HICF1 - Final Report v6

Dr. Susanne Weller

September 30, 2014

Contents

1 Methods and Programmes used

This can go into a paper:
Statistical analysis was carried out using the programme R (version 3.0.1). Survival data was analysed using the additional package ”survival” (version 2.37-7). The code for this analysis is publicly available on github: https://github.com/Suska/HICF1
citation:
R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

2 Univariate Analysis

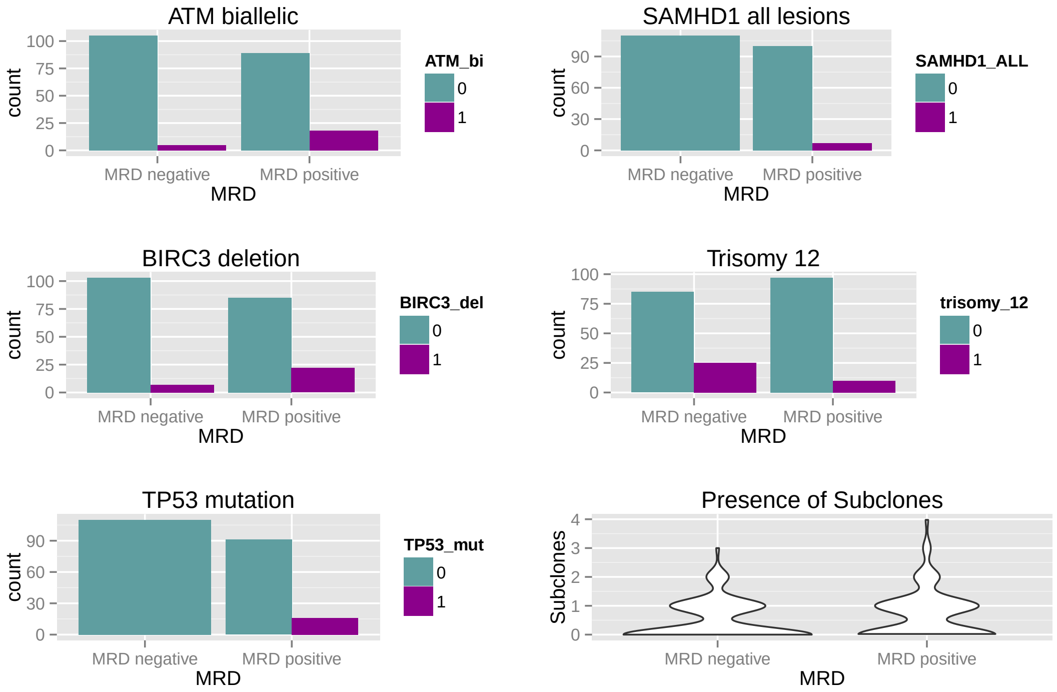
This can go into a paper:
Univariate analysis was done using Fisher’s Exact test for binary genetic variables and wilcoxon signed rank test for continuous variables (Number of CNAs and Subclones). Correction for multiple testing was done using False Discovery Rate.
citation:
Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289–300.

Note that TP53_mut are only mutation with >5%VAF! Univariate p-values change dramatically if you add more variables, this is due to the multiple testing problem.

Table 1: Univariate Analysis against MRD outcome

p	p	sig	corr.p	sig.corr	MRDpos_0	MRDneg_0	MRDpos_1	MRDneg_1
p ATM_bi	0.004	**	0.011	*	41%	48%	8%	2%
p ATM_del	0.001	***	0.005	**	36%	46%	13%	5%
p ATM_mono	0.502	n.s.	0.554	n.s.	45%	45%	4%	6%
p BIRC3_bi	0.365	n.s.	0.465	n.s.	48%	50%	1%	0%
p BIRC3_del	0.002	**	0.007	**	39%	47%	10%	3%
p BIRC3_mono	0.065	trend	0.101	n.s.	49%	47%	0%	3%
pNOTCH1_mut	0.068	trend	0.101	n.s.	45%	42%	4%	9%
pSAMHD1_ALL	0.006	**	0.014	*	46%	51%	3%	0%
p SF3B1_mut	0.514	n.s.	0.554	n.s.	37%	41%	12%	10%
p TP53_bi	0.001	***	0.005	**	45%	51%	4%	0%
p TP53_mut	0	***	0	***	42%	51%	7%	0%
p trisomy_12	0.009	**	0.018	*	45%	39%	5%	12%
p CNAs	0.574	n.s.	0.574	n.s.	NA%	NA%	NA%	NA%
p Subclones	0.072	trend	0.101	n.s.	NA%	NA%	NA%	NA%

p



3 Associations

To test for associations, I first counted the number of patients that have a particular mutation, and derived the probablity of having this lesion:
Example:
8 out of 217 patients have mutation X -> probability estimate for this mutation is 8/217
15 out of 217 patients have mutation Y -> probability estimate for this mutation is 15/217
The expected probablity of having both mutations is then 8/217 x 15/217

I then compared this expected probability to the observed probability using Exact Binomial Tests. This test is the only one that I could find that can deal with low numbers AND allows for testing agains expected frequencies. Fisher's Exact test is often used that way by constructing the expected frequencies from the expected probabilities, but does not allow for integers, which is a problem with the low numbers we are dealing with.
I again used False Discovery Rate to correct the p-values.

This can go into a paper:
We compared expected and observed probabilities using Exact Binomial Tests and corrected for mulitple testing using False Discovery Rates.

variables	TP53_del	TP53_cnLOH	TP53_mut	ATM_mut	ATM_del	ATM_cnLOH	BIRC3_mut	BIRC3_del	NOTCH1_mut	SF3B1_mut	X6q_del	X13q_ALL	trisomy_12	trisomy_18	trisomy_19	XPO1_gain	SAMHD1_ALL	MYD88mut	MED12mut	X8q_ALL	Subclones	CNAs
TP53_del		1.00	0.00	0.18	0.43	1.00	1.00	0.65	0.65	0.29	0.43	1.00	0.42	1.00	1.00	0.01	0.30	1.00	1.00	0.06	0.02	0.00
TP53_cnLOH			0.00	1.00	0.42	1.00	1.00	1.00	0.33	0.48	1.00	0.19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.11	0.30	0.10
TP53_mut				0.34	0.20	1.00	0.65	0.12	1.00	1.00	0.33	0.63	0.09	1.00	1.00	0.01	0.53	1.00	1.00	0.01	0.00	0.00
ATM_mut					0.00	0.01	0.28	0.02	0.26	0.24	0.56	0.48	0.12	0.65	0.41	0.49	1.00	0.59	0.39	0.73	0.25	0.19
ATM_del						1.00	0.01	0.00	0.84	0.41	0.33	0.66	0.06	0.63	0.64	1.00	0.22	1.00	1.00	0.70	0.01	0.00
ATM_cnLOH							1.00	1.00	1.00	0.29	1.00	0.32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.18	0.74	0.21
BIRC3_mut								0.02	0.06	1.00	1.00	0.86	0.04	0.06	0.01	1.00	1.00	1.00	1.00	0.14	0.90	0.07
BIRC3_del									0.34	1.00	0.71	0.80	0.07	1.00	1.00	1.00	0.12	1.00	1.00	1.00	0.05	0.06
NOTCH1_mut										0.13	1.00	0.12	0.02	0.55	0.60	0.38	1.00	1.00	0.55	0.38	0.10	0.84
SF3B1_mut											0.55	0.61	0.00	1.00	1.00	0.29	0.27	1.00	1.00	0.17	0.94	0.17
X6q_del												0.85	1.00	1.00	1.00	0.43	1.00	1.00	1.00	1.00	0.58	0.00
X13q_ALL													0.00	0.77	1.00	0.82	0.81	0.46	0.13	0.36	0.00	0.00
trisomy_12														0.02	0.00	0.42	0.41	1.00	0.07	0.68	0.00	0.51
trisomy_18															0.00	1.00	1.00	0.09	1.00	1.00	0.51	0.01
trisomy_19																1.00	1.00	1.00	1.00	1.00	0.79	0.05
XPO1_gain																	0.30	1.00	1.00	0.33	0.04	0.00
SAMHD1_ALL																		1.00	1.00	1.00	0.68	0.18
MYD88mut																			1.00	1.00	0.65	0.90
MED12mut																				1.00	0.51	0.48
X8q_ALL																					0.35	0.01
Subclones																						0.00
CNAs																						

Table 2: Association chart, uncorrected pvalues, Fisher's test

variables	TP53_del	TP53_cnLOH	TP53_mut	ATM_mut	ATM_del	ATM_cnLOH	BIRC3_mut	BIRC3_del	NOTCH1_mut	SF3B1_mut	X6q_del	X13q_ALL	trisomy_12	trisomy_18	trisomy_19	XPO1_gain	SAMHD1_ALL	MYD88mut	MED12mut	X8q_ALL	Subclones	CNAs
TP53_del		1.00	0.00	0.69	1.00	1.00	1.00	1.00	1.00	0.91	1.00	1.00	1.00	1.00	1.00	0.09	0.91	1.00	1.00	0.35	0.13	0.00
TP53_cnLOH			0.04	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.53	0.91	0.49
TP53_mut				0.94	0.70	1.00	1.00	0.54	1.00	1.00	0.93	1.00	0.45	1.00	1.00	0.09	1.00	1.00	1.00	0.09	0.05	0.00
ATM_mut					0.05	0.07	0.90	0.17	0.88	0.82	1.00	1.00	0.54	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.84	0.70
ATM_del						1.00	0.09	0.00	1.00	1.00	0.93	1.00	0.34	1.00	1.00	1.00	0.78	1.00	1.00	1.00	0.09	0.01
ATM_cnLOH							1.00	1.00	1.00	0.91	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.69	1.00	0.75
BIRC3_mut								0.16	0.35	1.00	1.00	1.00	0.30	0.35	0.09	1.00	1.00	1.00	1.00	0.56	1.00	0.39
BIRC3_del									0.94	1.00	1.00	1.00	0.39	1.00	1.00	1.00	0.54	1.00	1.00	1.00	0.31	0.35
NOTCH1_mut										0.55	1.00	0.54	0.17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.49	1.00
SF3B1_mut											1.00	1.00	0.04	1.00	1.00	0.91	0.90	1.00	1.00	0.69	1.00	0.69
X6q_del												1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
X13q_ALL													0.00	1.00	1.00	1.00	1.00	1.00	0.55	0.96	0.00	0.00
trisomy_12														0.13	0.02	1.00	1.00	1.00	0.39	1.00	0.06	1.00
trisomy_18															0.00	1.00	1.00	0.46	1.00	1.00	1.00	0.09
trisomy_19																1.00	1.00	1.00	1.00	1.00	1.00	0.30
XPO1_gain																	0.91	1.00	1.00	0.93	0.25	0.01
SAMHD1_ALL																		1.00	1.00	1.00	1.00	0.69
MYD88mut																			1.00	1.00	1.00	1.00
MED12mut																				1.00	1.00	1.00
X8q_ALL																					0.95	0.09
Subclones																						0.00
CNAs																						

Table 3: Association chart, corrected pvalues, Fisher's test with FDR correction

Odds ratios and p-values for associations between genes are represented in this heatmap. Note that odds ratios 0-1 (the first bar in the colour key) are mutually exclusive, everything else already counts as co-occuring.

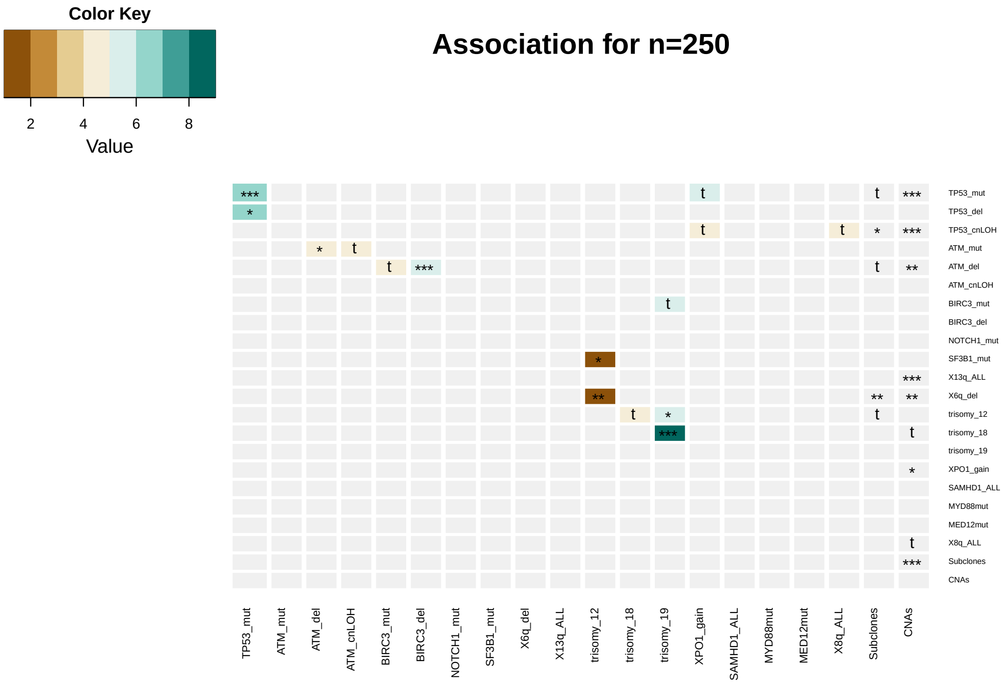
Note: Colour key still needs be adjusted to a somewhat funny scale to see this properly.

This can go into a paper:

Odds ratios and significant values for associations between genes are represented in this graph. Odds ratios between 0 and 1 indicate mutually exclusive genes, while odds ratios above 1 indicate increasing cooccurence. P-values are defined as follows: ***: p<0.001, **: p<0.01, *: p<0.05, t:trend, p<0.15

variables	TP53_del	TP53_cnLOH	TP53_mut	ATM_mut	ATM_del	ATM_cnLOH	BIRC3_mut	BIRC3_del	NOTCH1_mut	SF3B1_mut	X6q_del	X13q_ALL	trisomy_12	trisomy_18	trisomy_19	XPO1_gain	SAMHD1_ALL	MYD88mut	MED12mut	X8q_ALL	Subclones	CNAs
TP53_del		0.00	12.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.79	0.80	0.00	0.00	0.00	7.58	2.78	0.00	0.00	5.03		
TP53_cnLOH			12.04	0.00	1.86	0.00	0.00	0.00	2.53	1.55	0.00	2.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.36		
TP53_mut				0.42	0.26	0.00	0.00	0.00	0.72	0.88	1.71	1.16	0.00	0.00	0.00	4.82	1.32	0.00	0.00	4.82		
ATM_mut					2.07	4.53	0.28	1.95	0.53	1.34	1.28	1.13	0.44	0.00	0.00	1.34	0.99	1.12	1.49	0.44		
ATM_del						0.00	2.84	6.25	1.01	1.25	1.60	1.08	0.27	0.00	0.00	0.55	1.86	0.00	0.93	1.11		
ATM_cnLOH							0.00	0.00	0.00	1.86	0.00	1.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.02		
BIRC3_mut								2.89	2.40	0.87	0.00	1.01	2.38	5.24	6.77	0.00	0.00	0.00	0.00	3.14		
BIRC3_del									0.45	0.98	1.08	1.04	0.19	0.00	0.00	0.76	2.54	0.00	0.00	0.76		
NOTCH1_mut										0.41	0.54	0.59	2.13	1.26	1.08	1.52	0.84	0.00	1.26	1.52		
SF3B1_mut											1.33	1.09	0.11	0.77	0.66	0.00	0.00	0.00	0.77	1.87		
X6q_del												1.01	0.89	0.00	0.00	1.79	0.00	0.00	0.00	0.00		
X13q_ALL													0.24	1.01	0.86	1.01	1.12	1.52	0.00	1.42		
trisomy_12														4.22	5.46	0.00	0.00	0.00	3.15	1.25		
trisomy_18															30.35	0.00	0.00	10.45	0.00	0.00		
trisomy_19																0.00	0.00	0.00	0.00	0.00		
XPO1_gain																	2.78	0.00	0.00	2.51		
SAMHD1_ALL																		0.00	0.00	0.00		
MYD88mut																			0.00	0.00		
MED12mut																				0.00		
X8q_ALL																						
Subclones																						
CNAs																						

Table 4: Odds ratios for association between genes



3.1 Multiple logistic regression models

This can go into a paper:
As significantly more MRD positive patients have progressed during the trial (Chi Square test, ChiSquare=10.26, n=104, p=0.001), we use MRD status as proxy for progression free survival.
Mulitvariate analysis was done using multiple logistic regression models. We selected only variables that were significant in the univariate analysis to go into the multiple logistic regression. One specific goal was to see if ATM biallelic is a better predictor for MRD positivity than ATM deletions.

Table 5: Multiple log regression, n=217

	Dependent variable:						
				MRD			
	genetic1	genetic2	genetic3	genetic4	genetic5	genetic6	genetic7
TP53_ALL1	2.65*** (0.77)	2.59*** (0.76)	2.56*** (0.76)	2.56*** (0.76)	2.46*** (0.76)	2.46*** (0.76)	
ATM_del1	1.40*** (0.41)						
ATM_bi1		1.51*** (0.54)	1.55*** (0.55)	1.59*** (0.55)	1.55*** (0.55)		
trisomy_121	−0.66 (0.42)	−0.66 (0.42)	−0.45 (0.43)	−0.52 (0.43)	−0.61 (0.43)		
BIRC3_mono1			−0.96 (1.33)	−1.70 (1.15)	−1.74 (1.15)		
SAMHD1_ALL1	16.64 (854.98)	16.71 (873.55)	17.69 (1,438.54)	16.69 (872.11)			
trisomy_121:BIRC3_mono1			−15.83 (1,769.26)				
vh_mutation_statusunmutated							0.16 (0.50)
Constant	−0.44** (0.18)	−0.34* (0.17)	−0.32* (0.18)	−0.32* (0.18)	−0.22 (0.17)	−0.20 (0.14)	−0.22 (0.47)
Observations	217	217	217	217	217	217	196
Log Likelihood	−127.05	−128.86	−126.86	−127.39	−132.64	−141.44	−135.64
Akaike Inf. Crit.	264.11	267.72	267.71	266.79	275.29	286.88	275.29
Note:	*p<0.1; **p<0.05; ***p<0.01						

3.2 Missclassification Error

Table 6: Missclassification for summarized models

p	model	correct_MRD_neg	false_MRD_neg	correct_MRD_pos	false_MRD_pos	missclasserr	unclassified
p1	fit.sum.gen1	98	56	51	12	0.313	0
p2	fit.sum.gen2	103	65	42	7	0.332	0
p3	fit.sum.gen3	103	65	42	7	0.332	0
p4	fit.sum.gen4	103	66	41	7	0.336	0
p5	fit.sum.gen5	103	72	35	7	0.364	0
p6	fit.sum.gen6	108	88	19	2	0.415	0
p7	fit.vhmut	102	94				0.097

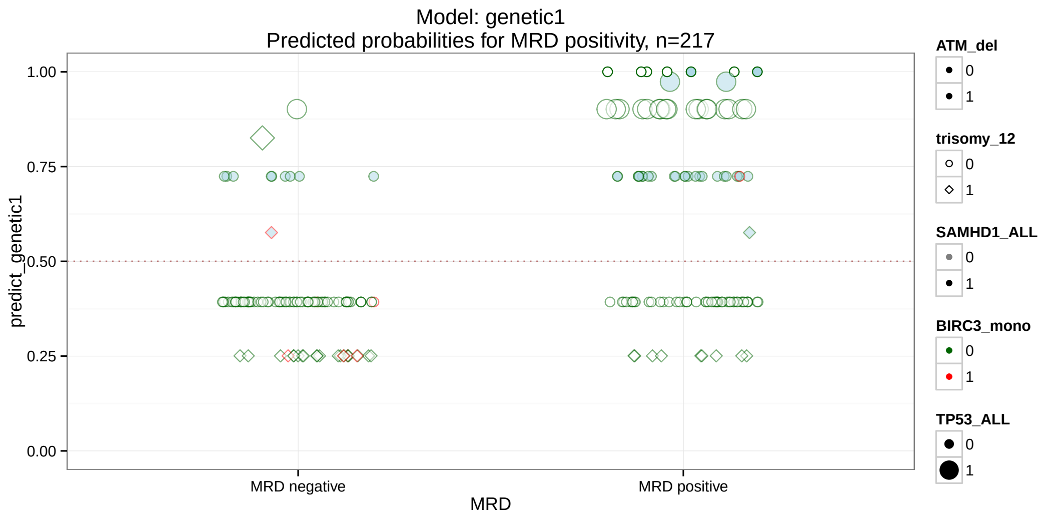
p

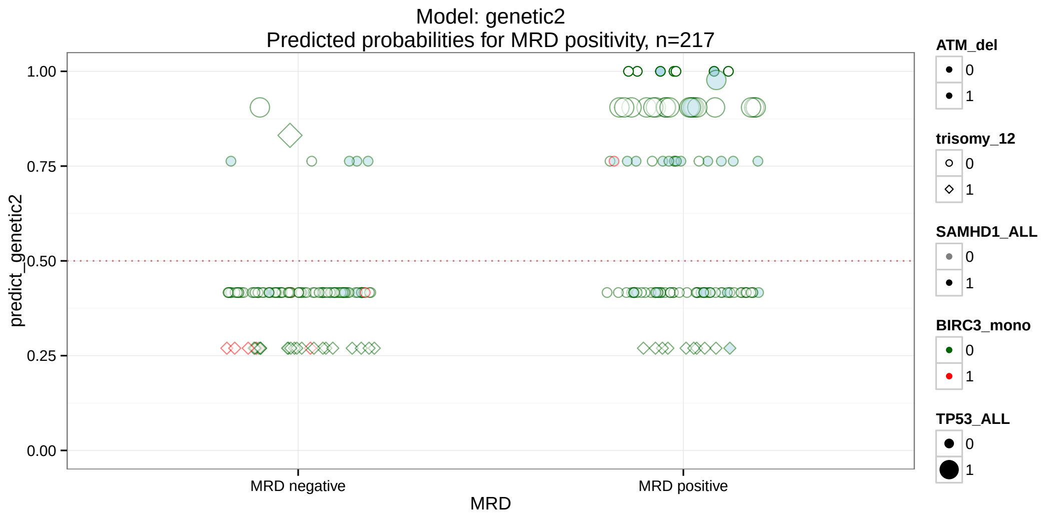
3.2.1 Model probabilities

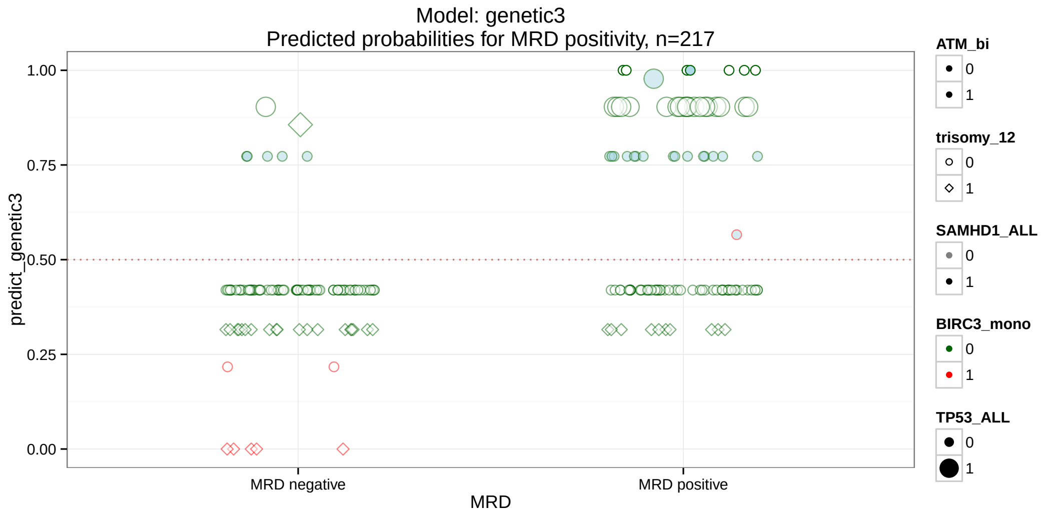
The following graphs show the predicted probability for MRD positivity of the different models, with the x-axis showing the real MRD status. Note again that the final model only contains 181 data points.
The graph depicts the following variables (note:Not all of them are necessarily in the model depicted):

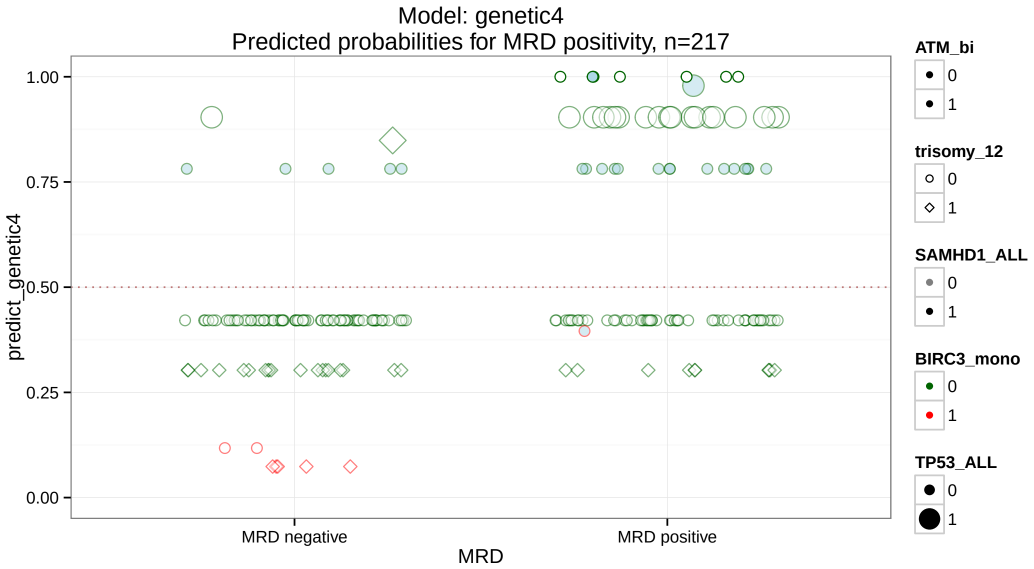
- Trisomy12 is depicted by the shape of the points (cirle=0, square=1).
- SAMHD1 is depicted by translucent points (translucent=mutated)
- ATM biallelic is depicted by light blue filling.
- BIRC3 is depicted by green(0) and red(1) point outline.
- TP53 is depicted by point size (large=1)

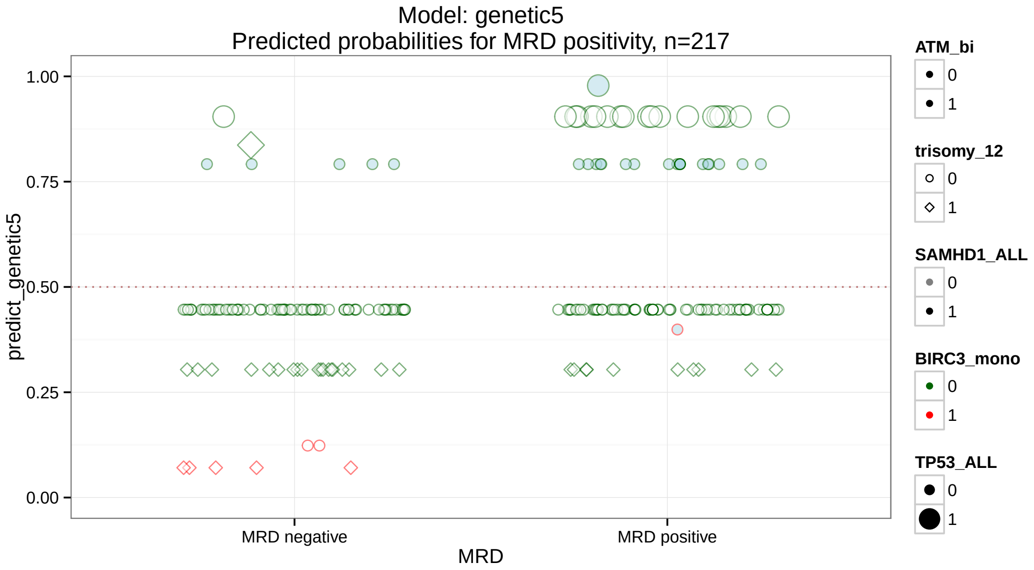
The dashed red line shows the 0.5 line. Everything above is classified by the model as MRD positive, below is classified as MRD negative.

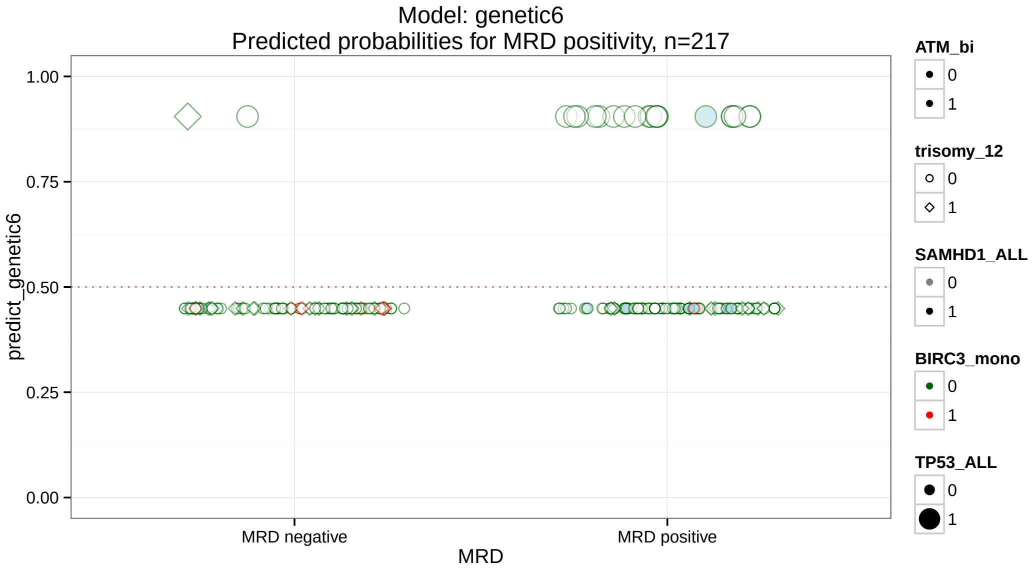

Model: genetic1
Predicted probabilities for MRD positivity, n=217
predict_genetic1
1.00
0.75
0.50
0.25
0.00
MRD negative
MRD positive
MRD
ATM_del
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1

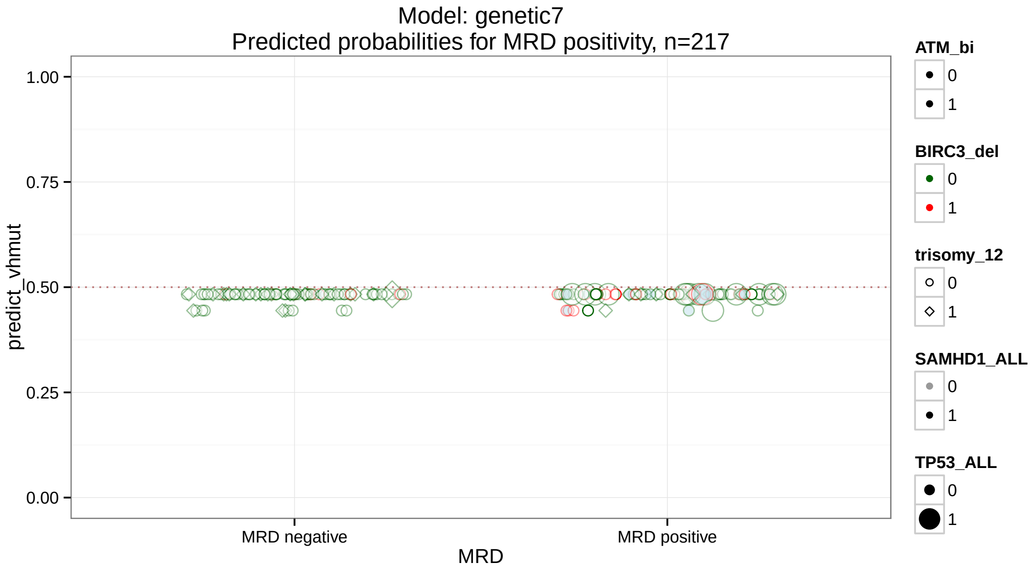

Model: genetic2
Predicted probabilities for MRD positivity, n=217
predict_genetic2
1.00
0.75
0.50
0.25
0.00
MRD negative
MRD positive
MRD
ATM_del
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1


Model: genetic3
Predicted probabilities for MRD positivity, n=217
predict_genetic3
1.00
0.75
0.50
0.25
0.00
MRD negative
MRD positive
MRD
ATM_bi
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1


Model: genetic4
Predicted probabilities for MRD positivity, n=217
predict_genetic4
1.00
0.75
0.50
0.25
0.00
MRD negative
MRD positive
MRD
ATM_bi
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1


Model: genetic5
Predicted probabilities for MRD positivity, n=217
predict_genetic5
1.00
0.75
0.50
0.25
0.00
MRD negative
MRD positive
MRD
ATM_bi
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1


Model: genetic6
Predicted probabilities for MRD positivity, n=217
predict_genetic6
1.00
0.75
0.50
0.25
0.00
MRD negative
MRD positive
MRD
ATM_bi
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1



3.3 Model Accuracy

To estimate model accuracy, we selected all patients that were correctly classified and looked at their model probablities. You can see nicely that model 1, despite being the model with the best missclassification errors, is not as accurate as model 2 and 4 (both using ATM bi) for MRD positivity.

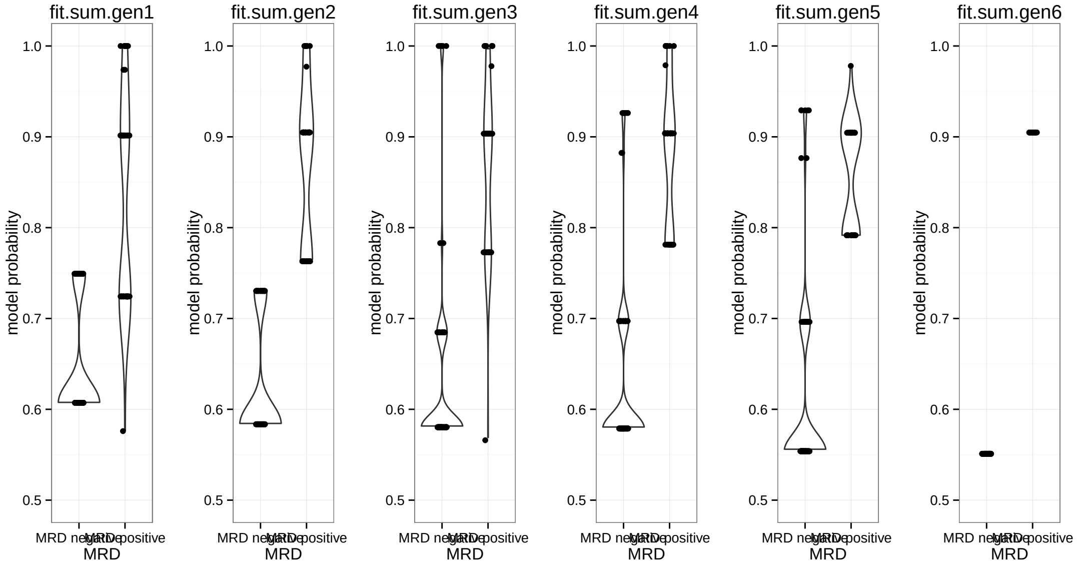


Table 7: Summary model probabilities

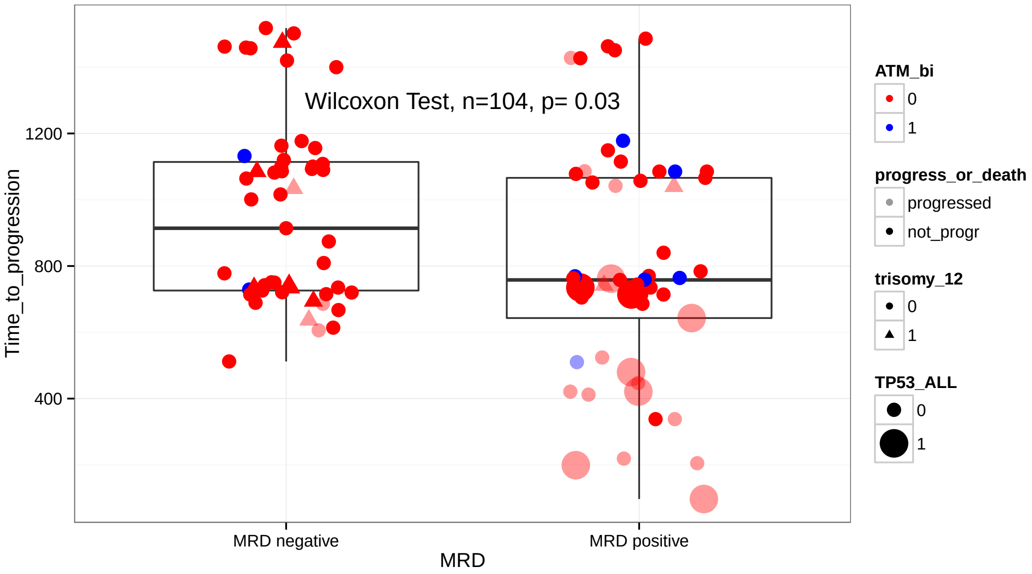
p	model	+mean	+median	+min	+max	-mean	-median	-min	-max
p1	fit.sum.gen1	83%	90%	58%	100%	64%	61%	61%	75%
p2	fit.sum.gen2	87%	90%	76%	100%	62%	58%	58%	73%
p3	fit.sum.gen3	87%	90%	57%	100%	62%	58%	58%	100%
p4	fit.sum.gen4	88%	90%	78%	100%	62%	58%	58%	93%
p5	fit.sum.gen5	85%	90%	79%	98%	60%	55%	55%	93%
p6	fit.sum.gen6	90%	90%	90%	90%	55%	55%	55%	55%

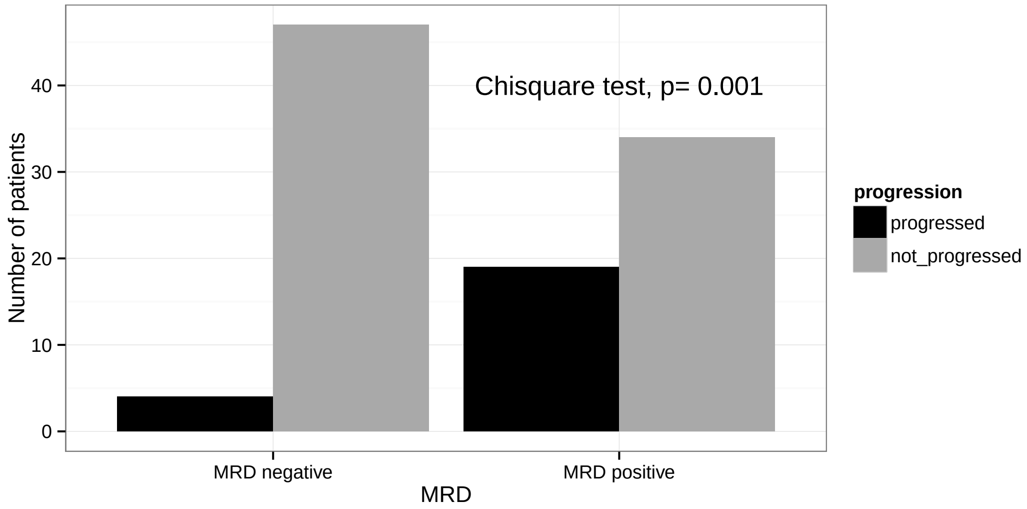
p

4 Progression Free Survival

4.1 MRD as Proxy for PFS

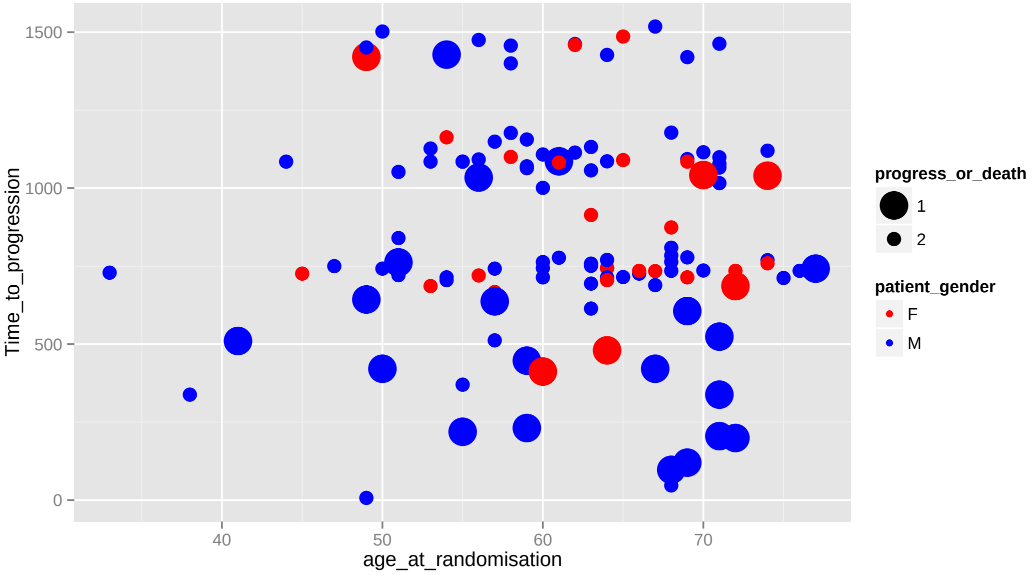
We first assess if MRD is a good proxy for survival via simple univariate testing:





We can conclude that MRD is a good proxy for PFS.

Next, we want to check if progression is biased towards a certain gender or age:



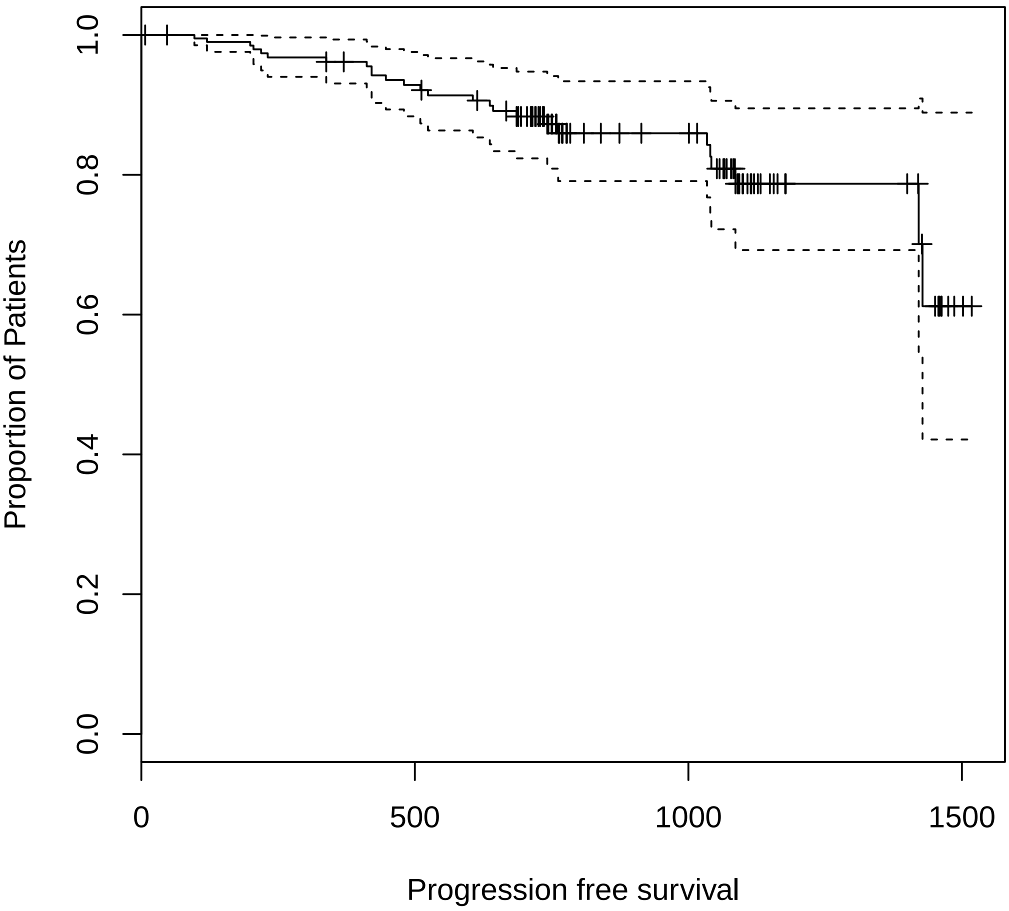
Fortunately, this is not the case, although we have double thenumber of males compared to females, both age and gender does not confound with pregression and time to progression.

4.2 Cox Hazard Regression Model

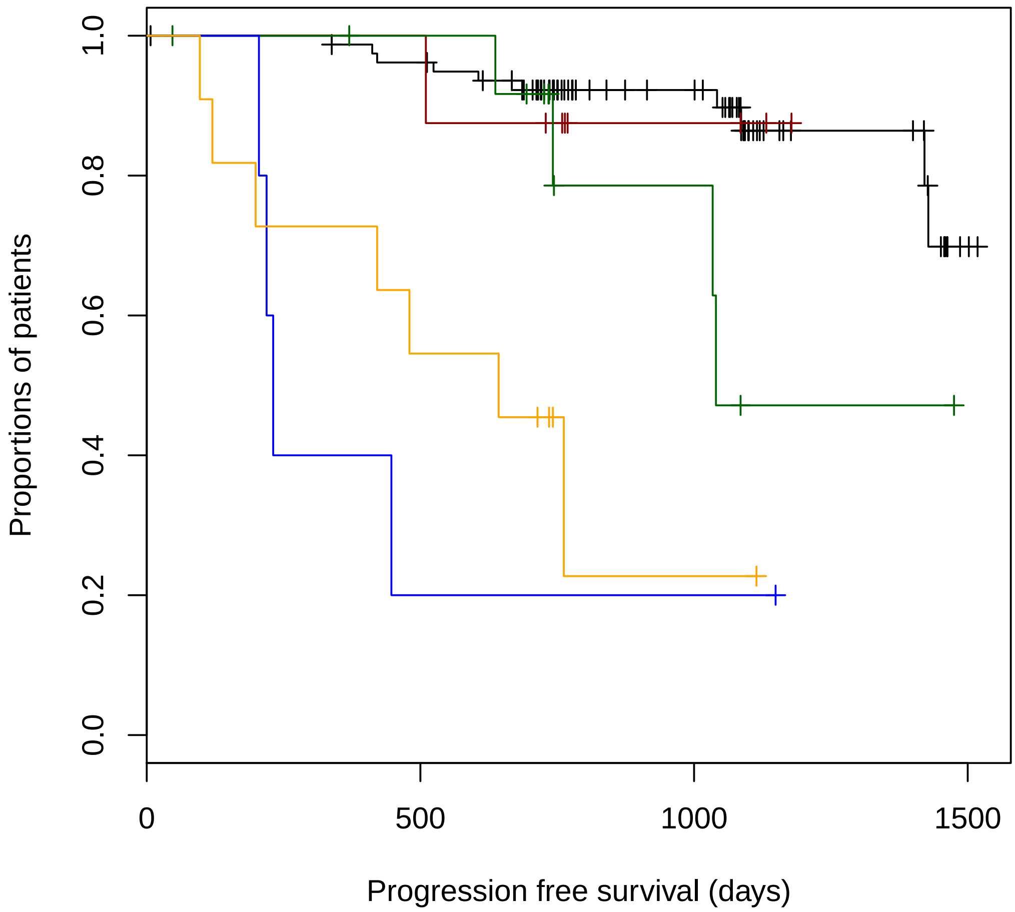
First, we plot all our data to see how it looks like in a Kaplan-Meier Curve:

Table 8: Survival model

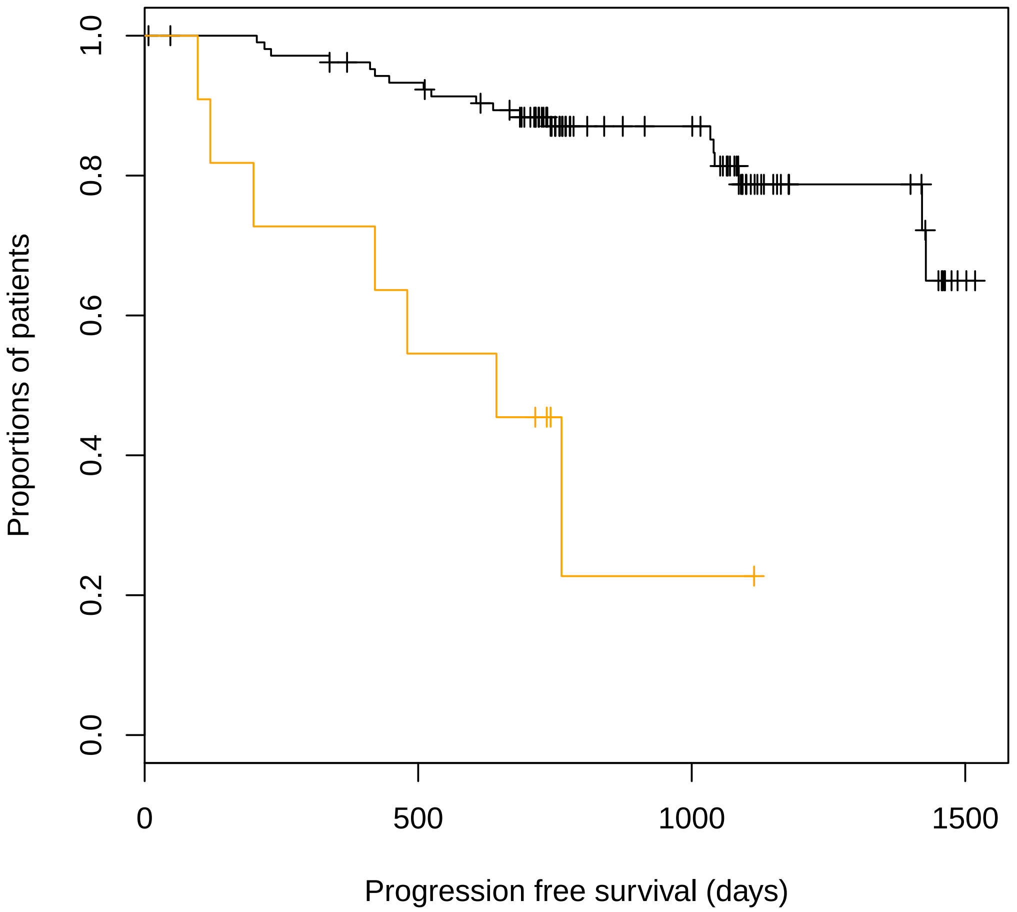
	Dependent variable:
	Time_to_progression
TP53_ALL1	2.38*** (0.51)
ATM_bi1	0.19 (1.06)
SAMHD1_ALL1	2.82*** (0.61)
trisomy_121	1.08* (0.59)
Observations	118
R^2	0.21
Max. Possible R^2	0.85
Log Likelihood	−98.03
Wald Test	32.56*** (df = 4)
LR Test	27.90*** (df = 4)
Score (Logrank) Test	50.83*** (df = 4)
Note:	*p<0.1; **p<0.05; ***p<0.01



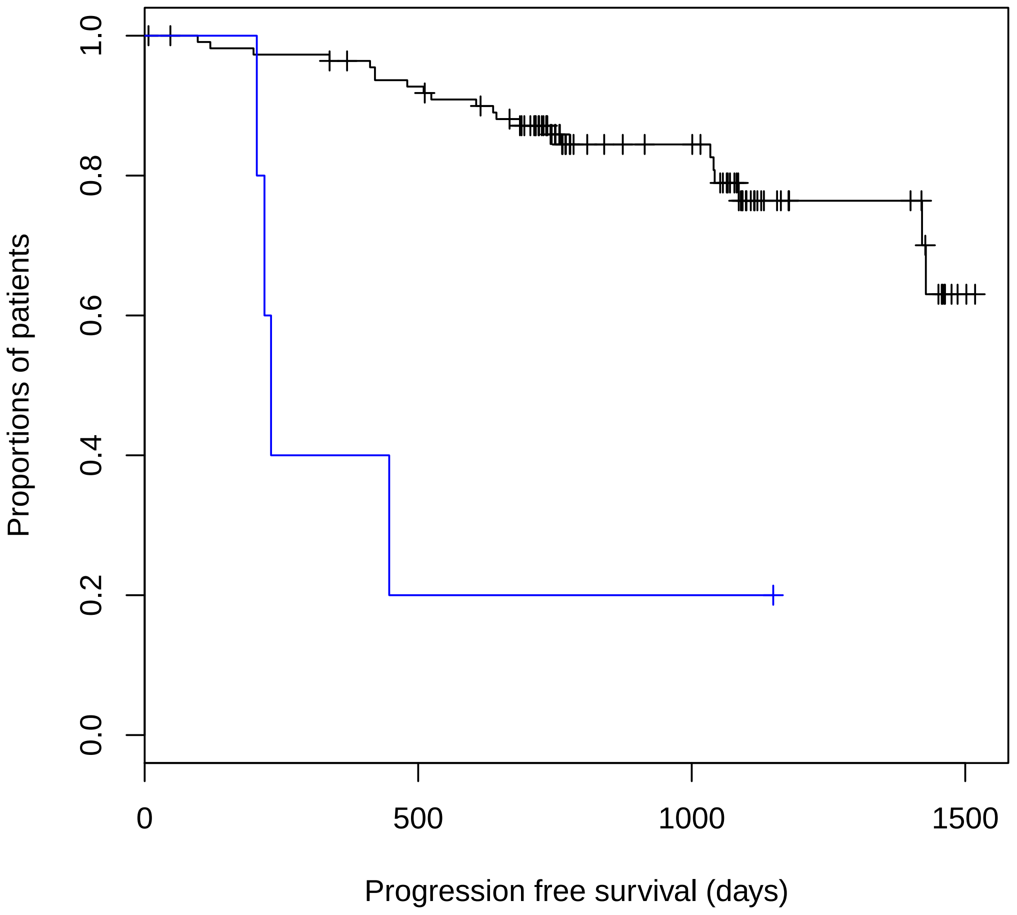
We fitted a Cox Proportional Hazard Model using the survival package (R), with TP53, ATM biallelic, SAMHD1 and trisomy 12 as predictors.



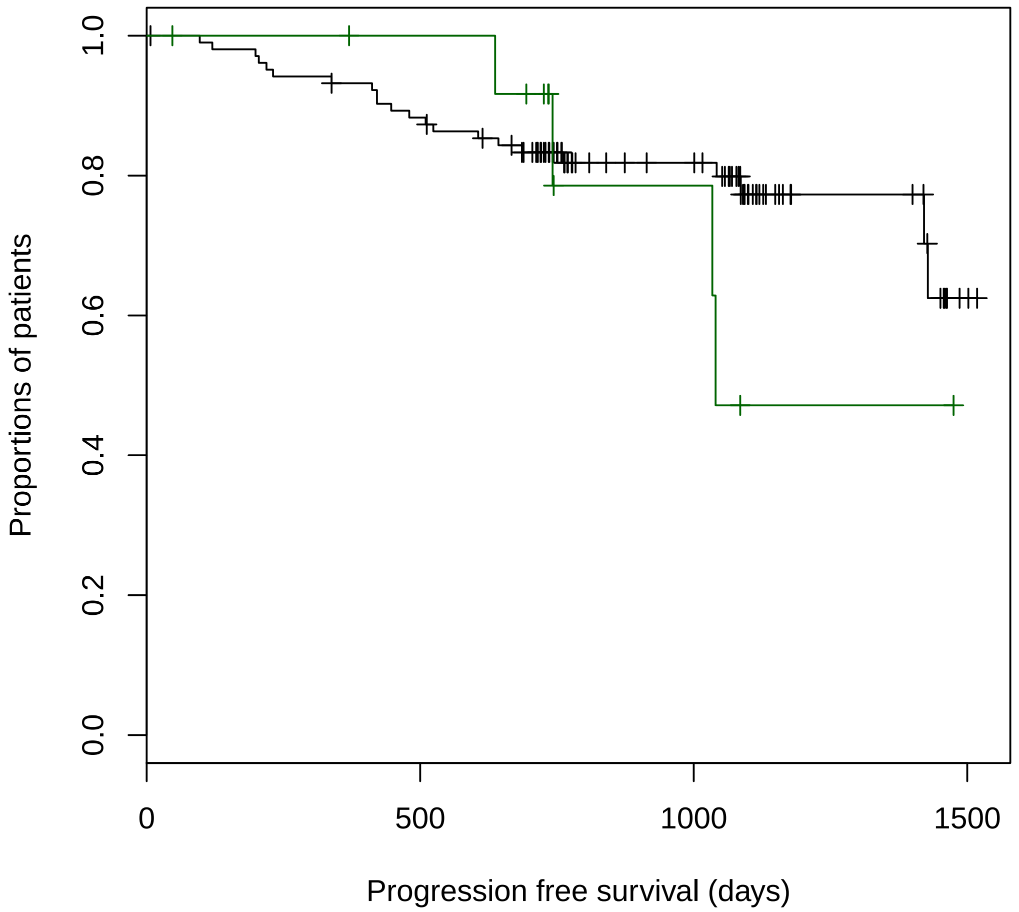
For TP53


Proportions of patients
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
Progression free survival (days)

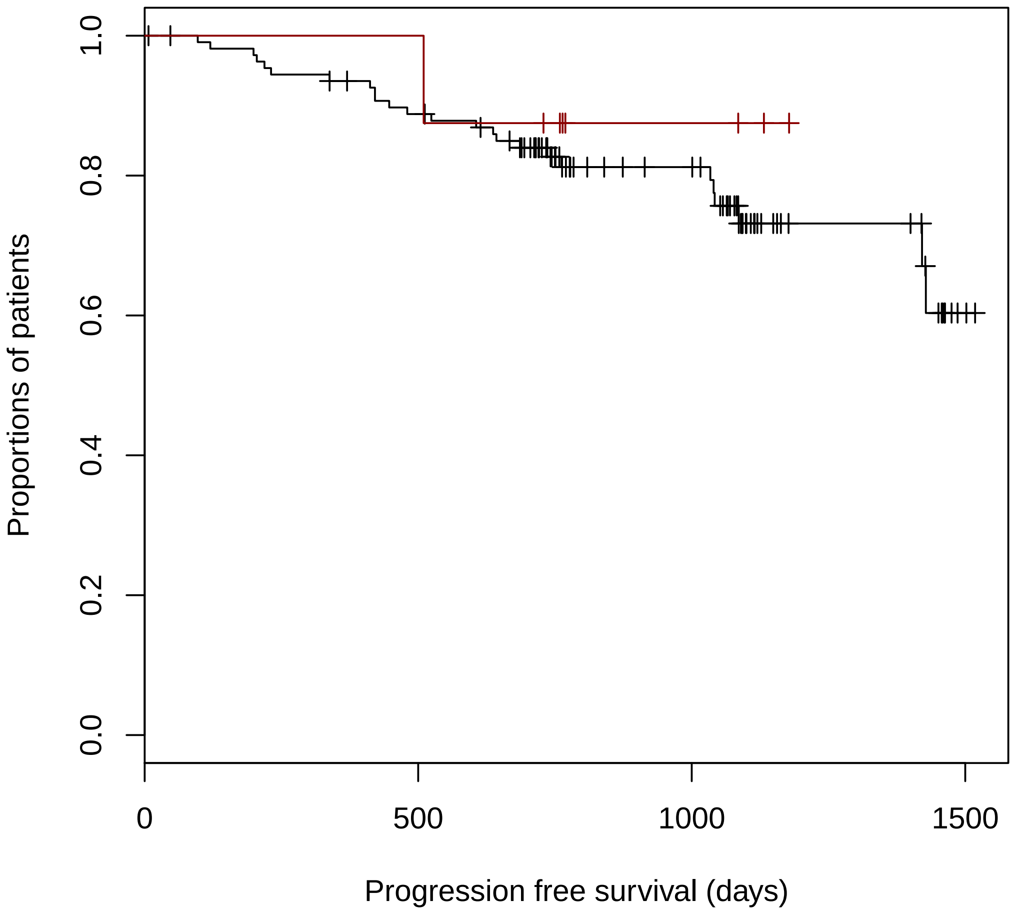
For SAMHD1


Proportions of patients
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
Progression free survival (days)

For Trisomy 12



For ATM bi

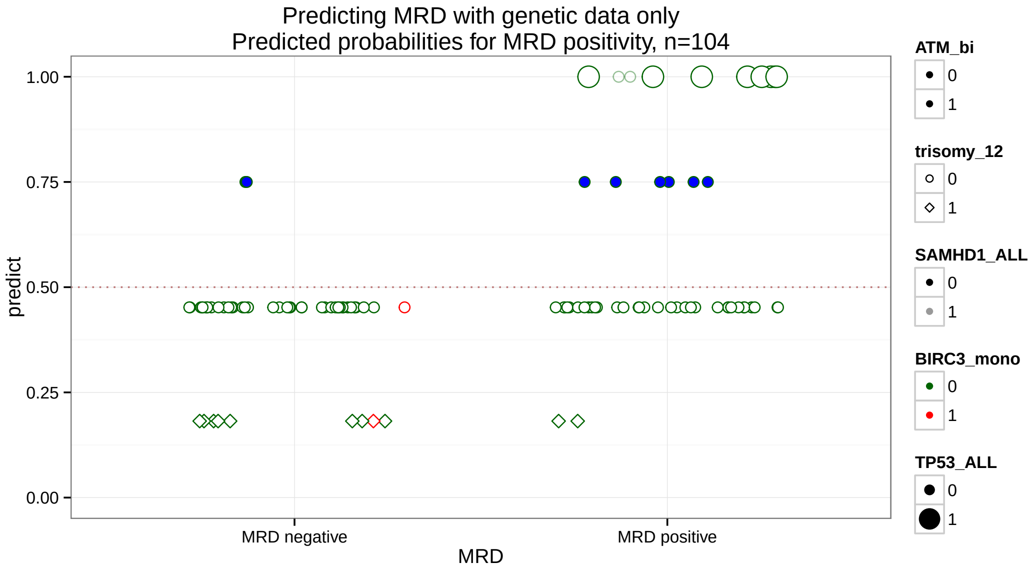


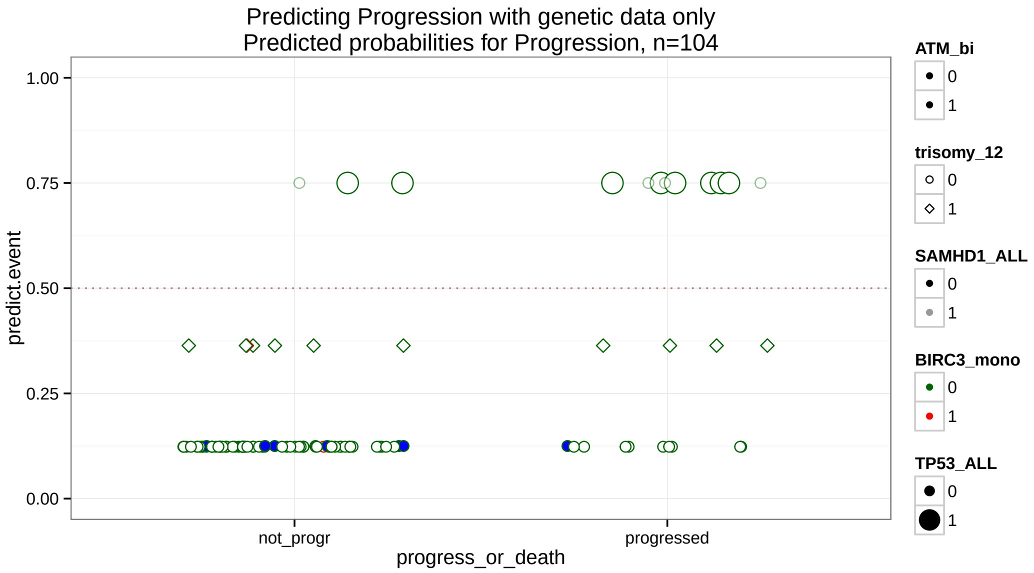
4.3 Logistic regression for patients with survival data

We can now use the subset of patients for which we have both MRD and PFS data to compare the logistic regression models.

The first model uses MRD as response variable and is comparable to the models that we built with the whole data set. The second model uses Progression as response variable. The third model (Combi) uses Progression as response, but includes MRD as predictor.

We can see that using MRD as response is fairly unstable and does not give a good prediction with this small data set. Combining both MRD and genetic data however seems to be a very good predictor for progression. Note that we have quite a number of patients with ATM bi that are MRD positive, but did not progress (yet).


Predicting MRD with genetic data only
Predicted probabilities for MRD positivity, n=104
predict
1.00
0.75
0.50
0.25
0.00
MRD negative
MRD positive
MRD
ATM_bi
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1


Predicting Progression with genetic data only
Predicted probabilities for Progression, n=104
predict.event
1.00
0.75
0.50
0.25
0.00
not_progr
progressed
progress_or_death
ATM_bi
0
1
trisomy_12
0
1
SAMHD1_ALL
0
1
BIRC3_mono
0
1
TP53_ALL
0
1

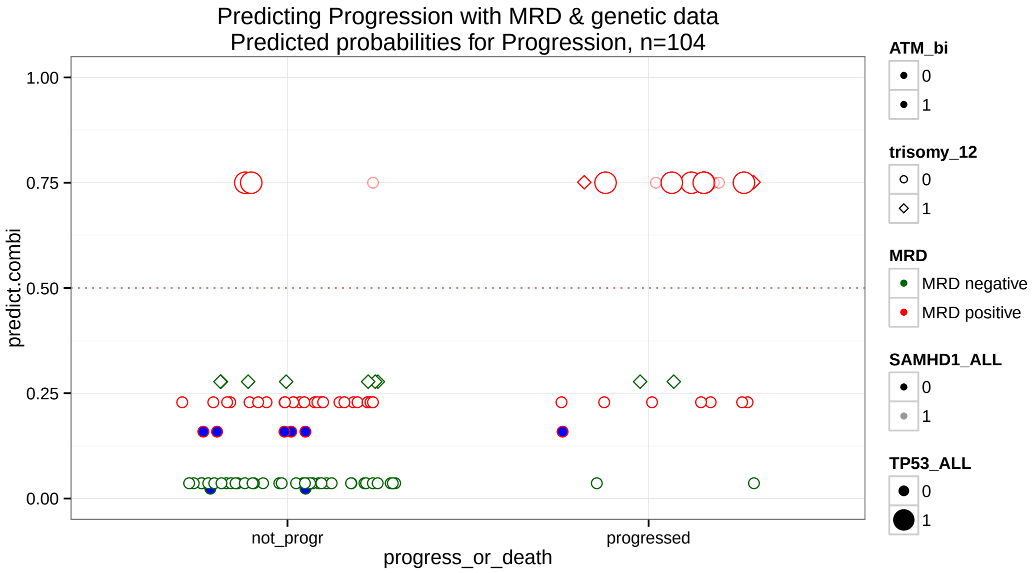

Predicting Progression with MRD & genetic data
Predicted probabilities for Progression, n=104
predict.combi
1.00
0.75
0.50
0.25
0.00
not_progr
progressed
progress_or_death
ATM_bi
0
1
trisomy_12
0
1
MRD
MRD negative
MRD positive
SAMHD1_ALL
0
1
TP53_ALL
0
1

Table 9: Compare MRD and Progression Models, n=104

	Dependent variable:		
	MRD	progress_or_death	
	MRD	Progression	Combi
TP53_ALL1	17.76 (1,398.72)	3.06*** (0.89)	2.31** (0.91)
ATM_bi1	1.29 (0.85)	0.02 (1.13)	−0.45 (1.15)
trisomy_121	−1.31 (0.82)	1.40* (0.72)	2.32** (0.93)
SAMHD1_ALL1	17.76 (1,978.09)	3.06** (1.21)	2.31* (1.22)
MRDMRD positive			2.06** (0.82)
Constant	−0.19 (0.24)	−1.96*** (0.36)	−3.28*** (0.76)
Observations	104	104	104
Log Likelihood	−59.98	−44.23	−40.03
Akaike Inf. Crit.	129.96	98.46	92.05
Note:	*p<0.1; **p<0.05; ***p<0.01		

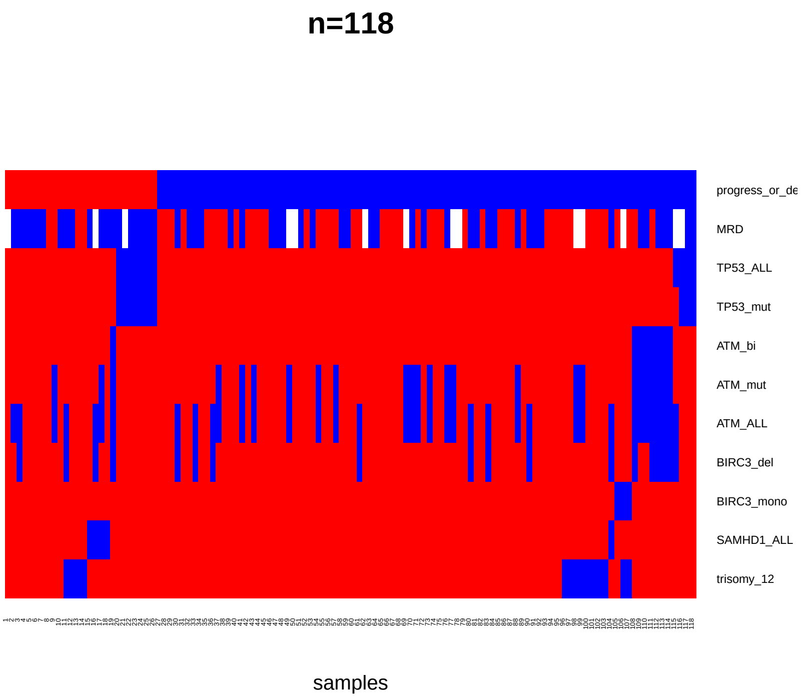
Table 10: Missclassification for pfs models, n=104

p	model	true positive	false positive	true negative	false negative	missclasserr
p1	fit.survlogreg	49	35	18	2	0.356
p2	fit.survlogreg.event	71	13	10	10	0.221
p3	fit.survlogreg.combi	71	13	10	10	0.221

p

5 Patient distributions

You can use this part of the script to generate a nice distribution of patients by filtering for specific traits first, then ordering by these traits in the order you desire.


n=118
progress_or_de
MRD
TP53_ALL
TP53_mut
ATM_bi
ATM_mut
ATM_ALL
BIRC3_del
BIRC3_mono
SAMHD1_ALL
trisomy_12
samples