HICF1 - Final Report v4

Dr. Susanne Weller

August 6, 2014

Contents

1 Univariate Analysis

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. I have used False Discovery Rate, currently the least stringent correction method that was specifically designed for genetics.

Table 1: Univariate Analysis against MRD outcome

p	p	sig	corr.p	sig.corr	MRDpos_0	MRDneg_0	MRDpos_1	MRDneg_1
p ATM_ALL	0.002	∗∗	0.005	∗∗	28%	40%	21%	11%
p ATM_bi	0.002	∗∗	0.005	∗∗	40%	49%	9%	2%
p ATM_del	0	∗∗∗	0	∗∗∗	35%	47%	13%	4%
p ATM_mono	0.836	n.s.	0.836	n.s.	43%	44%	6%	7%
p BIRC3_ALL	0.095	trend	0.129	n.s.	38%	45%	11%	6%
p BIRC3_bi	0.360	n.s.	0.428	n.s.	47%	51%	1%	0%
p BIRC3_del	0.002	∗∗	0.005	∗∗	39%	48%	10%	3%
p BIRC3_mono	0.066	trend	0.101	n.s.	48%	48%	0%	3%
p NOTCH1_mut	0.069	trend	0.101	n.s.	44%	42%	4%	9%
p SAMHD1_ALL	0.054	trend	0.093	trend	45%	50%	4%	1%
p SF3B1_mut	0.415	n.s.	0.464	n.s.	36%	41%	12%	11%
p TP53_ALL	0	∗∗∗	0	∗∗∗	40%	50%	9%	1%
p TP53_bi	0.009	∗∗	0.021	∗	44%	51%	4%	0%
p TP53_lowVAF	0.163	n.s.	0.206	n.s.	46%	50%	3%	1%
p TP53_mut	0	∗∗∗	0	∗∗∗	42%	51%	7%	0%
p Trisomy_12	0.002	∗∗	0.005	∗∗	45%	39%	4%	12%
p X11q_mono	0.046	∗	0.093	trend	44%	50%	5%	1%
p Subclones	0.050	∗	0.093	trend	NA%	NA%	NA%	NA%
pTotal_num_CNAs	0.483	n.s.	0.510	n.s.	NA%	NA%	NA%	NA%

p

2 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 209 patients have mutation X -> probability estimate for this mutation is 8/209
15 out of 209 patients have mutation Y -> probability estimate for this mutation is 15/209
The expected probablity of having both mutations is then 8/209 x 15/209

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.

variables	TP53_ALL	TP53_del	TP53_cnLOH	TP53_mut	ATM_ALL	ATM_mut	ATM_del	ATM_cnLOH	BIRC3_ALL	BIRC3_mut	BIRC3_del	NOTCH1_mut	SF3B1_mut	X6q._del_ALL	X13q_ALL	Trisomy_12	Trisomy_18	Trisomy_19	XPO1_gain	SAMHD1_ALL	MYD88_mut	MED12mutation	X8q_ALL	Subclones	Total_num_CNAs
TP53_ALL		0.00	0.00	0.00	0.00	0.51	0.57	0.50	1.00	0.41	0.34	1.00	0.59	0.39	0.44	1.00	0.63	0.64	0.43	0.42	0.10	0.69	0.08	0.66	0.07
TP53_del			1.00	0.00	0.00	0.27	0.53	0.43	1.00	1.00	0.42	1.00	0.64	0.63	1.00	1.00	0.67	1.00	0.65	0.18	0.47	0.82	0.63	1.00	0.18
TP53_cnLOH				0.01	0.00	1.00	0.63	0.51	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.21	1.00	1.00	0.10	0.60	1.00	0.17	1.00	0.09	0.61
TP53_mut					0.00	0.17	0.19	0.28	1.00	0.63	0.09	0.64	0.19	0.18	1.00	1.00	1.00	1.00	1.00	1.00	0.36	0.88	0.18	0.38	0.10
ATM_ALL						0.64	0.59	0.53	1.00	0.51	0.18	1.00	0.43	0.42	1.00	0.56	0.48	1.00	1.00	0.39	0.58	1.00	0.42	0.54	0.39
ATM_mut							0.00	0.00	0.03	0.00	0.00	0.37	0.00	0.00	0.03	0.00	0.00	0.59	0.44	0.55	0.66	0.80	0.74	0.85	0.19
ATM_del								0.00	0.01	0.00	0.11	0.20	0.02	0.00	0.65	0.06	0.00	0.53	0.16	0.58	0.60	0.52	0.25	0.92	0.35
ATM_cnLOH									1.00	0.28	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.70	0.83	0.40	0.34	1.00	0.34	0.54	0.19
BIRC3_ALL										1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.31	1.00	0.35	0.43	0.47	1.00
BIRC3_mut											0.28	1.00	0.41	0.65	1.00	1.00	0.00	1.00	1.00	0.35	1.00	0.54	0.39	0.65	0.07
BIRC3_del												0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66	1.00	0.74	0.82	0.34	0.71	0.51
NOTCH1_mut													0.04	1.00	0.00	1.00	0.18	0.00	0.05	1.00	0.61	0.86	1.00	1.00	1.00
SF3B1_mut														0.00	0.00	0.00	0.02	0.64	0.45	0.84	0.71	0.90	0.27	0.66	0.44
X6q._del_ALL															1.00	0.42	0.03	1.00	0.59	0.22	0.24	0.50	1.00	0.45	0.42
X13q_ALL																1.00	1.00	1.00	0.49	1.00	1.00	1.00	1.00	1.00	1.00
Trisomy_12																	0.29	1.00	0.71	0.56	0.58	1.00	0.42	0.54	0.15
Trisomy_18																		0.41	0.25	0.86	1.00	0.91	0.21	0.60	0.60
Trisomy_19																			0.05	1.00	0.47	1.00	1.00	1.00	1.00
XPO1_gain																				0.13	0.73	0.08	0.28	0.26	0.46
SAMHD1_ALL																					0.16	0.84	0.54	0.44	0.14
MYD88_mut																						1.00	1.00	0.69	1.00
MED12mutation																							0.00	0.00	0.00
X8q_ALL																								0.00	0.15
Subclones																									0.00
Total_num_CNAs																									

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

variables	TP53_ALL	TP53_del	TP53_cnLOH	TP53_mut	ATM_ALL	ATM_mut	ATM_del	ATM_cnLOH	BIRC3_ALL	BIRC3_mut	BIRC3_del	NOTCH1_mut	SF3B1_mut	X6q._del_ALL	X13q_ALL	Trisomy_12	Trisomy_18	Trisomy_19	XPO1_gain	SAMHD1_ALL	MYD88_mut	MED12mutation	X8q_ALL	Subclones	Total_num_CNAs
TP53_ALL		0.00	0.01	0.00	0.00	1.00	1.00	1.00	1.00	0.99	0.98	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.99	0.99	0.46	1.00	0.41	1.00	0.37
TP53_del			1.00	0.00	0.00	0.85	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.68	1.00	1.00	1.00	1.00	0.68
TP53_cnLOH				0.03	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.71	1.00	1.00	0.46	1.00	1.00	0.68	1.00	0.43	1.00
TP53_mut					0.00	0.68	0.68	0.85	1.00	1.00	0.42	1.00	0.68	0.68	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.68	0.99	0.46
ATM_ALL						1.00	1.00	1.00	1.00	1.00	0.68	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	0.99
ATM_mut							0.00	0.00	0.18	0.01	0.00	0.99	0.00	0.00	0.18	0.00	0.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.68
ATM_del								0.02	0.07	0.00	0.50	0.69	0.10	0.00	1.00	0.30	0.00	1.00	0.65	1.00	1.00	1.00	0.82	1.00	0.98
ATM_cnLOH									1.00	0.86	0.00	0.11	0.00	0.00	0.01	0.00	0.02	1.00	1.00	0.99	0.98	1.00	0.98	1.00	0.68
BIRC3_ALL										1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.98	0.99	1.00	1.00
BIRC3_mut											0.85	1.00	0.99	1.00	1.00	1.00	0.00	1.00	1.00	0.98	1.00	1.00	0.99	1.00	0.34
BIRC3_del												0.00	0.00	0.00	0.01	0.00	0.02	0.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
NOTCH1_mut													0.23	1.00	0.00	1.00	0.68	0.00	0.28	1.00	1.00	1.00	1.00	1.00	1.00
SF3B1_mut														0.00	0.00	0.00	0.10	1.00	0.99	1.00	1.00	1.00	0.85	1.00	0.99
X6q._del_ALL															1.00	0.99	0.15	1.00	1.00	0.73	0.80	1.00	1.00	0.99	0.99
X13q_ALL																1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Trisomy_12																	0.86	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.64
Trisomy_18																		0.99	0.82	1.00	1.00	1.00	0.71	1.00	1.00
Trisomy_19																			0.25	1.00	1.00	1.00	1.00	1.00	1.00
XPO1_gain																				0.57	1.00	0.39	0.85	0.84	1.00
SAMHD1_ALL																					0.66	1.00	1.00	0.99	0.63
MYD88_mut																						1.00	1.00	1.00	1.00
MED12mutation																							0.01	0.00	0.00
X8q_ALL																								0.00	0.64
Subclones																									0.00
Total_num_CNAs																									

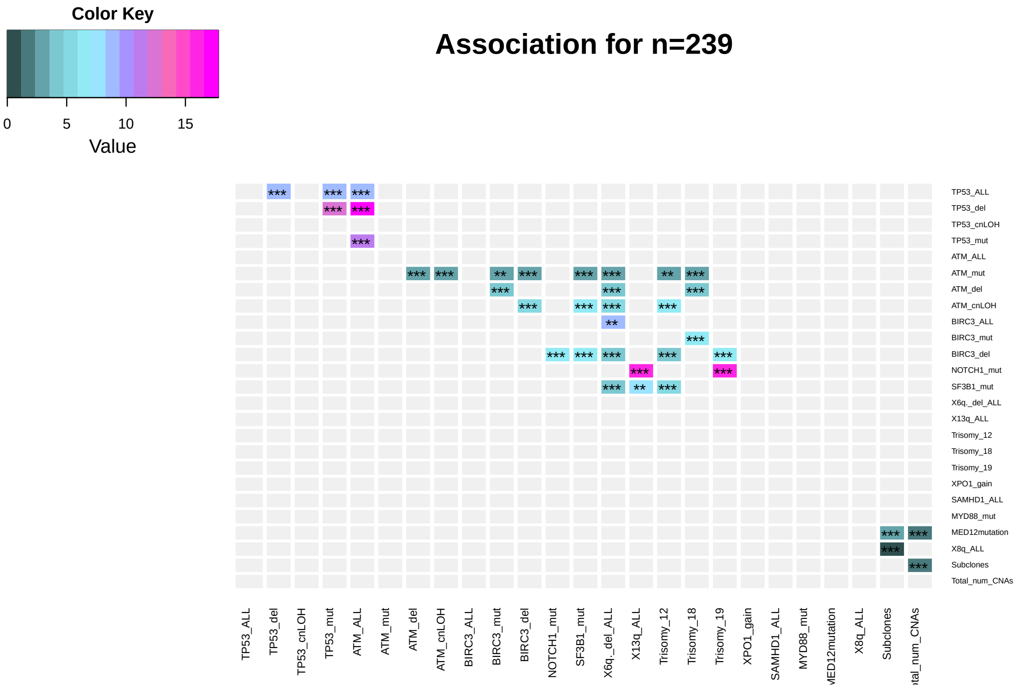
Table 3: Association chart, corrected pvalues, Fisher's test with FDR correction

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.

variables	TP53_ALL	TP53_del	TP53_cnLOH	TP53_mut	ATM_ALL	ATM_mut	ATM_del	ATM_cnLOH	BIRC3_ALL	BIRC3_mut	BIRC3_del	NOTCH1_mut	SF3B1_mut	X6q._del_ALL	X13q_ALL	Trisomy_12	Trisomy_18	Trisomy_19	XPO1_gain	SAMHD1_ALL	MYD88_mut	MED12mutation	X8q_ALL	Subclones	Total_num_CNAs
TP53_ALL		8.86	8.66	9.22	9.00	0.71	0.67	0.60	0.00	1.43	0.42	0.57	0.57	0.31	1.71	0.61	1.16	0.00	1.34	1.27	2.30	0.86	0.00	1.10	0.30
TP53_del			0.00	12.43	17.77	0.28	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.29	0.00	0.00	0.00	1.60	0.74	0.00	0.95	0.00
TP53_cnLOH				9.06	12.95	0.72	0.00	1.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.28	0.00	0.00	3.76	1.11	0.00	1.88	0.00	2.39	1.07
TP53_mut					11.68	0.42	0.37	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.97	0.00	0.74	0.88	1.60	1.03	0.00	1.32	0.21
ATM_ALL						0.61	0.54	0.40	0.00	1.42	0.00	0.00	0.00	0.00	0.00	1.22	1.39	0.00	0.53	0.31	1.14	0.93	0.00	1.19	0.30
ATM_mut							3.55	3.28	2.92	2.98	2.51	1.35	3.15	3.12	2.92	3.00	3.12	0.57	0.71	1.13	1.16	1.04	1.07	1.03	1.32
ATM_del								2.09	3.85	3.94	1.55	0.25	1.96	4.15	0.00	0.00	4.17	0.38	0.47	1.13	1.27	1.11	1.42	1.02	1.24
ATM_cnLOH									0.00	0.00	5.10	2.71	6.36	5.01	5.79	5.98	0.00	1.14	1.07	1.28	1.53	0.98	0.42	1.14	1.45
BIRC3_ALL										0.00	0.00	0.00	0.00	9.02	0.00	0.00	0.00	0.00	0.00	1.78	0.00	1.50	1.77	1.43	0.85
BIRC3_mut											0.00	0.00	0.00	0.00	0.00	0.00	6.75	0.00	0.62	1.48	0.00	1.25	1.48	1.19	2.16
BIRC3_del												6.31	6.69	3.72	6.08	4.41	0.00	6.19	1.12	1.00	1.20	0.93	0.44	1.07	1.18
NOTCH1_mut													2.69	0.59	16.25	0.00	0.00	16.59	2.52	0.88	1.06	0.87	0.59	0.95	0.85
SF3B1_mut														4.66	8.12	5.90	0.00	0.00	0.49	1.03	1.06	0.93	0.29	1.11	1.29
X6q._del_ALL															0.00	0.00	0.00	0.88	0.55	1.49	1.78	1.19	0.98	1.24	1.27
X13q_ALL																0.00	0.00	0.00	1.50	0.88	0.00	0.74	0.00	0.95	0.85
Trisomy_12																	0.00	0.00	1.07	1.27	1.14	0.93	0.00	1.19	1.85
Trisomy_18																		0.00	0.40	0.83	0.86	1.01	1.70	0.82	1.16
Trisomy_19																			3.02	0.88	1.60	0.93	0.88	0.95	0.85
XPO1_gain																				0.41	0.50	0.57	0.27	0.65	0.66
SAMHD1_ALL																					1.79	1.04	0.65	1.16	1.46
MYD88_mut																						1.00	0.59	1.11	0.85
MED12mutation																							1.98	2.50	2.13
X8q_ALL																								0.00	1.61
Subclones																									2.17
Total_num_CNAs																									

Table 4: Odds ratios for association between genes



3 Model building - from here, only 209 data points will be used

3.1 Multiple logistic regression models

The goal is to compare several different models and their quality, and eventually compare them to clinical parameters that are currently used.
We first built a model with parameters that come out significant in the univariate analysis or have been described in the literature (genetic1). We can see that Trisomy12, NOTCH1 and BIRC3mono do not contribute to the model. There could be two reasons for this:
(1) They really do not contribute to the model
(2) There is a colinearity (or in factors, co-occurence) that did not show up on the associaton chart.

In order to see which one contributes most to the model, I built three models with only one of the them in:

- genetic2 : NOTCH1
- genetic3: Trisomy12
- genetic4: BIRC3mono

You can see that NOTCH1 alone is not contributing to the model, Trisomy12 is contributing (and improving the AIC and Log Likelihood), and BIRC3mono contributes, but apparently towards MRD negativity, and with quite large variance(the number in brackets).

Table 5: Multiple log regression, n=209

	Dependent variable:						
				MRD			
	genetic1	genetic2	genetic3	genetic4	genetic5	vhmut	Binet
TP53_ALL1	2.62^{***} (0.77)	2.74^{***} (0.76)	2.64^{***} (0.77)	2.68^{***} (0.76)			
ATM_bi1	1.66^{***} (0.56)	1.67^{***} (0.54)	1.61^{***} (0.54)	1.85^{***} (0.55)			
BIRC3_mono1	−2.03 (1.24)			-2.27^{*} (1.18)			
Trisomy_121	−0.65 (0.47)		-0.87^{*} (0.45)				
NOTCH1_mut1	−0.51 (0.51)	−0.74 (0.48)					
SAMHD1_ALL1	1.95^{**} (0.89)	1.80^{**} (0.82)	1.65^{**} (0.82)	2.10^{**} (0.89)			
Subclones1					0.52^{*} (0.28)		
vh_mutation_statusunmutated						1.15^{***} (0.31)	
BinetC							0.12 (0.29)
Constant	−0.30 (0.19)	-0.42^{**} (0.18)	-0.37^{**} (0.18)	-0.47^{***} (0.17)	−0.29 (0.19)	-0.75^{***} (0.24)	−0.09 (0.17)
Observations	209	209	209	209	209	181	209
Log Likelihood	−121.59	−124.83	−124.07	−123.34	−143.10	−118.14	−144.73
Akaike Inf. Crit.	257.19	259.65	258.13	256.68	290.19	240.28	293.46
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	genetic1	99	59	43	8	0.321	0
p2	genetic2	98	58	44	9	0.321	0
p3	genetic3	98	58	44	9	0.321	0
p4	genetic4	99	59	43	8	0.321	0
p5	genetic5	64	48	54	43	0.435	0
p6	Binet	72	66	36	35	0.483	0
p7	vhmutation	55	26	60	40	0.365	0.139
p							

4 Random Forest for variable importance

We will use the variables of the multivariate regression model *genetic3*:

- TP53_ALL
- ATM_bi
- Trisomy_12
- SAMHD1_ALL

Here is our model:

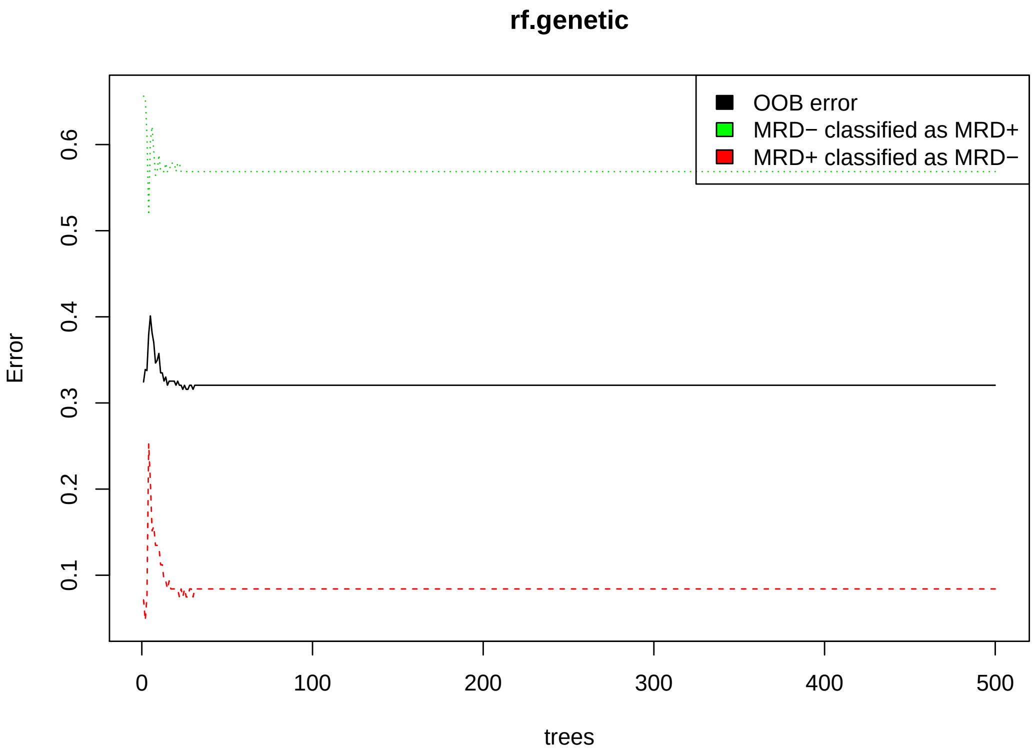
```
Call:
 randomForest(formula = MRD ~ ., data = treegenetic, importance = TRUE,      ntree = 500, mtry =
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 32.06%
Confusion matrix:
                 MRD_MRD negative MRD_MRD positive class.error
MRD_MRD negative               98                9  0.08411215
MRD_MRD positive               58               44  0.56862745
```

4.1 Determine performance and tuning the model

We could potentially improve our model or shift its focus by changing the main tuning parameters:

- Number of trees
- Weighted classes
- Decision cutoff

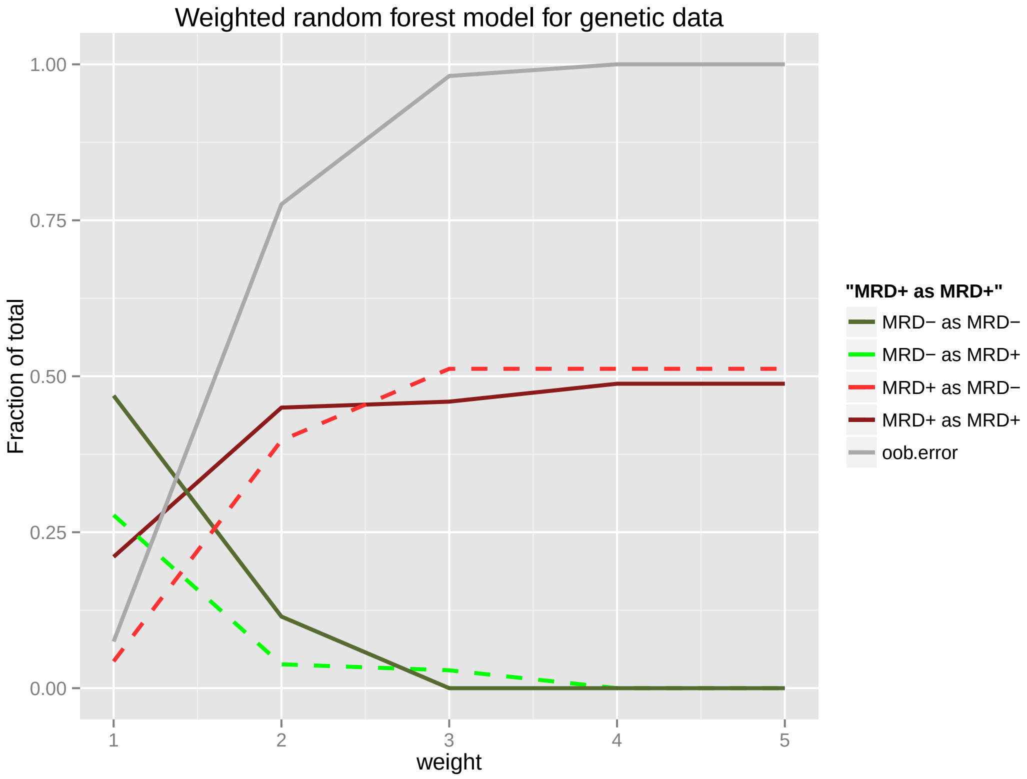
4.1.1 Number of trees



Conclusion: The number of trees does not seem to play an important role, 500 should be sufficient.

4.1.2 Weighted class

The focus of our study is to find predictors for ”MRD positive”(to give him/her access to the more expensive drug). There is cost of increasing true negative findings (real ”MRD positives”) as we will also generate more false positive findings (”MRD negatives” that are classified as positives).
We can incorporate class weights into the random forest classifier, thus making it more sensitive to find MRD positives. The resulting errors are shown below:



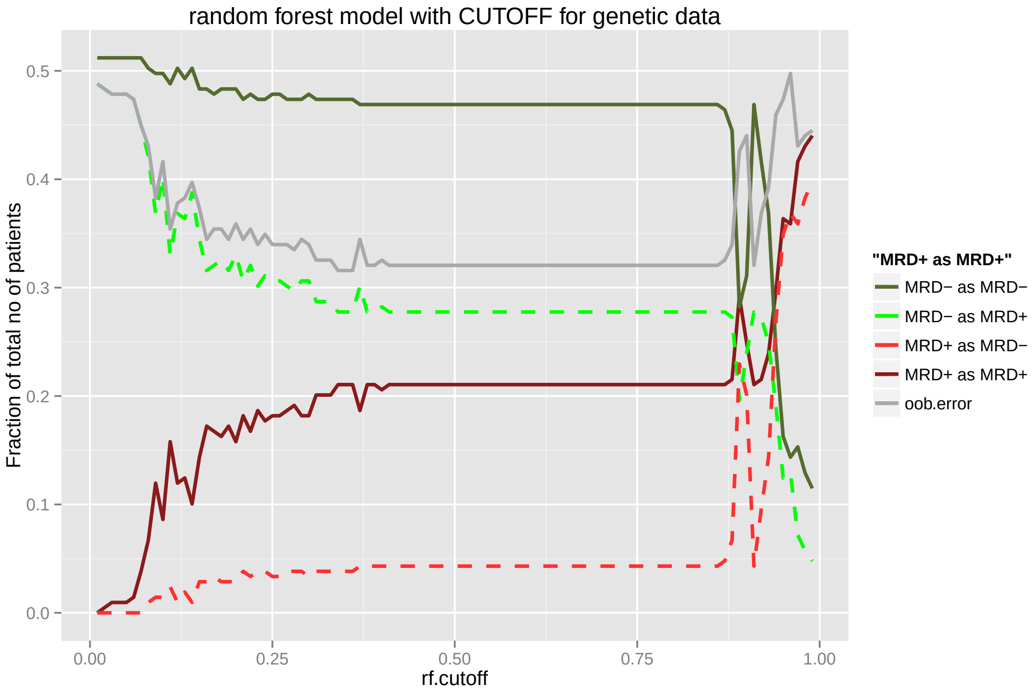
Here is an example for a model with weight slightly shifted towards finding more MRD positives:

```
Call:
 randomForest(formula = MRD ~ ., data = treegenetic, importance = TRUE,      classwt = c(1, 2), n
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 43.54%
Confusion matrix:
                 MRD_MRD negative MRD_MRD positive class.error
MRD_MRD negative               24               83  0.77570093
MRD_MRD positive                8               94  0.07843137
```

Conclusion: Introducing more MRD positive findings makes the model produce more false positives as well. Not very helpful...

4.1.3 Cutoff selection

We can vary the cutoff that is used in the single decision trees (that are combined in the forest model) such that there is an emphasis on putting patients into MRD+ groups.



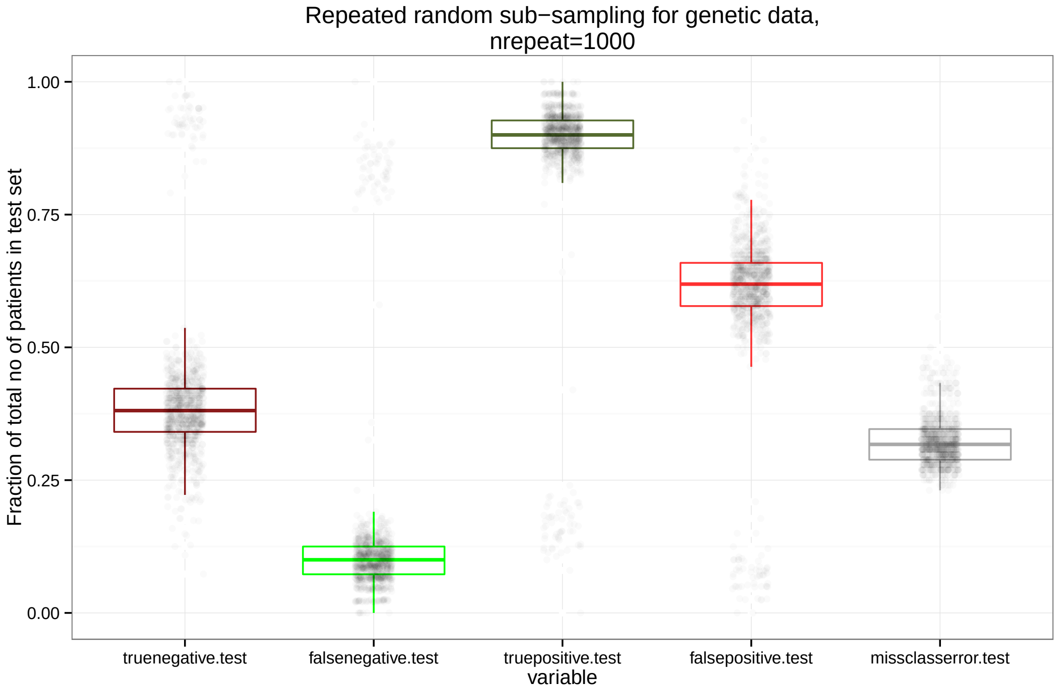
Conclusion: The model seems to be pretty stable for a cutoff between 0.5 and 0.75 and gets messy afterwards, so a change in cutoff is not advisable.

5 Estimating error and variation of Random forest

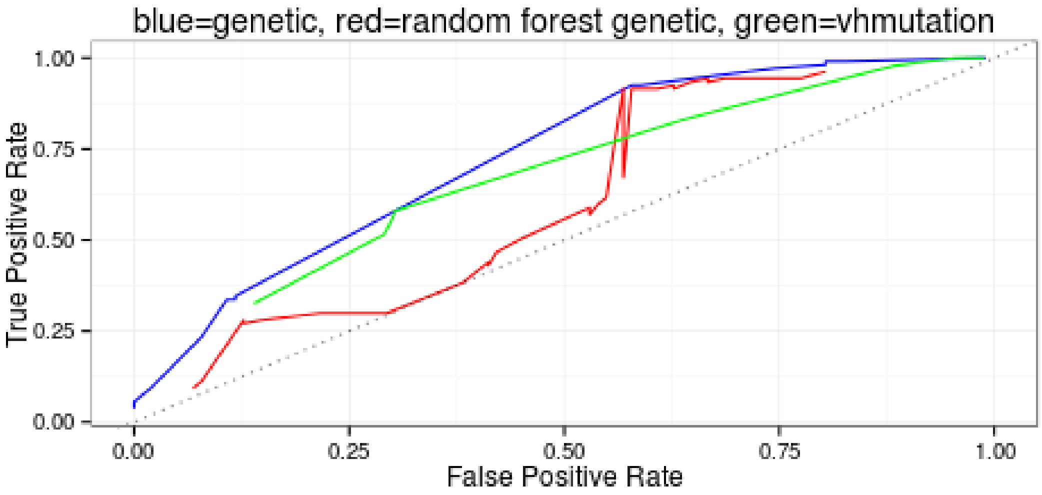
To estimate error and variaton of the model, I devide our dataset at random 1:10 *(ask Chris about this)* into a a training set and a test set. I then train a model on the training set and check it's performance on the test set. This process is repeated 100 times.

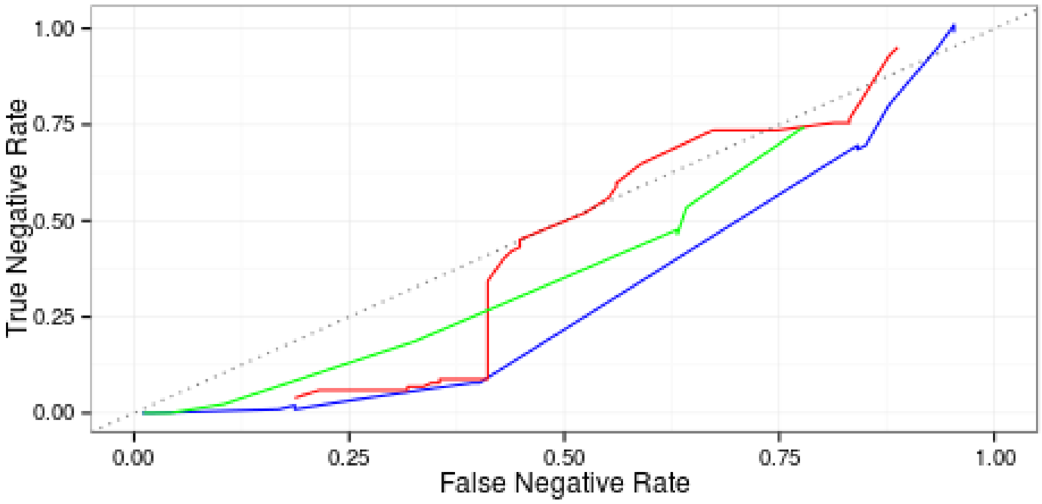
5.1 Repeated random sub-sampling

The next figure shows how the different errors vary when choosing random subsets for training the model



5.2 ROC comparison of different models





Conclusion:
(1) The log regression model does better than the vhmutation model as it is more specific (still not very good though, it is supposed to hug the top right corner more).
(2) The random forest is still doing some really funny things, probably because it is highly overfitted to our data Apparently, calculating a ROC for random forests without deviding into train and test set is highly overoptimistic...