Variant calling algorithm

decision making based on
binomial distribution

Miona Rankovic 2019/3126

Variant Calling Algorithm

```
21     9590334     C     184     A$A$A$A$AAAAAAAAAA..A,AaaA..AAAAAAAaA.Aaa.AAA.AAAa.A.A.A...AA.A.aAAaAA.aaaa.a.A.a....Aaa
AaaA..AAAA,,.A.,a.a.,Aa,a.A.aAAaaaaAAAa.AaaAaAaa,A,a.aaaaa,A.,aaaAa,,aaaaaAaa.A,.,A,aaaAa,a,a,aa,aaa     AFAE./1R=?^==
@mD_C??@@Ij=A@@@>B@@s@AAGA@?I>>@AK=r@dAJJg@@I>j@>>@@@t@@@@r@s>J?sqJJ>@AAA@ABoAAA?H>s>tJ@e@IH@BJ>r?i@@@A@BB>?A@E<C@>B@
CBJ>JBmAB@>@J@JJA?C@CJJ@@@BA@BAA@J=J@IAAB?@IBIAI@0@//.
```

PILEUP FILE

Variant Calling

VCF FILE

```
21      9590334 .       C       A       93.0077 .       DP=214;VDB=0.822087;SGB=-0.693147;RPB=0.713057;MQB=5.22737e-2
0;MQSB=0.124555;BQB=1.31323e-19;MQ0F=0.172897;AF1=0.5;AC1=1;DP4=36,19,68,61;MQ=32;FQ=96.0159;PV4=0.143592,1.60789e-2
3,1.55109e-30,0.0654315 GT:PL   0/1:123,0,255
```

Binomial distribution

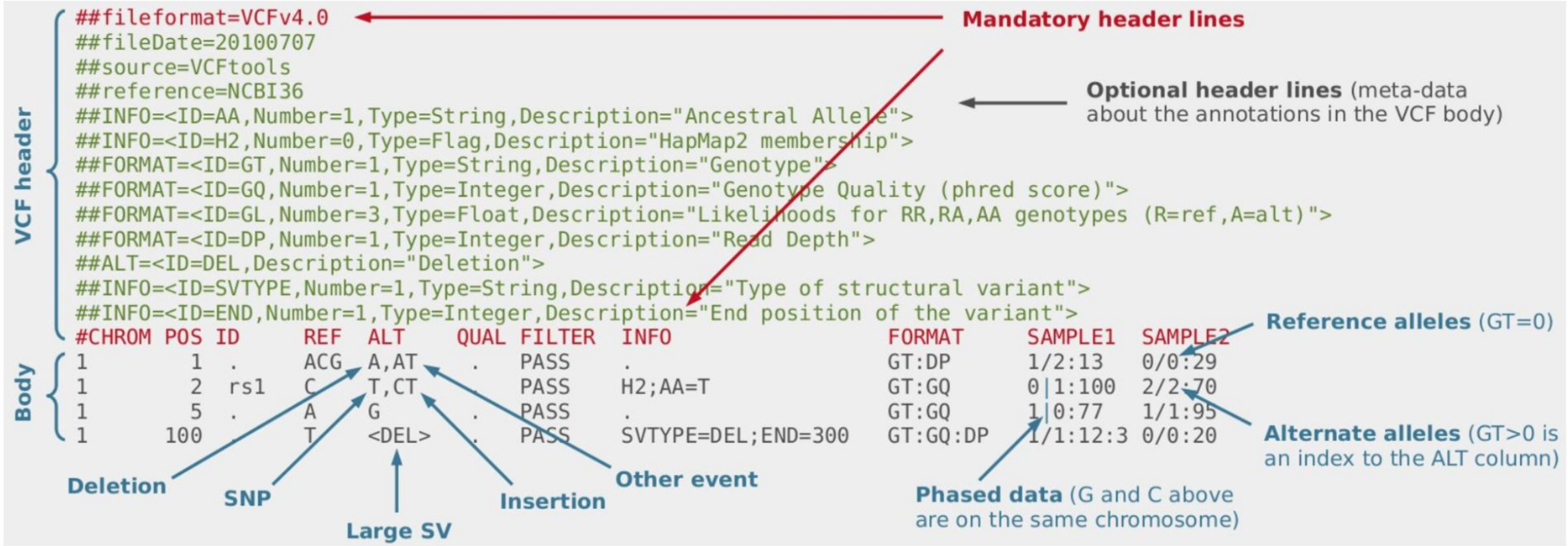
number of trials - total number of sequences with bases mapped to current position

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

probability of success in a single trial

- calculate probabilities of each possible genotype
- choose the most probable one

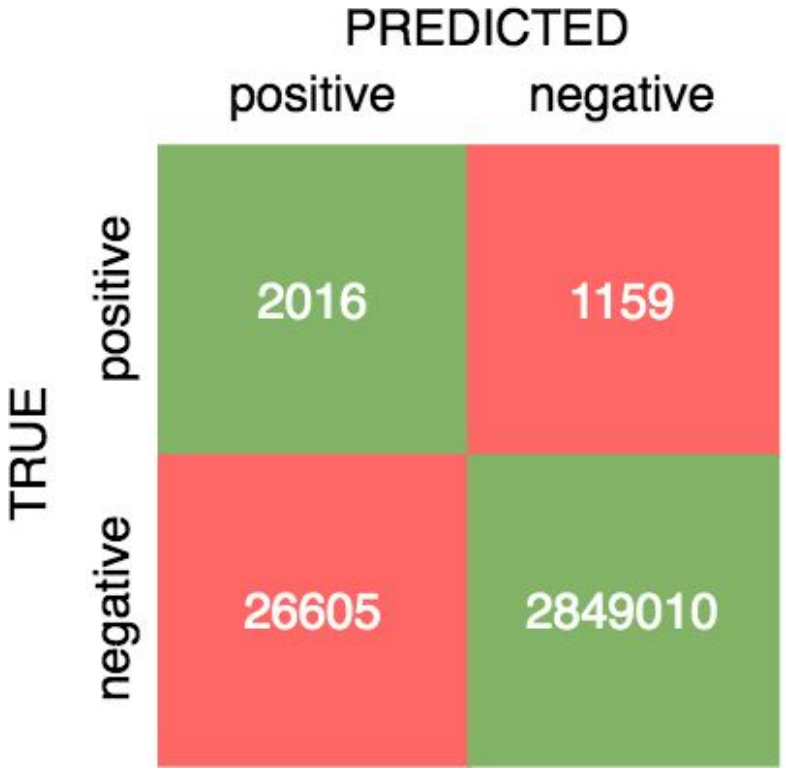
Results ⟶ VCF format



Data

- Pileup file generated with samtools pileup tool from bam and reference file
  - bam file created from the exome portions of chromosomes 21, 22, Y and MT
  - fasta file used in 1000 genomes phase 3 project
- Test file for the purpose of comparing results generated with bcftools call tool

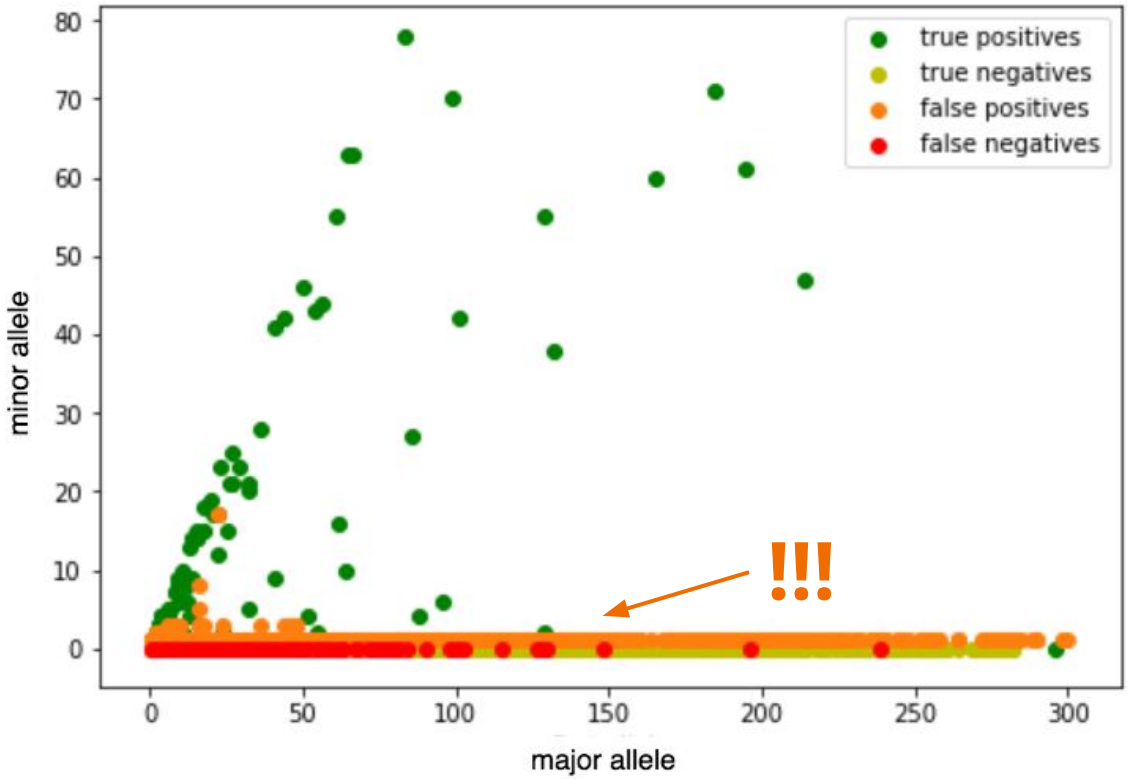
p = 0.9



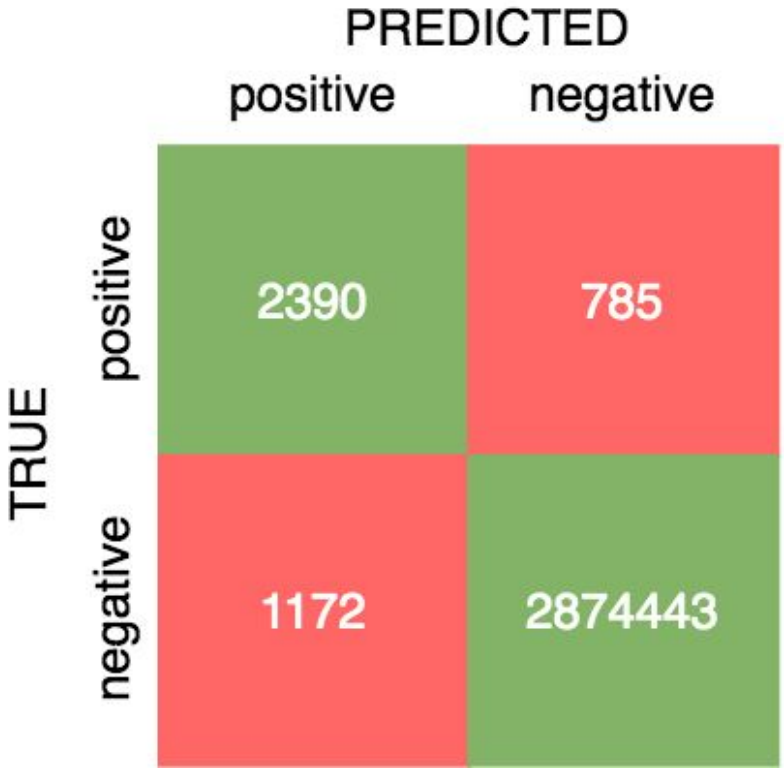
- bad results
- unacceptably low precision

precision: 7.0437 %
recall: 63.4940 %
F1 score: 12.5994 %

p = 0.9


true positives
true negatives
false positives
false negatives
minor allele
major allele
!!!
0
10
20
30
40
50
60
70
80
0
50
100
150
200
250
300

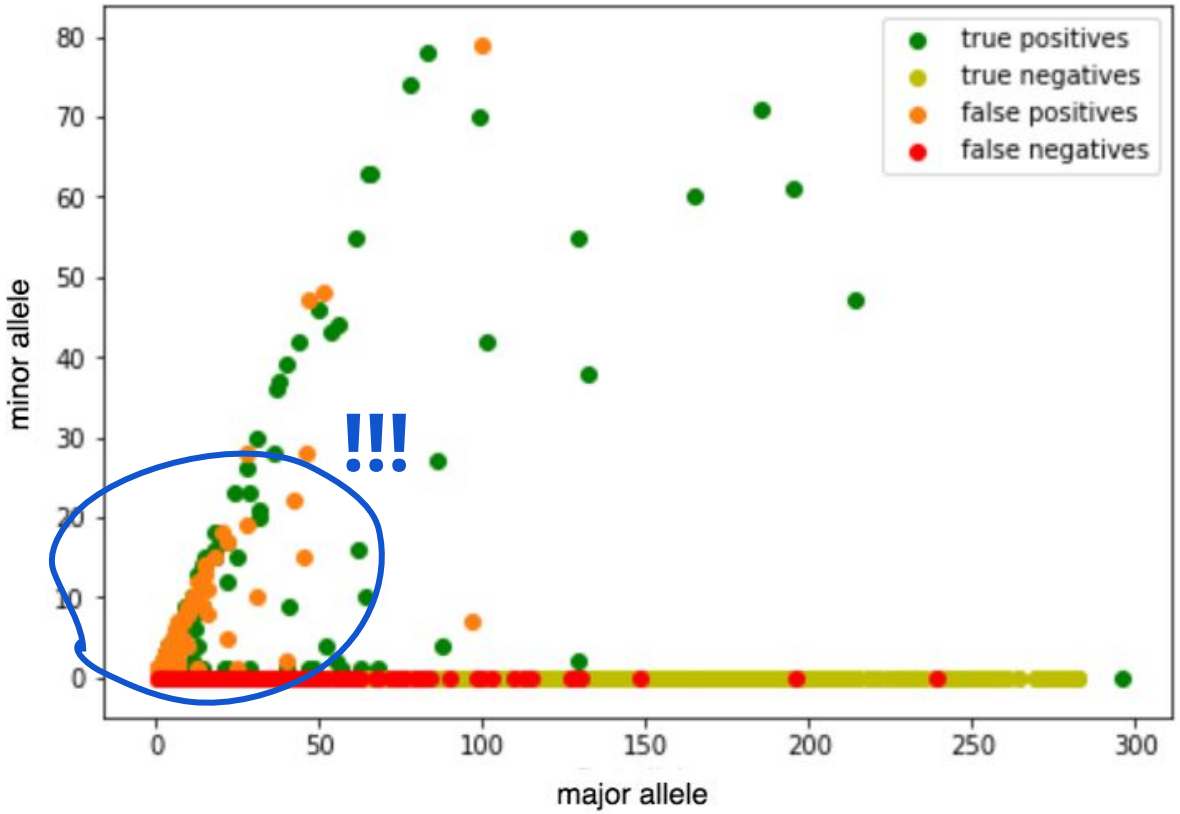
p = 0.8



- much better
- number of good predictions increased, number of bad predictions decreased

precision: 67.0971 %
recall: 75.2756 %
F1 score: 70.9515 %
F2 score: 73.4842 %

p = 0.8

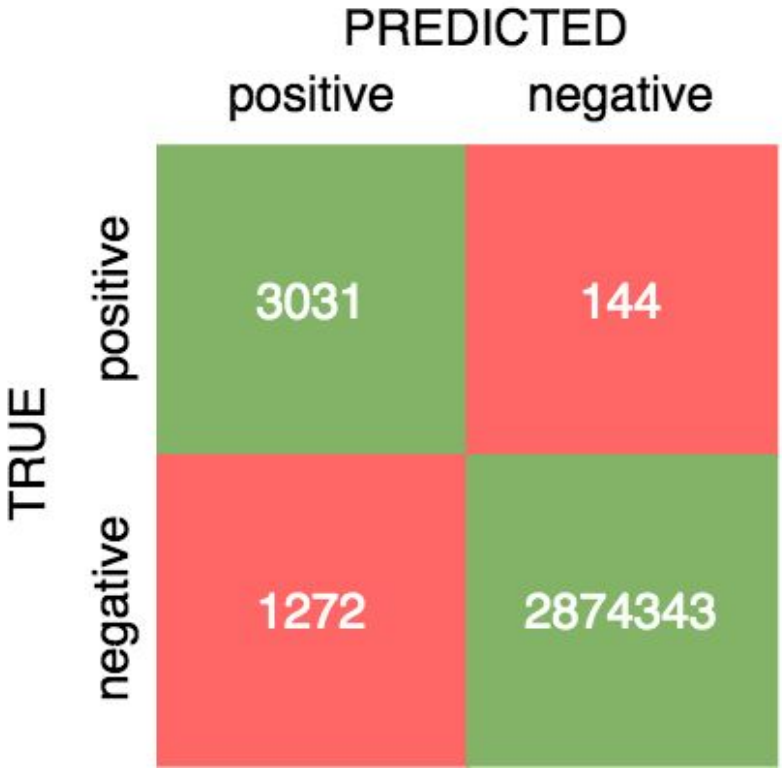

true positives
true negatives
false positives
false negatives
minor allele
major allele
!!!
0
10
20
30
40
50
60
70
80
0
50
100
150
200
250
300

Modified p

- overlapping regions - existing information not enough
- Indels appear less frequently!

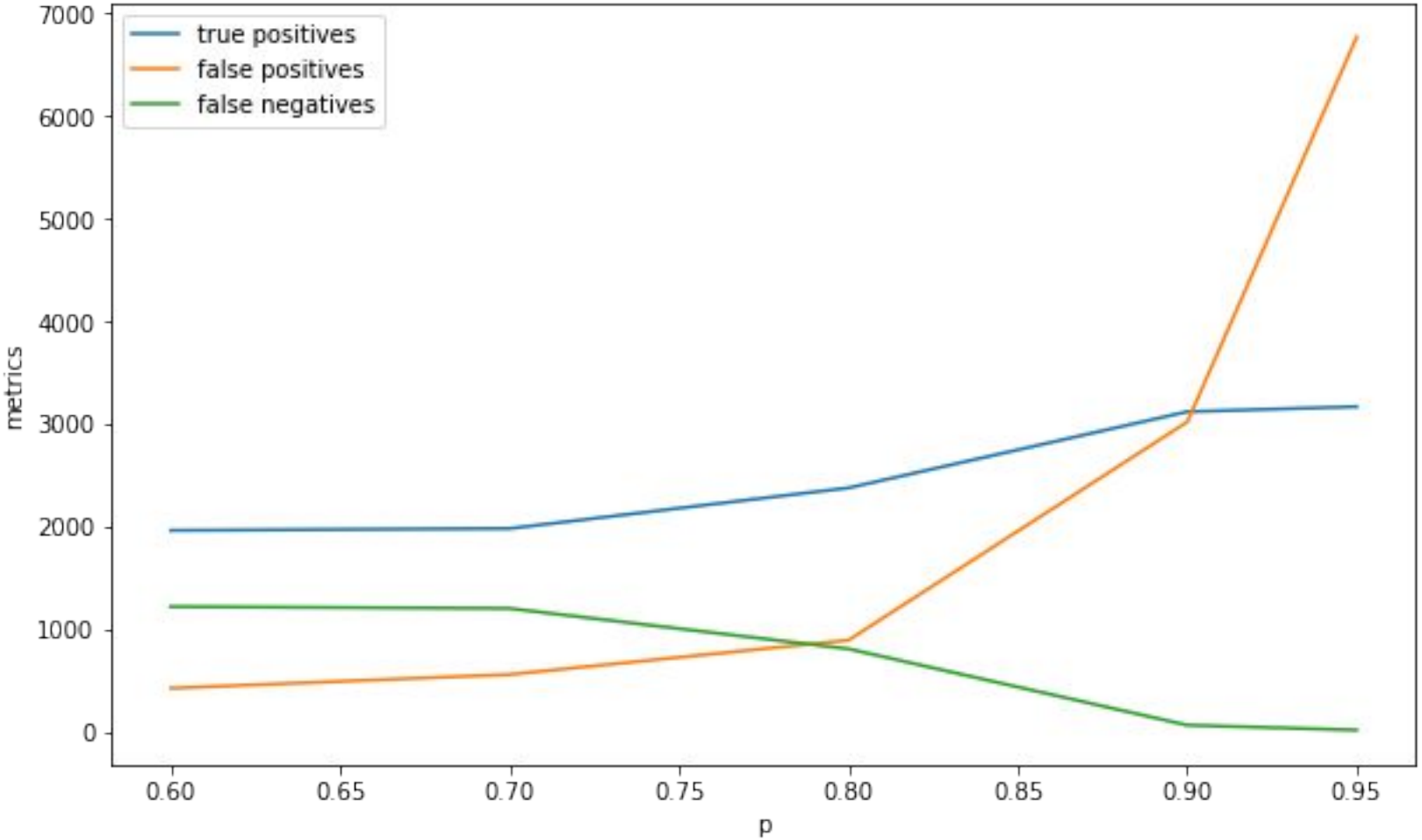
$$p = \begin{cases} 0.8, & \text{for SNVs and matches} \\ 0.6, & \text{for indels} \end{cases}$$

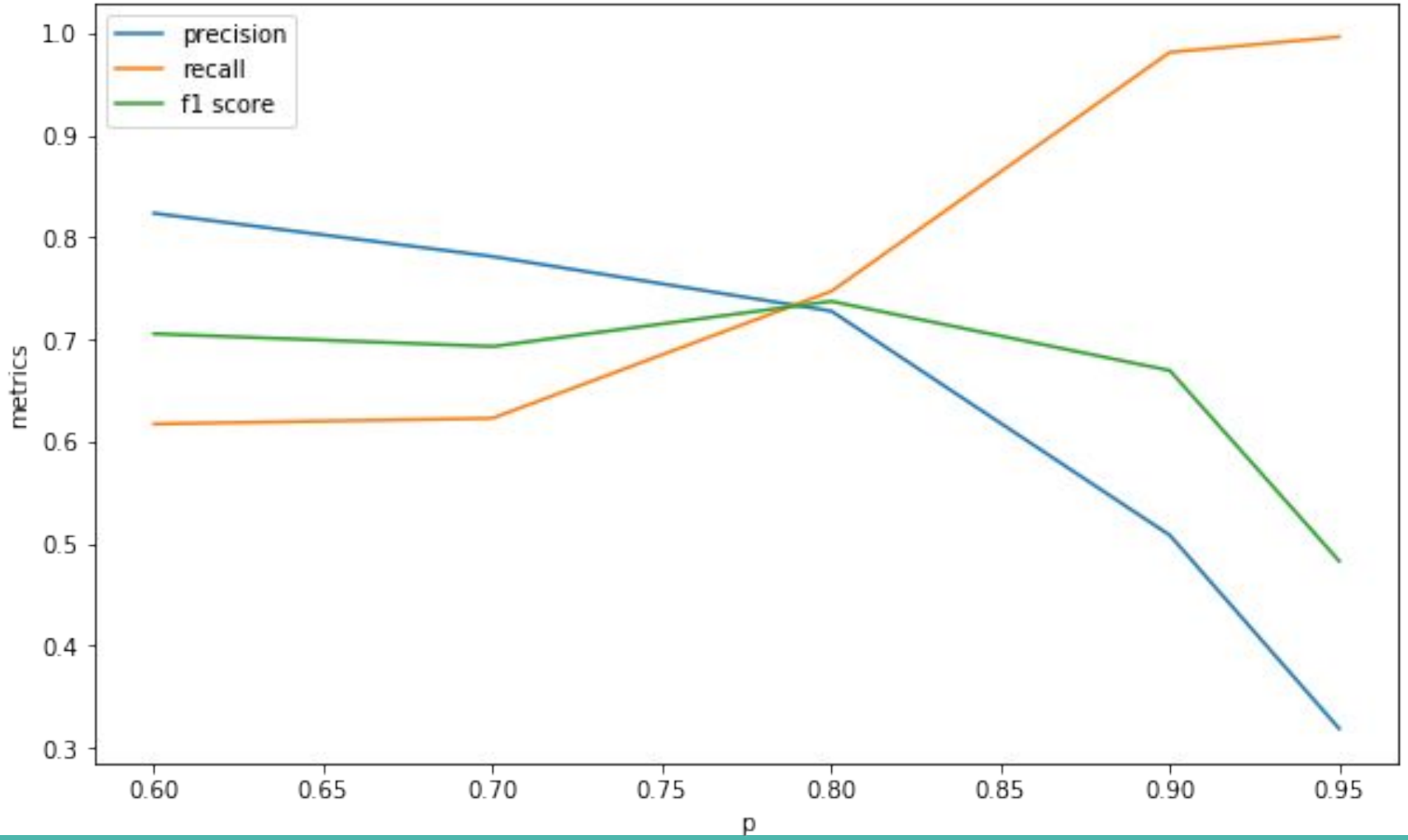
Modified p



- Almost 96% of all mutations found!!!
- Low coverage reads - could probably improve with filtering

precision: 70.4392 %
recall: 95.4646 %
F1 score: 81.0645 %
F2 score: 84.3076 %


true positives
false positives
false negatives
7000
6000
5000
4000
3000
2000
1000
0
metrics
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
p


precision
recall
f1 score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
metrics
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
p