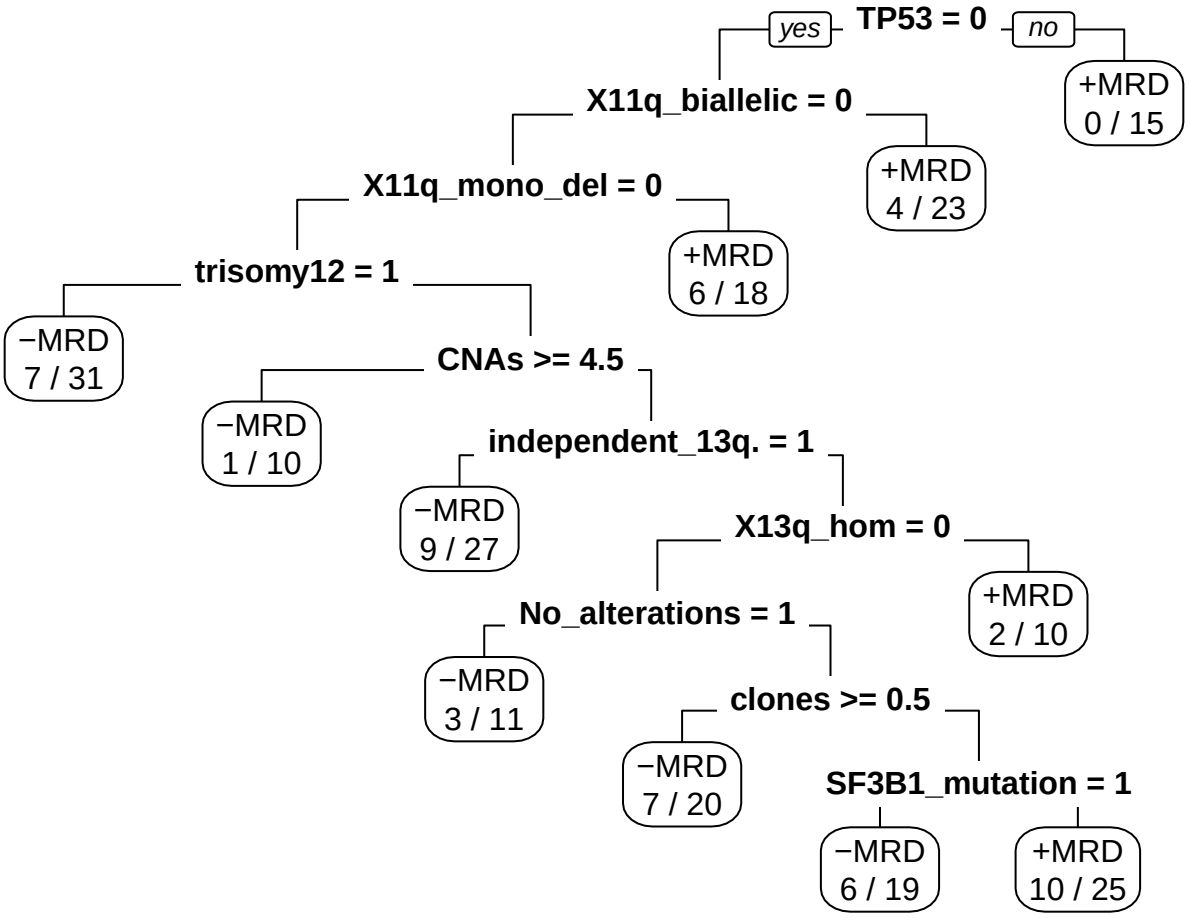
All branches same length, total patient numbers displayed, nodes labelled above split (type2)

2c) tree10:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



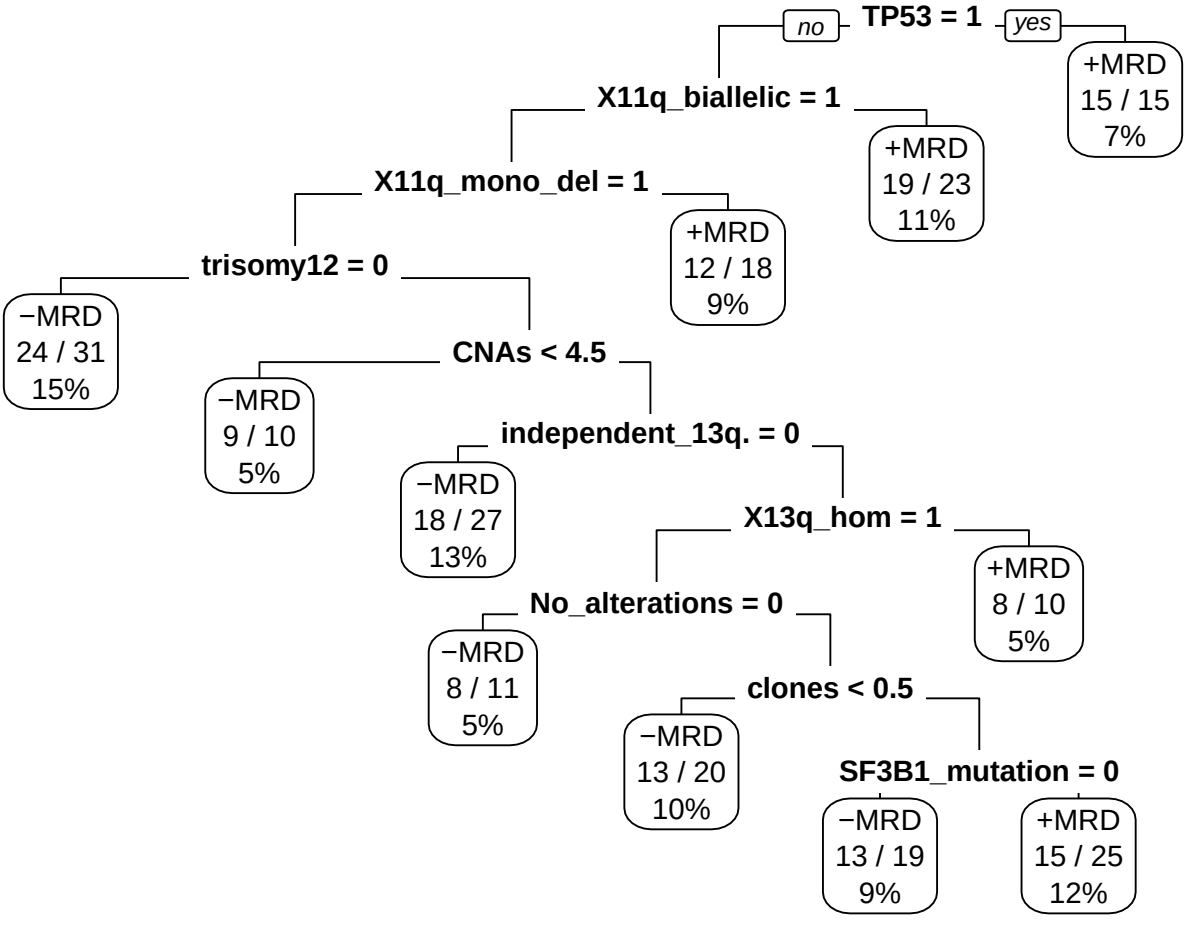
All branches same length, total patient numbers displayed, nodes and branches labelled (type4)

2c) tree11:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



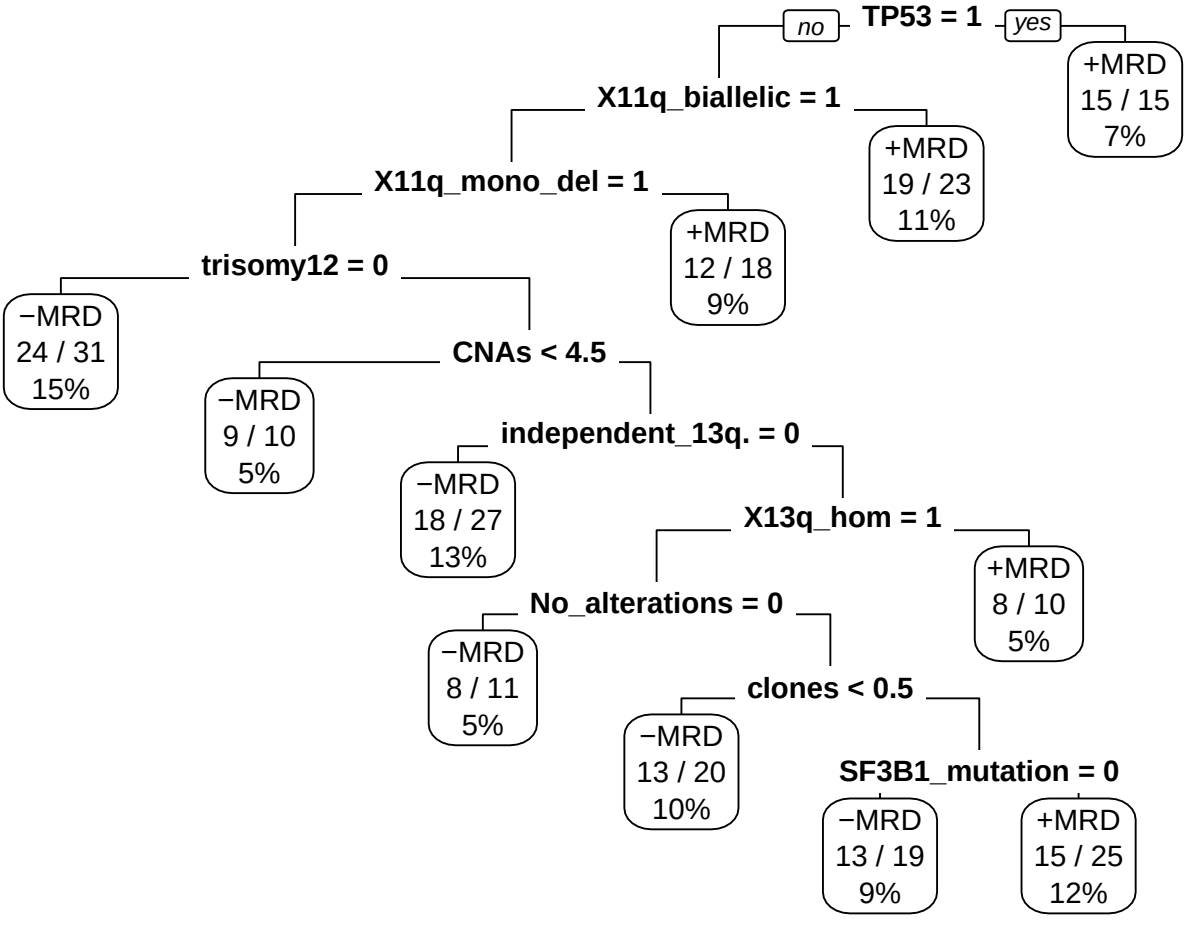
All branches same length, missclassified patients displayed (extra=3)

2c) tree12:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



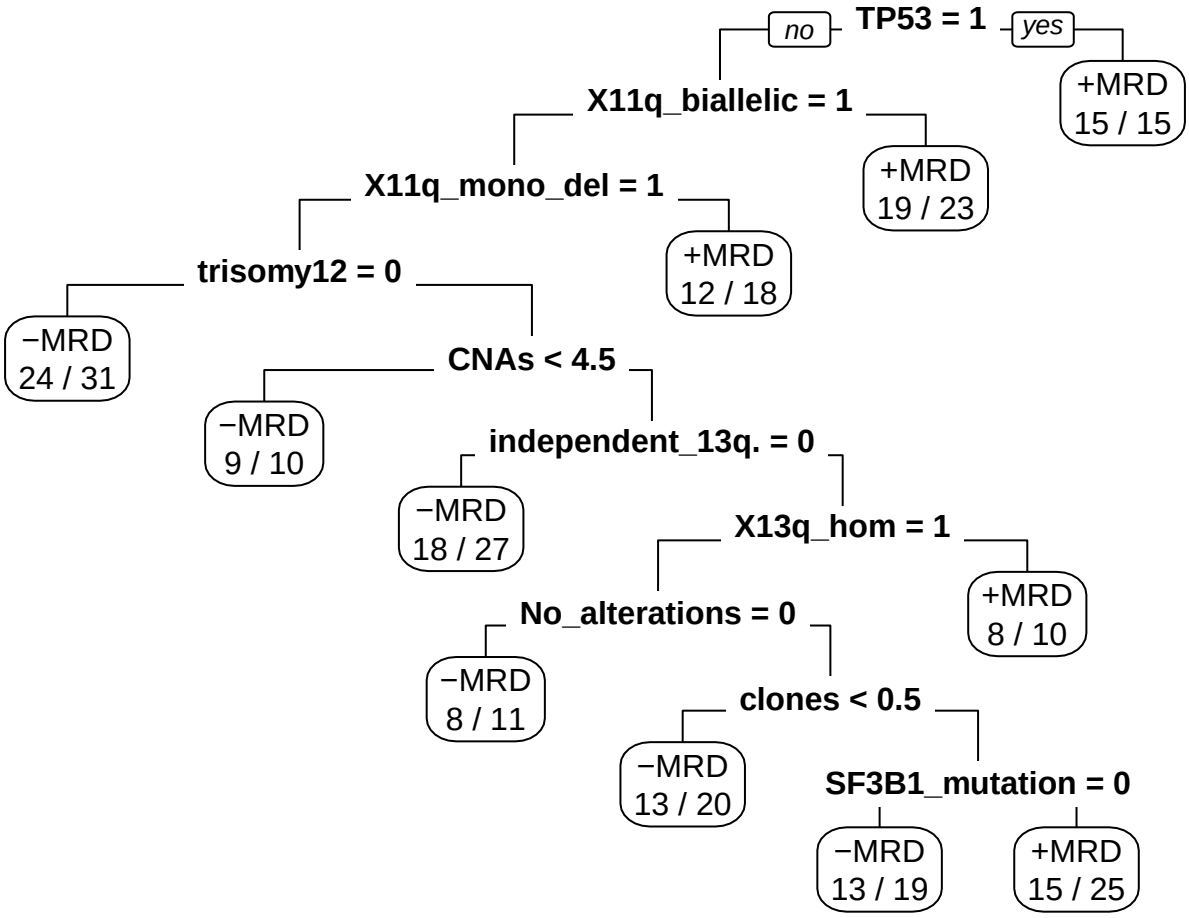
All branches same length,
correctly classified patients & total percentage in dataset displayed (extra=102)

2c) tree13:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



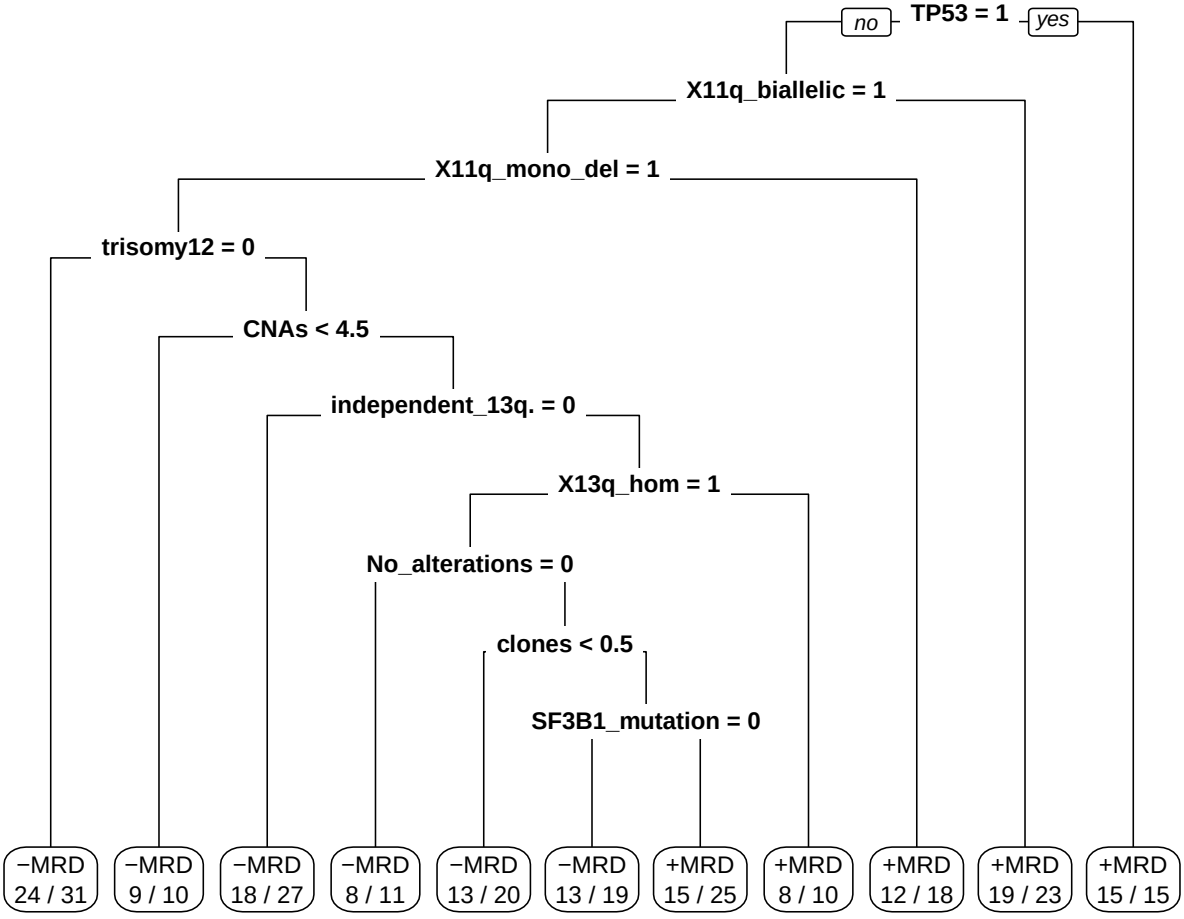
All branches same length,
correctly classified patients & total percentage in dataset displayed (extra=102)

2c) tree14:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



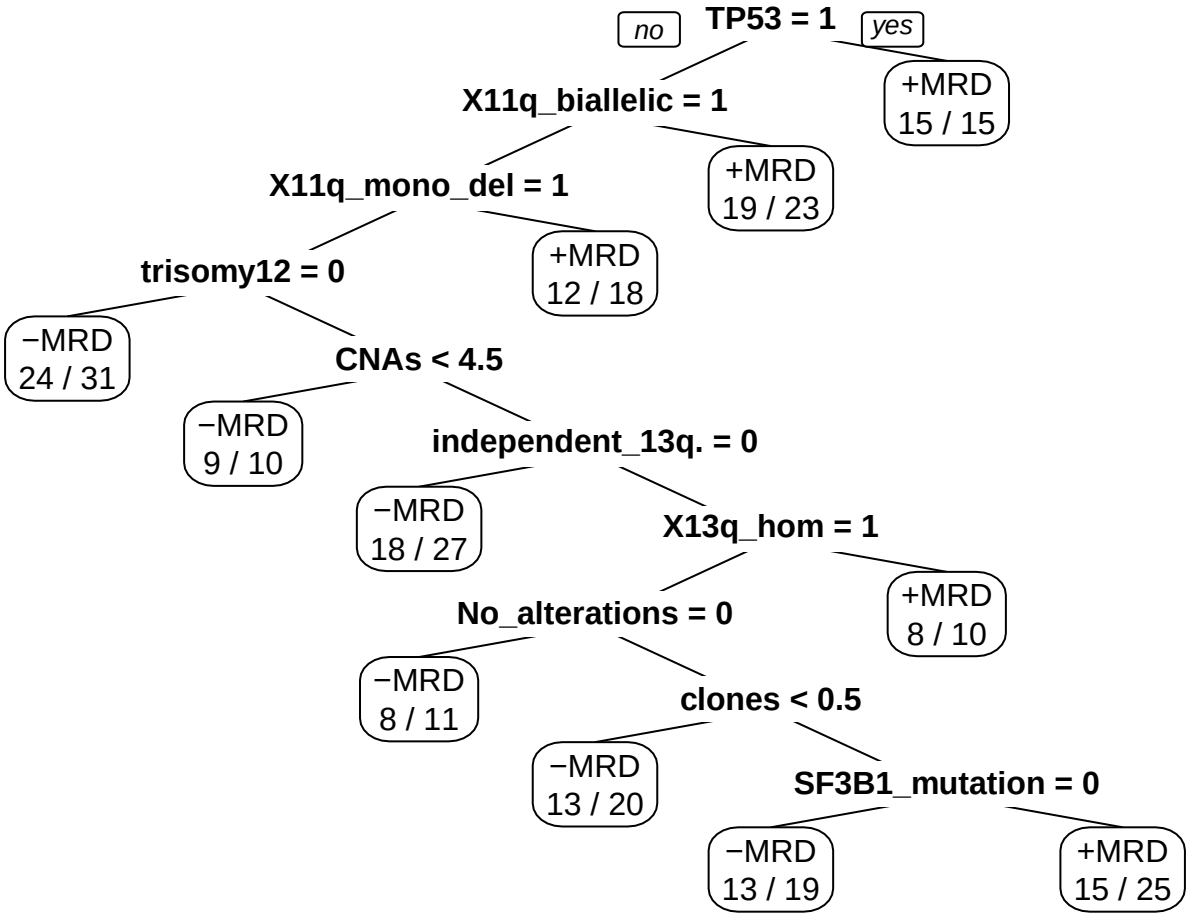
All branches same length, total patient numbers displayed

2c) tree15:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



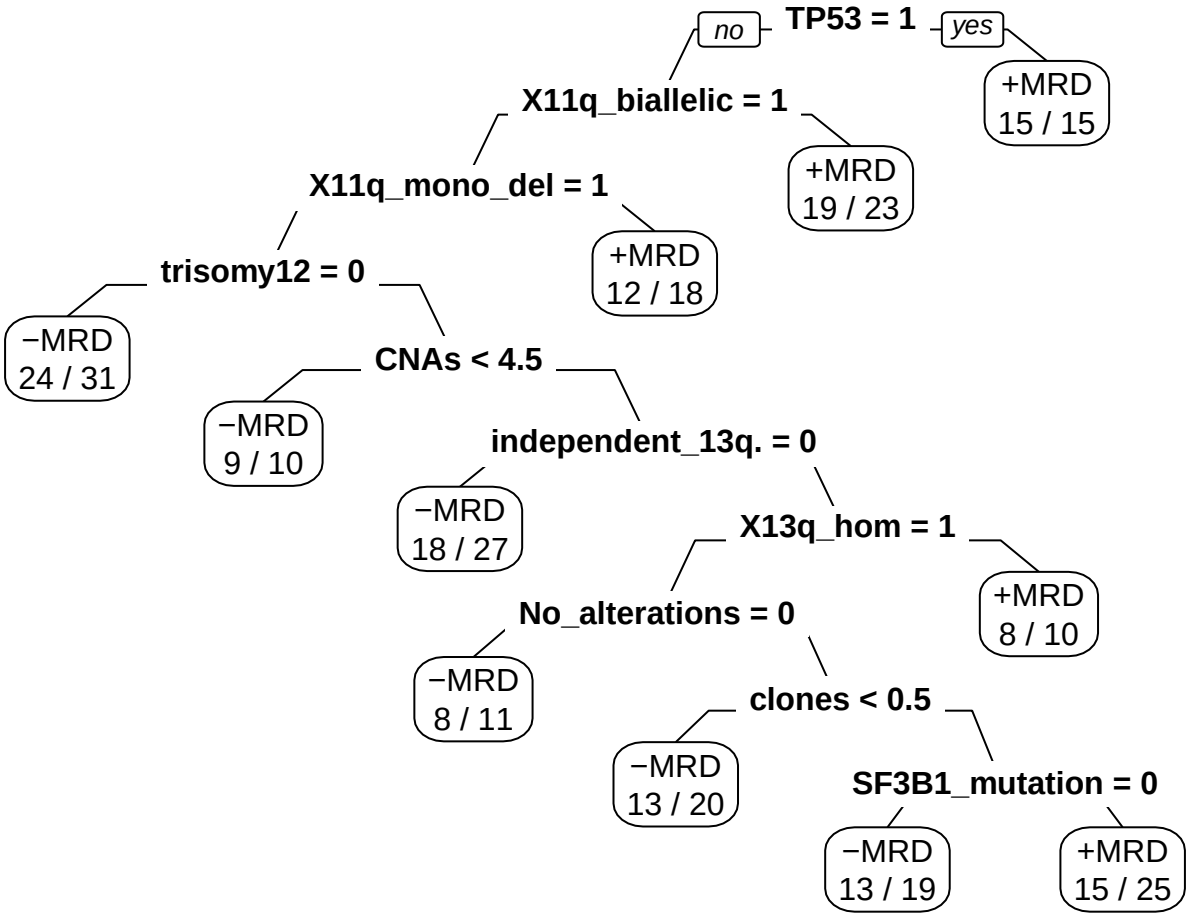
All branches same length, total patient numbers displayed, 'fallen leaves'

2c) tree16:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



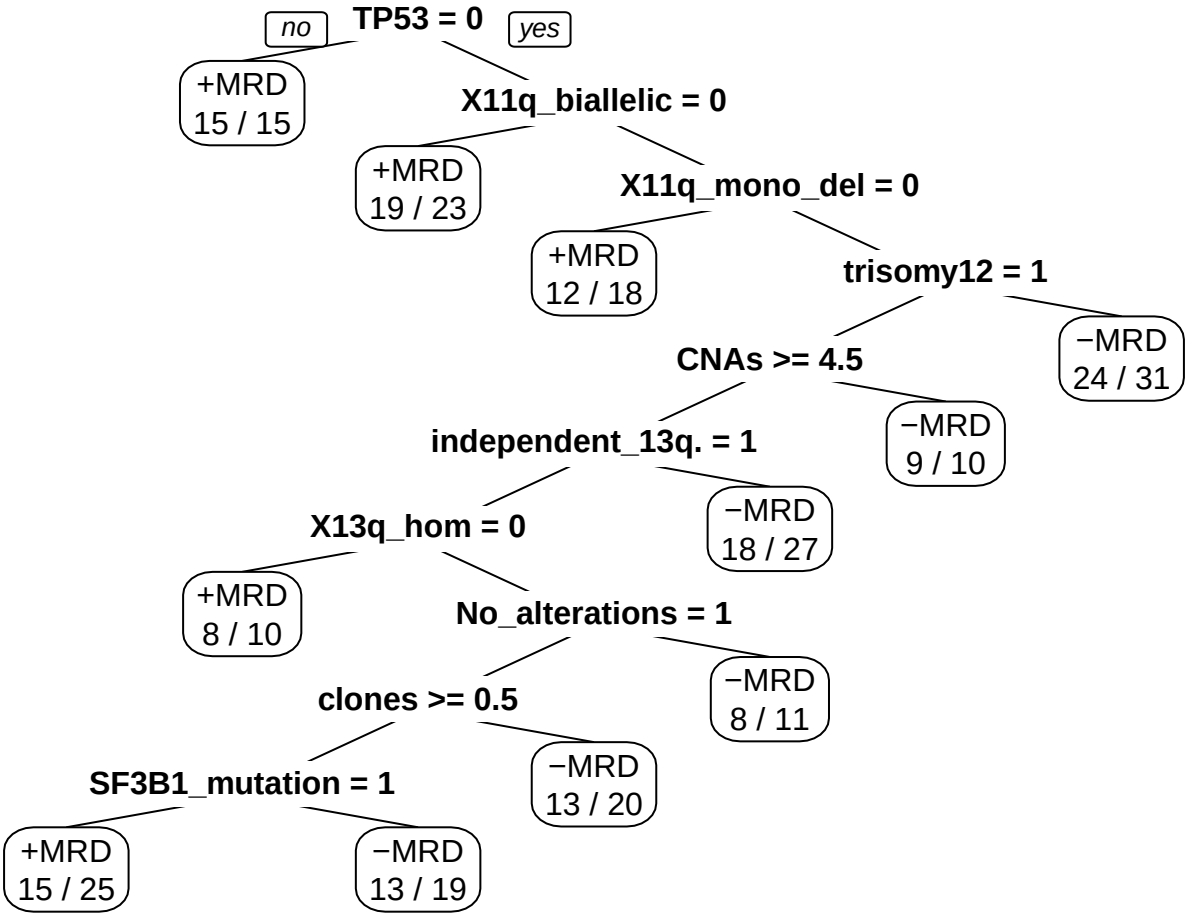
All branches same length, total patient numbers displayed, 'branch=0'

2c) tree17:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



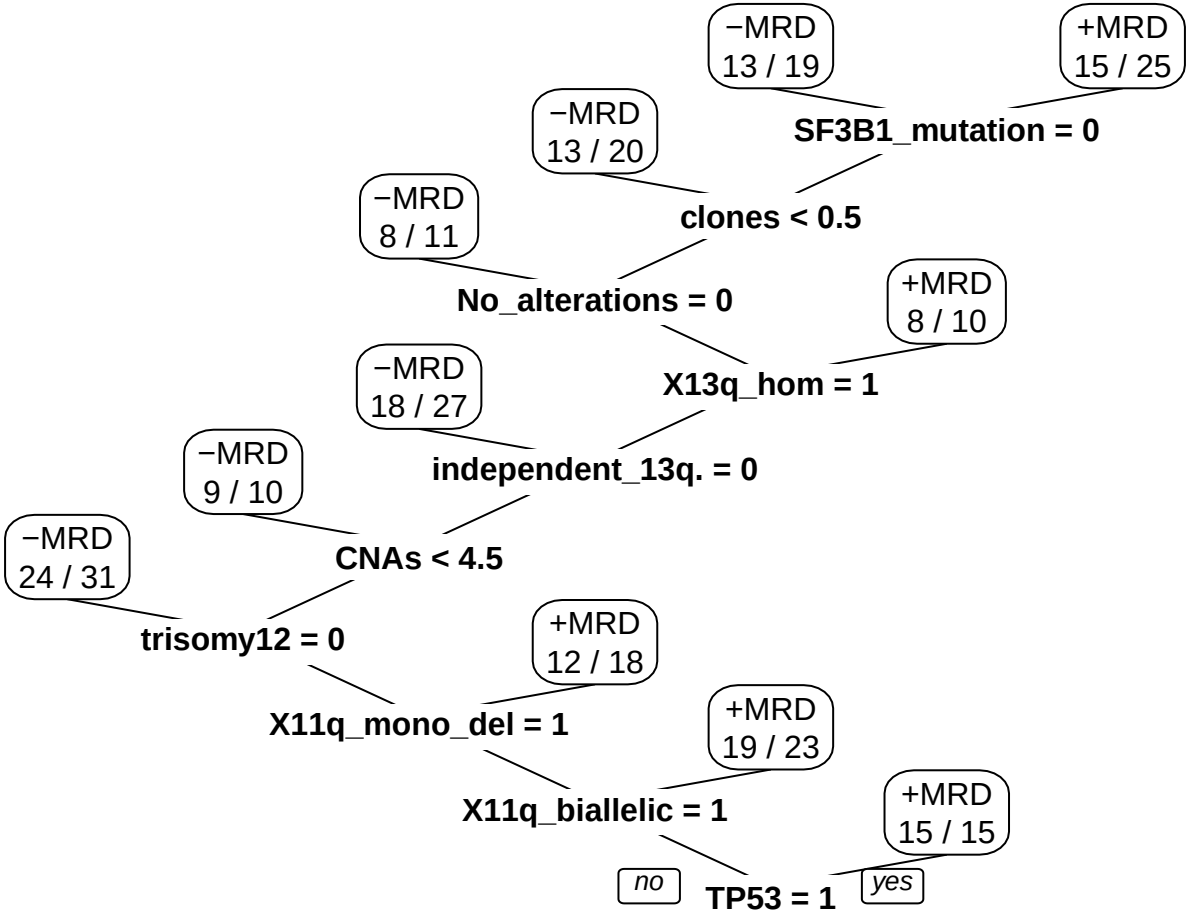
All branches same length, total patient numbers displayed, 'branch=0.8'

2c) tree18:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%



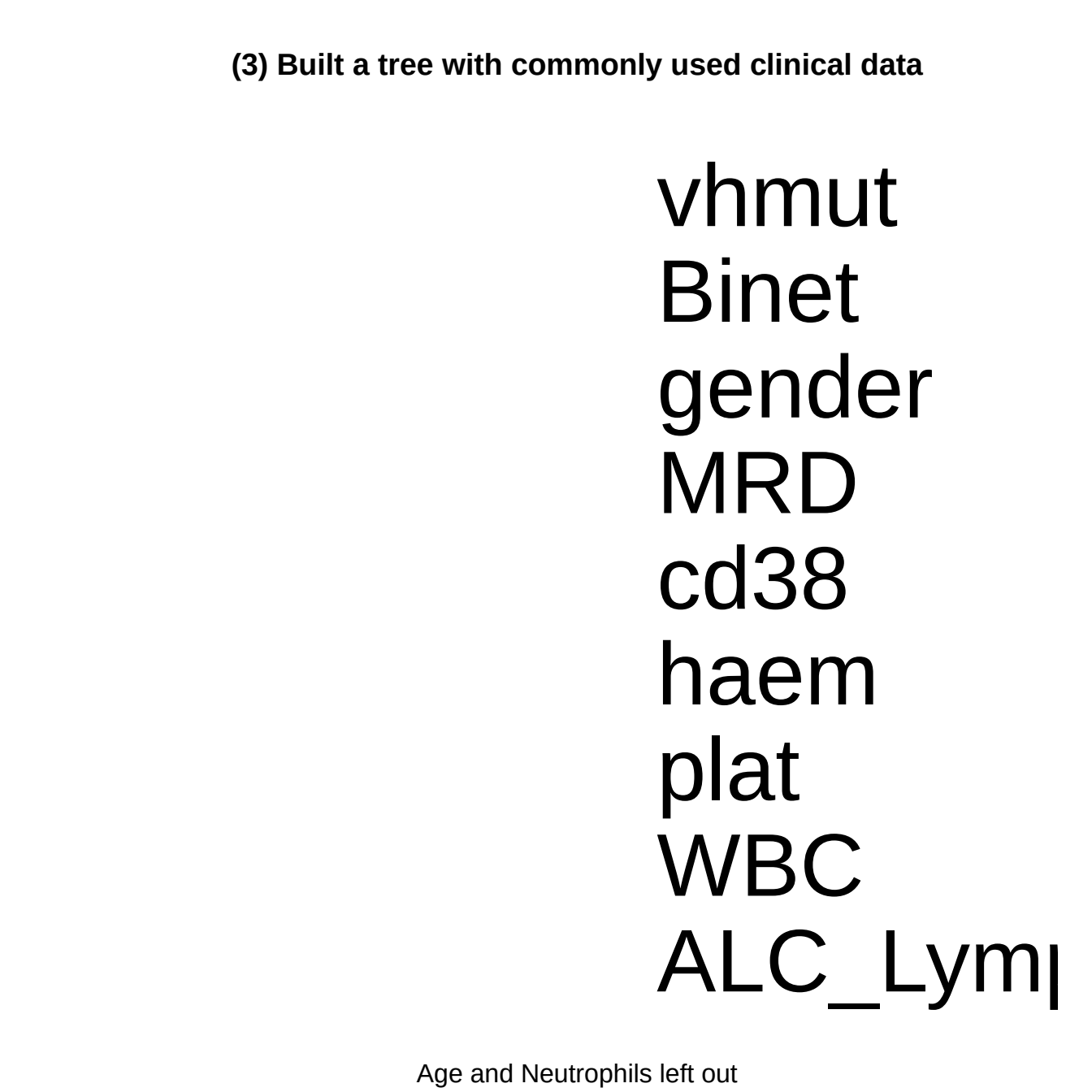
All branches same length, total patient numbers displayed, 'xflip=TRUE'

2c) tree19:Only our genetic findings
cp=0.01, endgroup=8, MissClassErr~27%

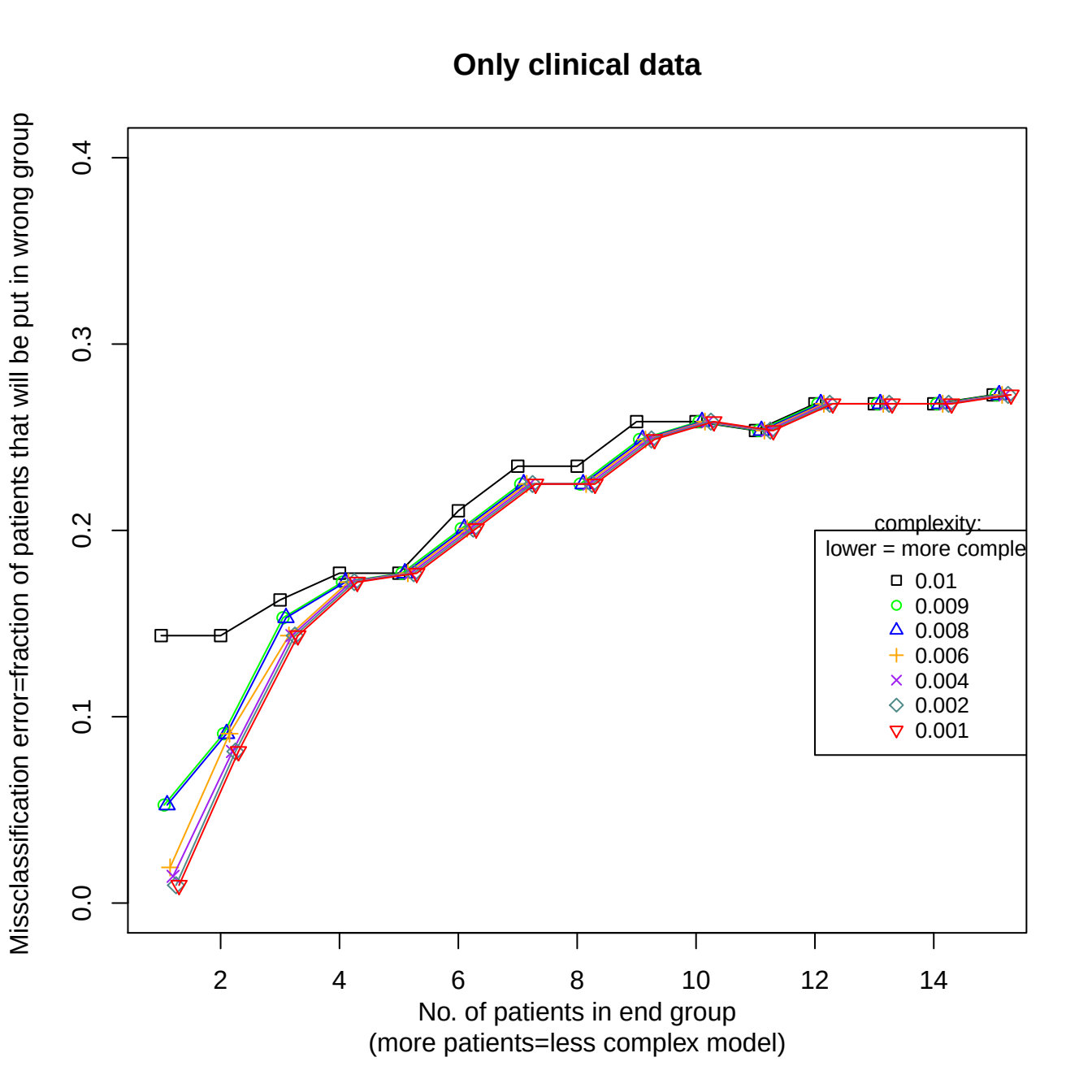


All branches same length, total patient numbers displayed, 'yflip=TRUE'

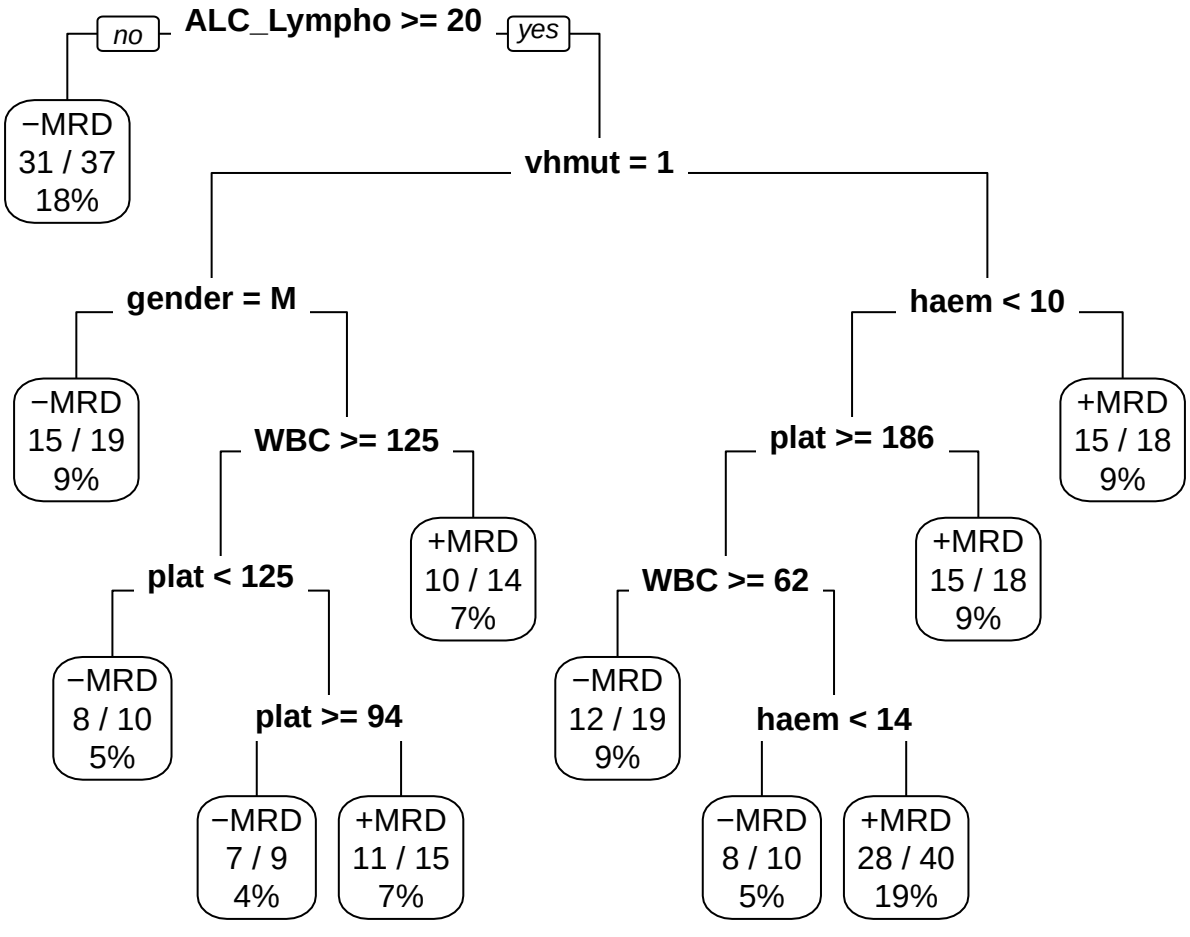
(3) Built a tree with commonly used clinical data



Age and Neutrophils left out


Only clinical data
Missclassification error=fraction of patients that will be put in wrong group
0.0
0.1
0.2
0.3
0.4
2
4
6
8
10
12
14
No. of patients in end group
(more patients=less complex model)
complexity:
lower = more comple
0.01
0.009
0.008
0.006
0.004
0.002
0.001

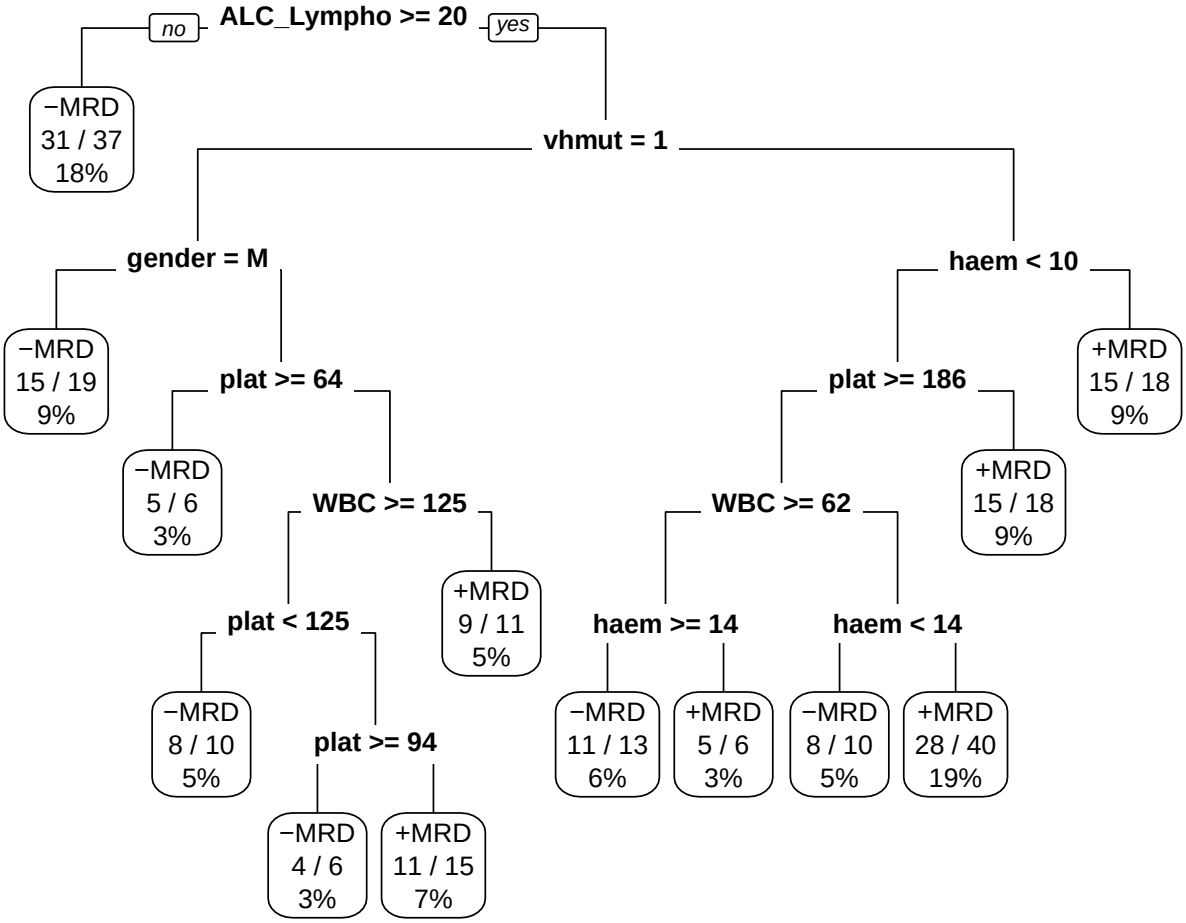
3c) tree20: All clinical data
cp=0.01, endgroup=8, MissClassErr=23.4%



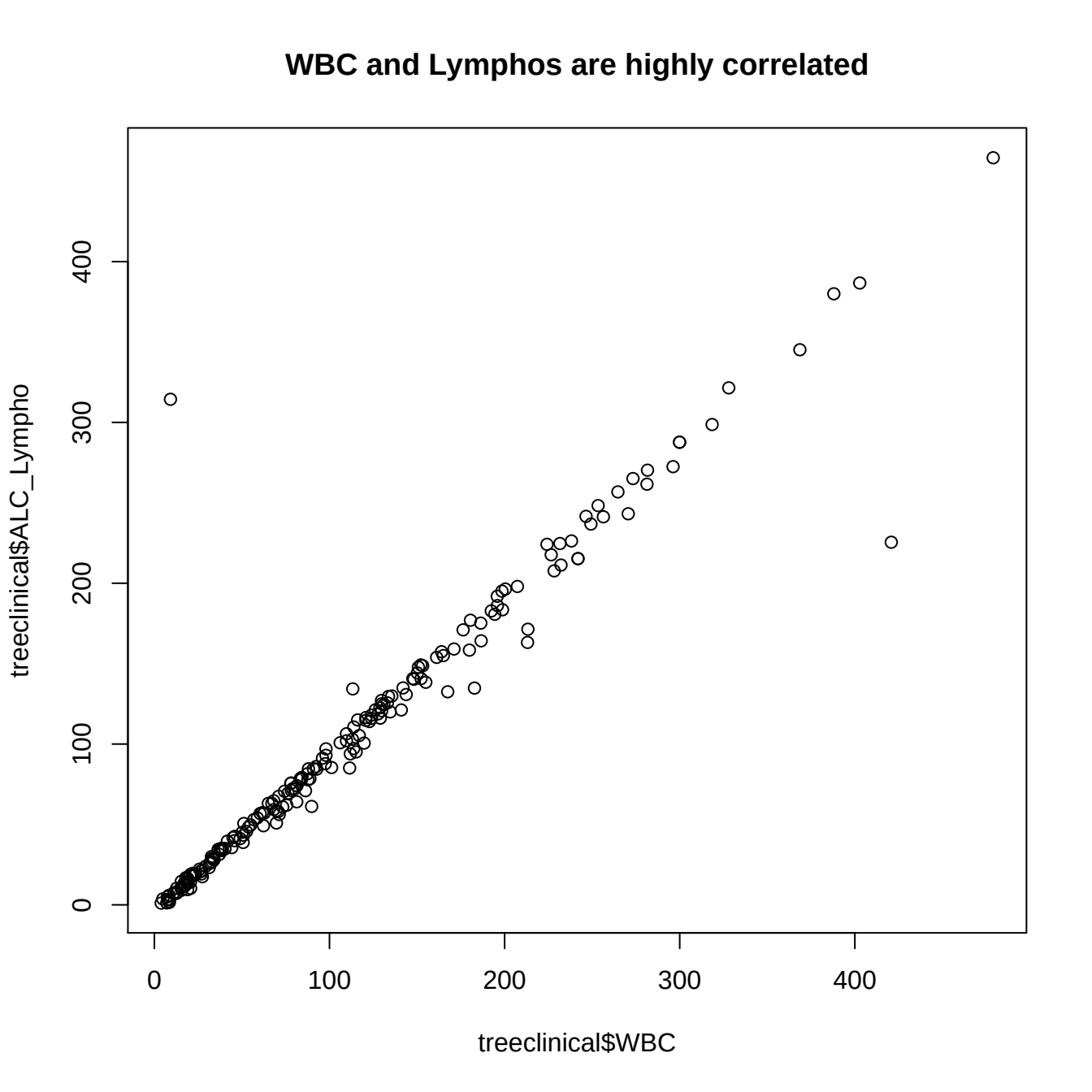
ree has quite a good Missclassification error, but does not make much sense clini

Note:vhmut=1 −> 98−100% mutated (which is good?

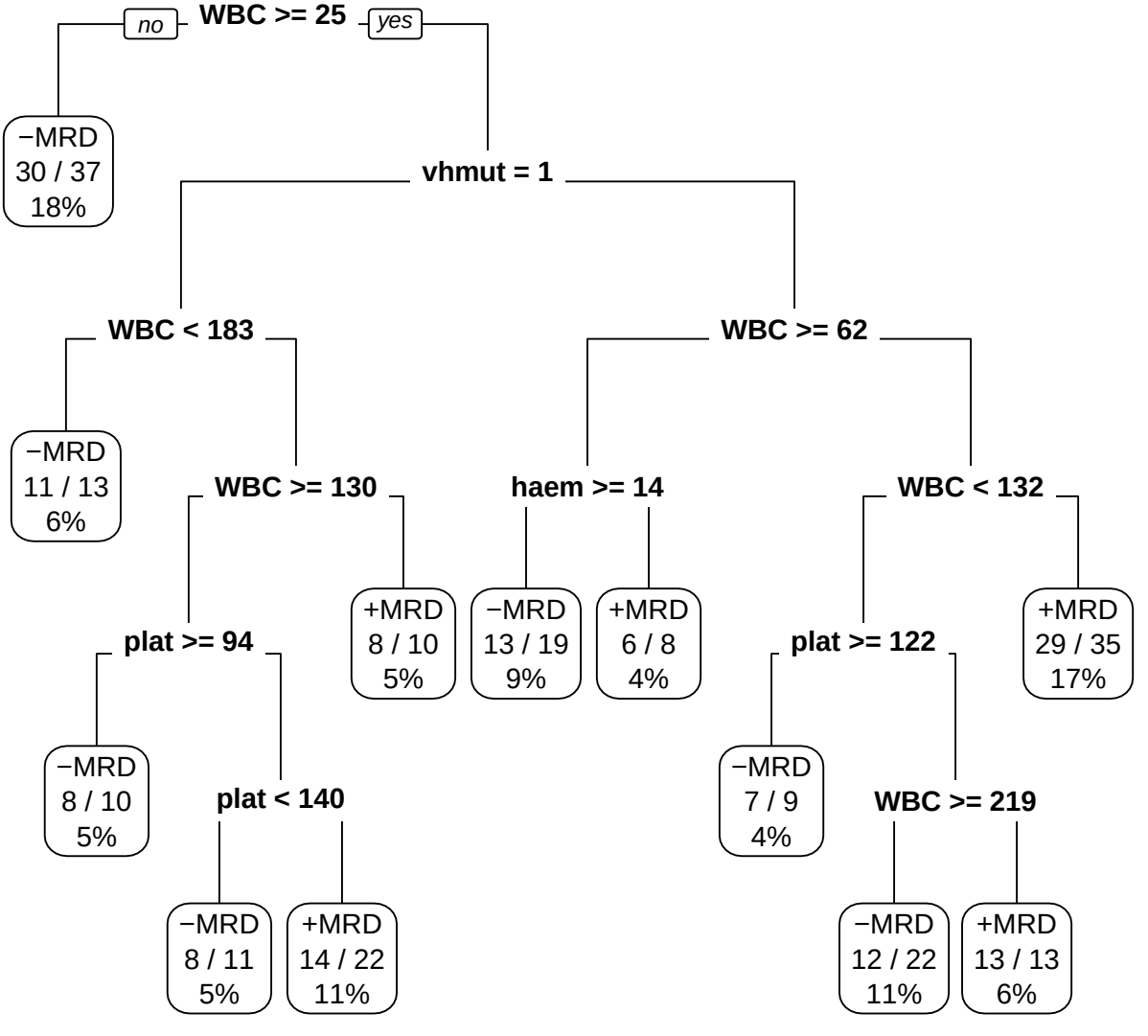
3c) tree21:All clinical data
cp=0.01, endgroup=6, MissClassErr=21%



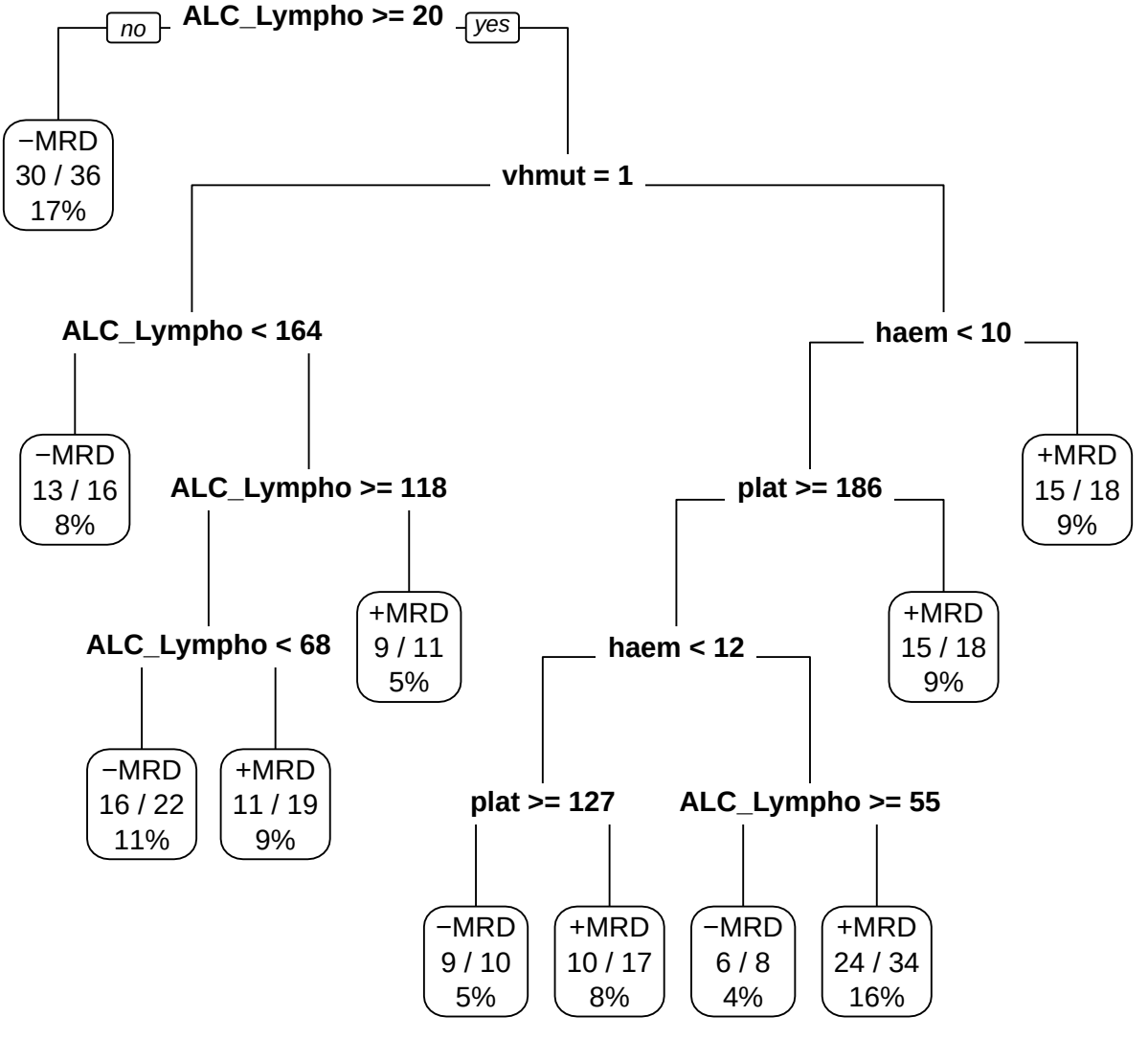
This tree has quite a good Missclassification error, but is quite complicated. One idea would be to use either Lymphos OR WBC and platelets OR haem (See next trees)


WBC and Lymphos are highly correlated
treeclinical$ALC_Lympho
0
100
200
300
400
treeclinical$WBC
0
100
200
300
400

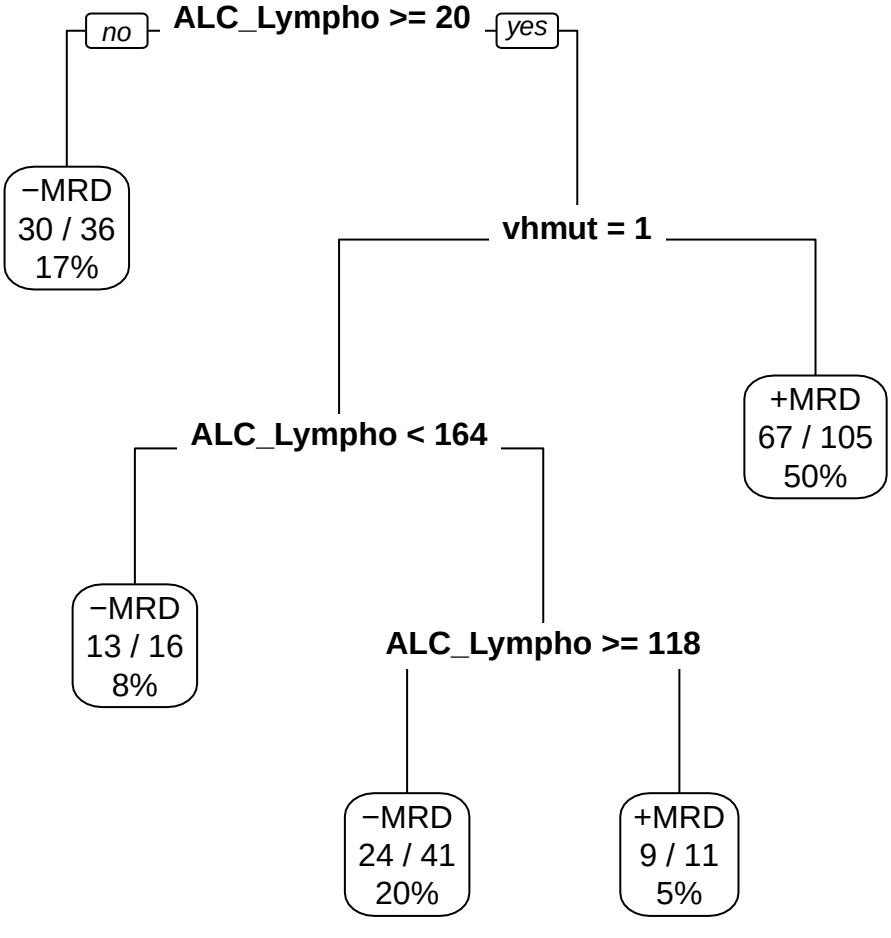
3c) tree22:All clinical data, using only WBC
cp=0.01, endgroup=8, MissClassErr=23.9%



3c) tree23:All clinical data, using only Lymphos
cp=0.01, endgroup=8, MissClassErr=24.4%

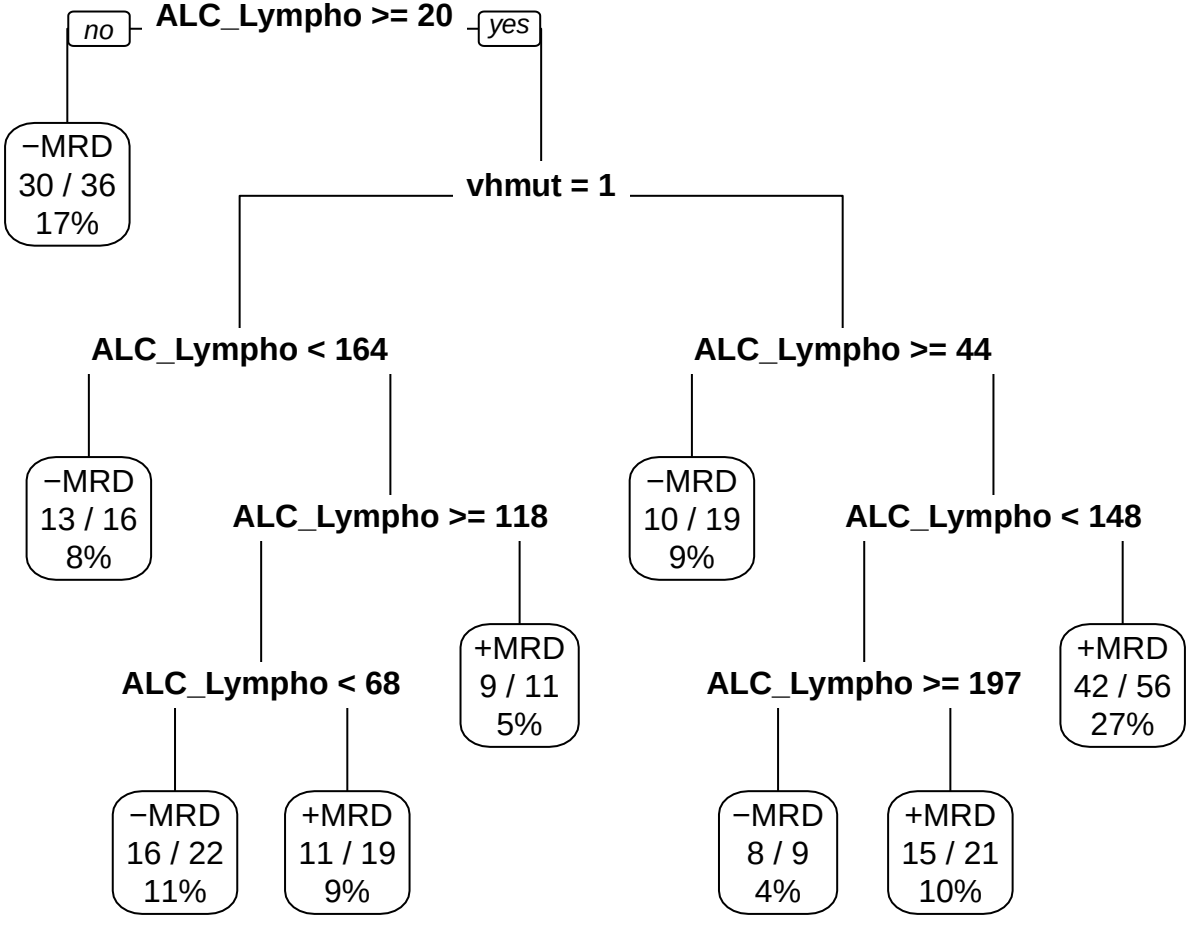


3c) tree24:All clinical data, using only Lymphos
cp=0.03, endgroup=8, MissClassErr=32%



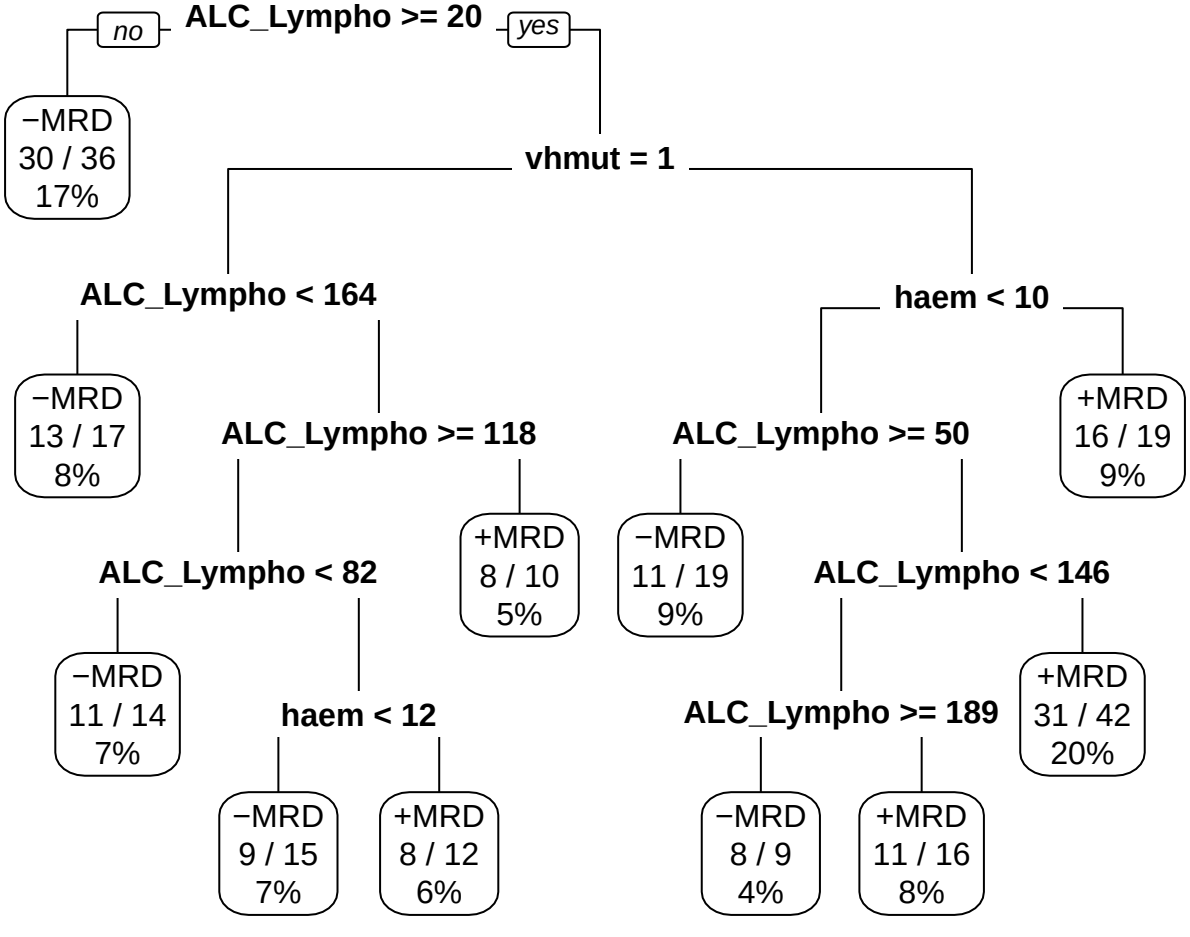
Trying to make the tree less complex results in increase in missclassification.
Note that platelets and haem are not important anymore.

3c) tree25:All clinical data, using only Lymphos and only plat
cp=0.01, endgroup=8, MissClassErr=26.3%



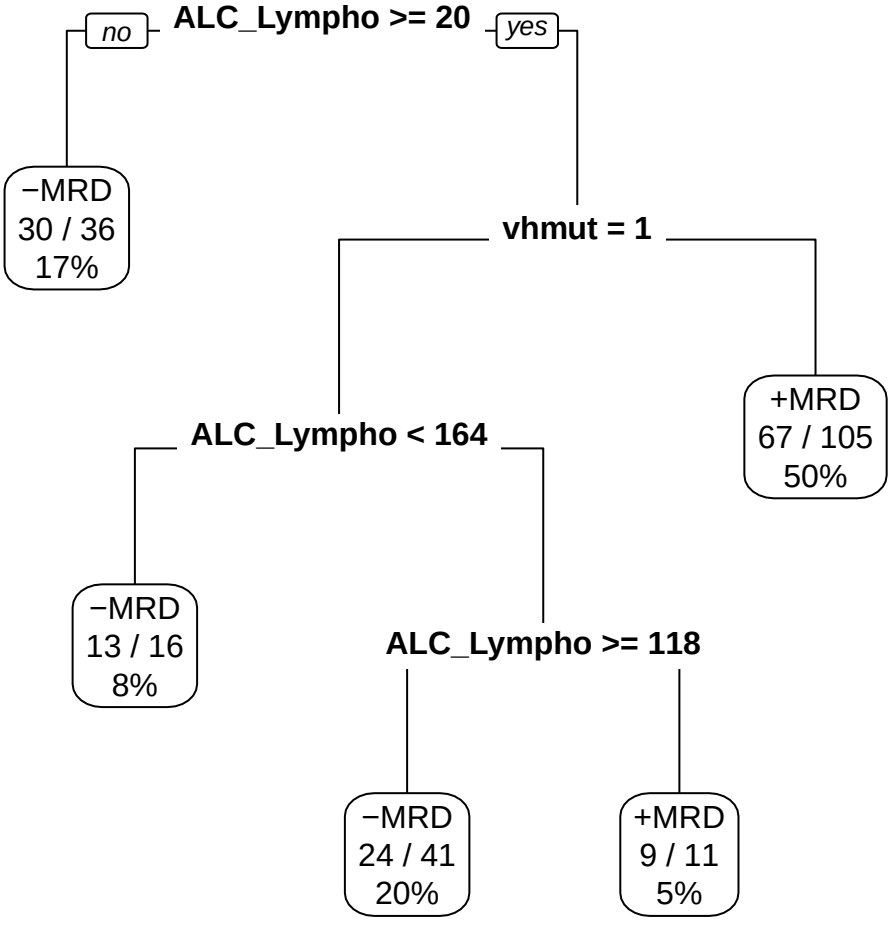
Note that platelets don't seem to be important anymore once you take out haem.

3c) tree26:All clinical data, using only Lymphos and only haem
cp=0.01, endgroup=8, MissClassErr=25.4%



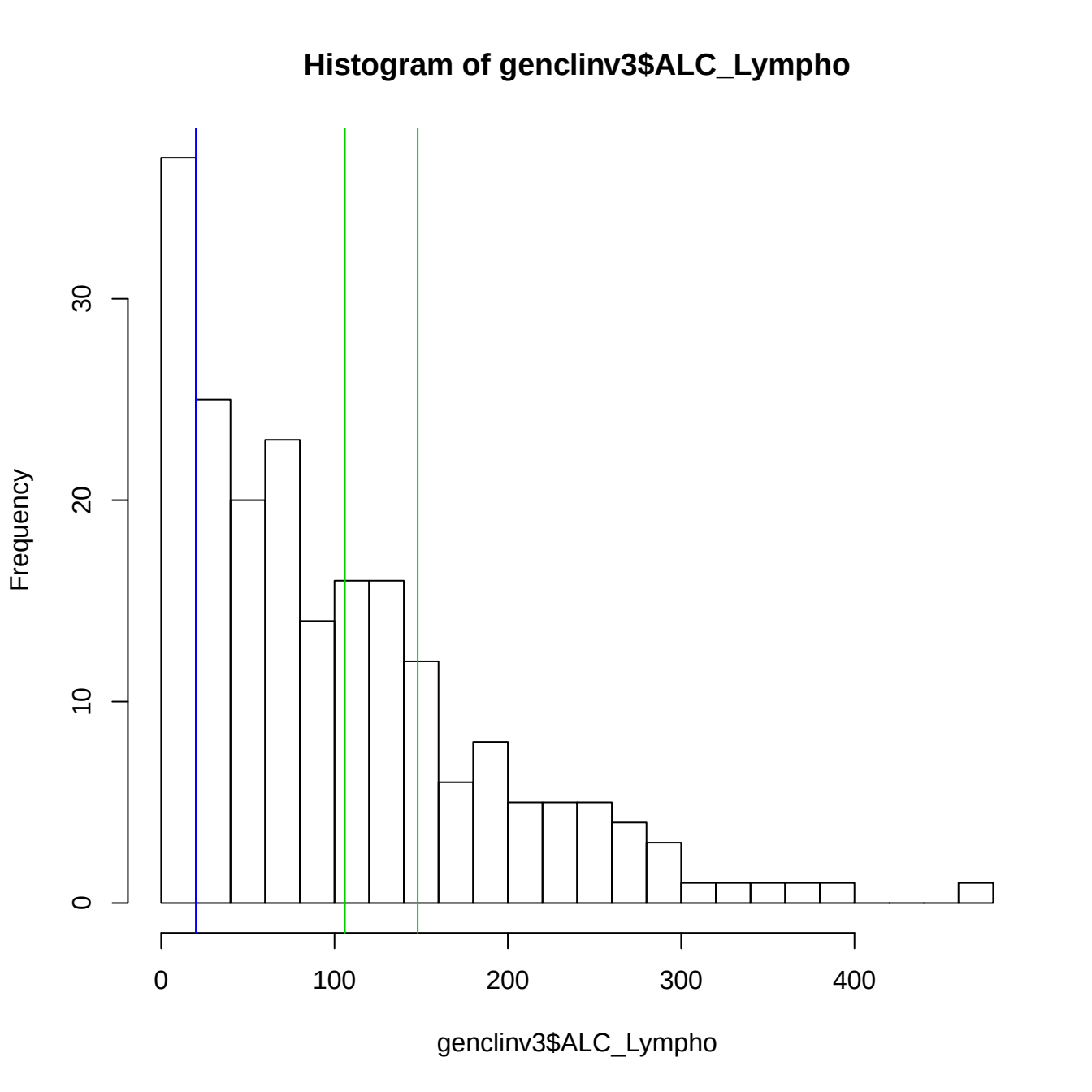
Note that haem doesn't seem to be important anymore once you take out platelets.

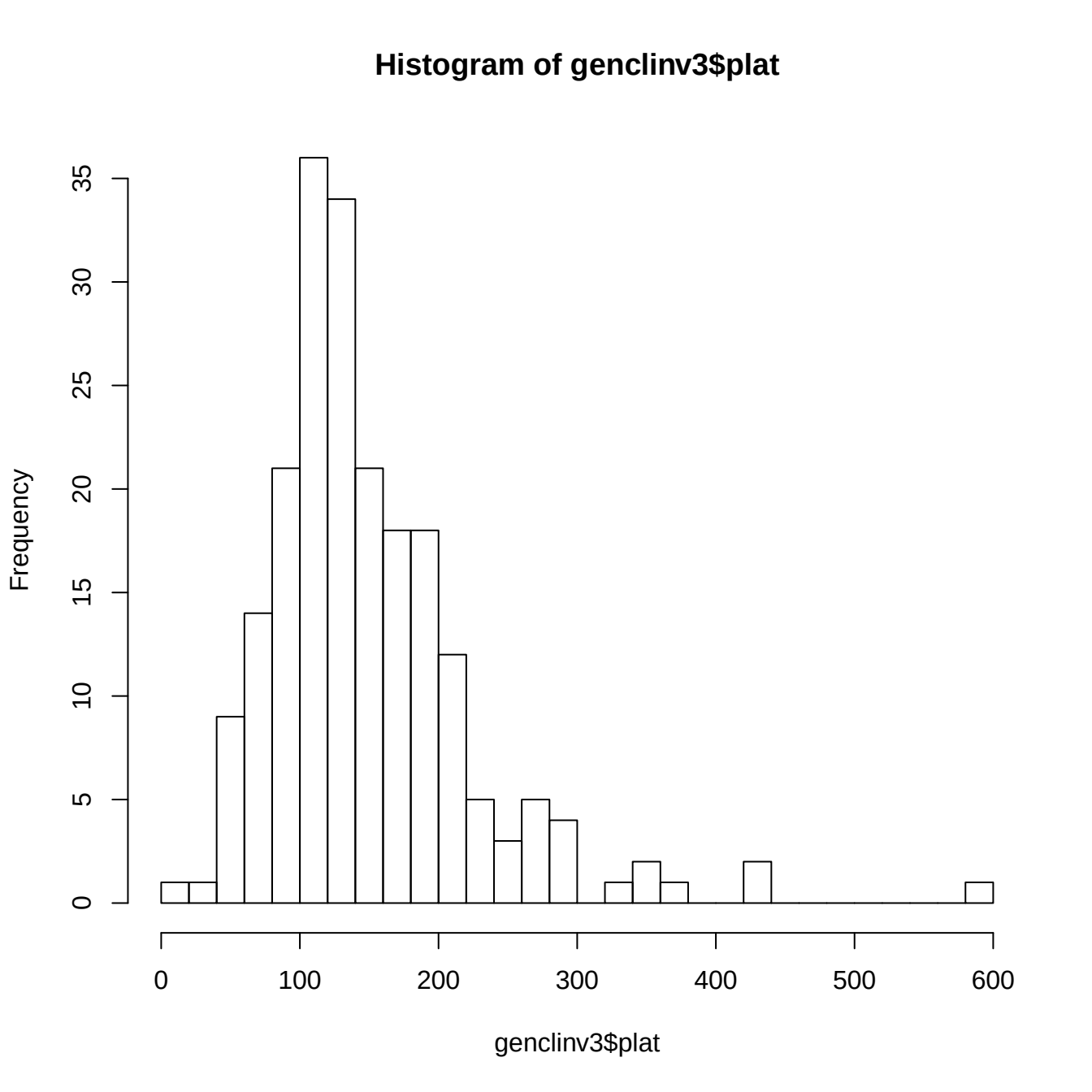
3c) tree27:All clinical data, using only Lymphos and only plat
cp=0.03, endgroup=8, MissClassErr=32%

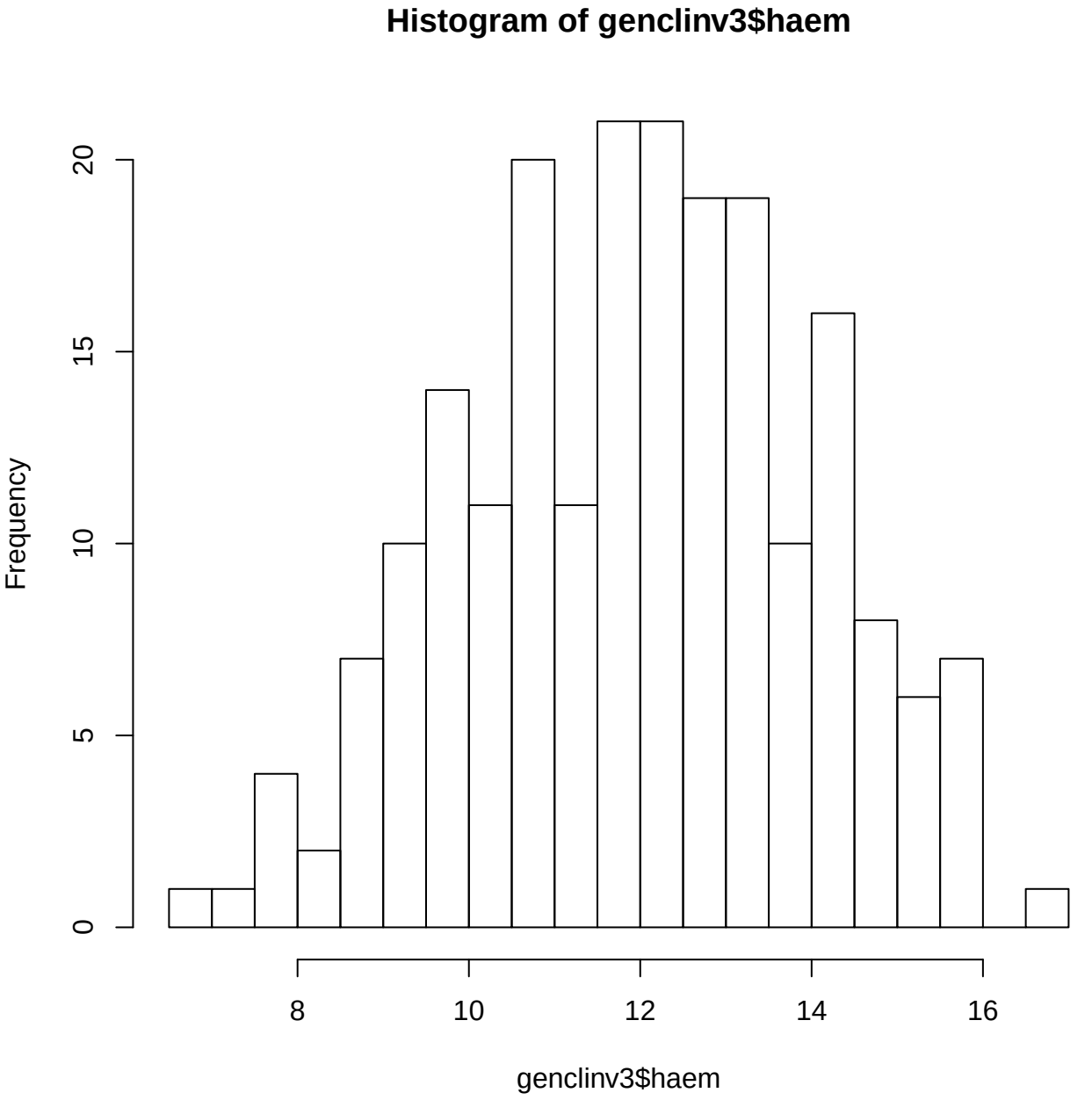


Trying to make the tree less complex results in increase in missclassification.
Note that platelets don't seem to be important anymore once you take out haem.

Question1:
Is there a way to group full blood count data?
Are there any common standards that are used for deciding if a blood count is abnormal?
I attach an overview of how the data is distributed.


Histogram of genclinv3$ALC_Lympho
Frequency
0
10
20
30
0
100
200
300
400
genclinv3$ALC_Lympho

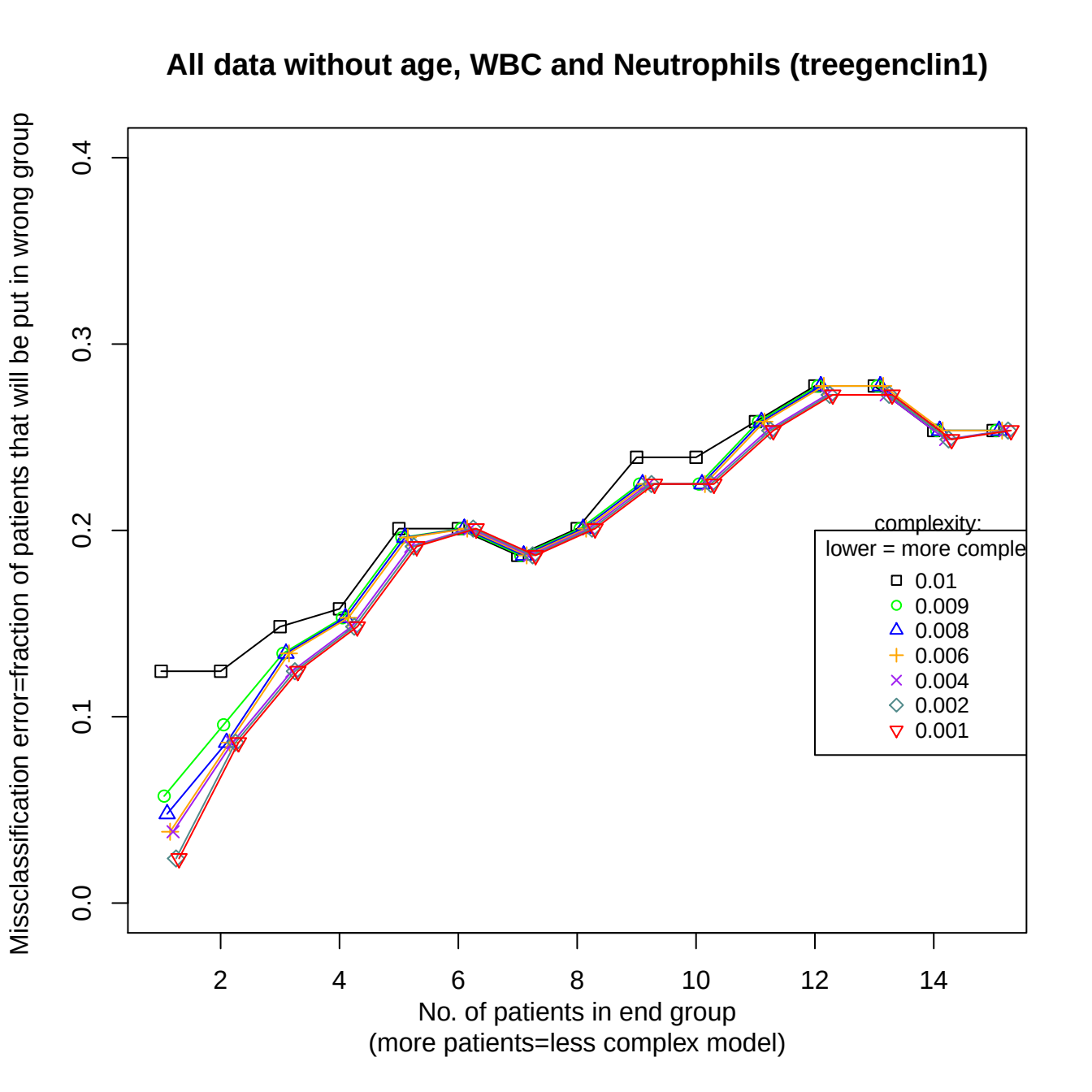

Histogram of genclinv3$plat
Frequency
0
5
10
15
20
25
30
35
0
100
200
300
400
500
600
genclinv3$plat


Histogram of genclinv3$haem
Frequency
0
5
10
15
20
8
10
12
14
16
genclinv3$haem

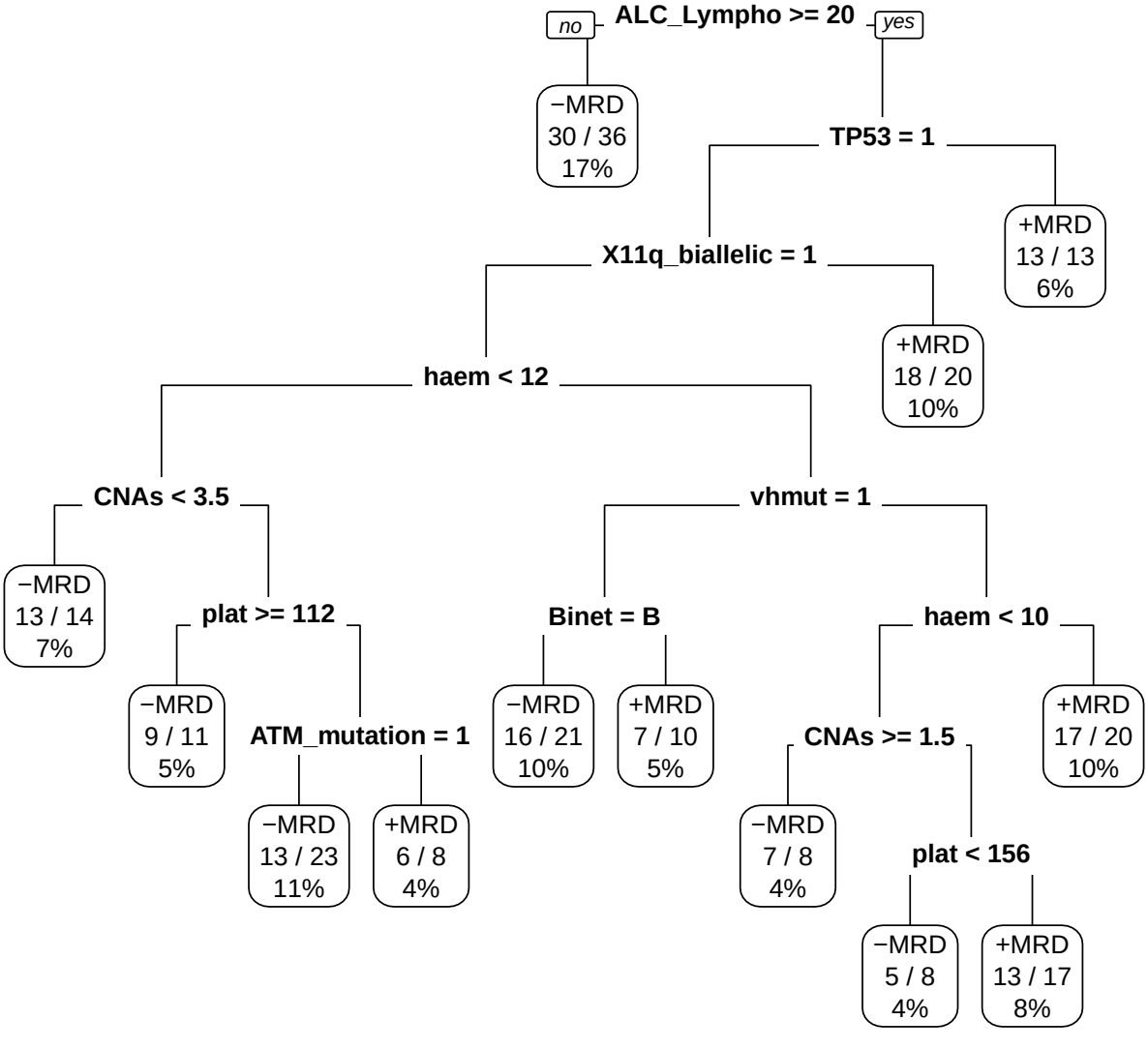
(4) Built a tree with clinical and genetic data (treegenclin1)

vhmut
Binet
gender
MRD
cd38
clones
CNAs
SAMHD1_1mutationONLY
SAMHD1_biallelic_all
TP53_mutationONLY
TP53biallelic
SF3B1_mutation
NOTCH1_mutation
trisomy12
trisomy18
trisomy19
XPO1amplificationALL
XPO1mutationALL
MYD88mutation
MED12mutation
POT1mutation
X8q24amplification
ZFPM2mutation
Del14qi
del8p
X2pgain
del4p
X8qgain
X6qMDR3
X17q11q12
X18q21p23
No_alterations
independent_13q.
X13q_hom
X13q_het
X13_refinedMDR_loss_4
X3_newMDR_gain
X9_refinedMDR_loss
ATM_mutation
BIRC3_mutation
X11q_mono_del
X11q_biallelic
SAMHD1
TP53
haem
plat
ALC_Lympho

Age, Neutrophils and WBC left out


All data without age, WBC and Neutrophils (treegenclin1)
Missclassification error=fraction of patients that will be put in wrong group
0.0
0.1
0.2
0.3
0.4
2
4
6
8
10
12
14
No. of patients in end group
(more patients=less complex model)
complexity:
lower = more comple
0.01
0.009
0.008
0.006
0.004
0.002
0.001

4c) tree 28:All data (treegenclin1)
cp=0.01, endgroup=8, MissClassErr=20%



5) Comparison

How much did we improve forecasting of MRD+/− with genetics?

−> This would at the moment mean that we go from 24.4% to 20% missclassification error.

How well do both models perform seperately?

−> Both models perform quite well, but are highly dependend on which variables we decide to have in.

Are there certain subgroups of patients that benefit particularly from genetic testing?

−> Will go into more detail of this question soon...