
Variant calling algorithm

— decision making based on —
binomial distribution

Variant Calling Algorithm

```
21      9590334      C      184      A$A$A$A$AAAAAAAAA..A,AaaA..AAAAAAaA.Aaa.AAA.AAAa.A.A.A...AA.A.aAAaAA.aaaa.a.A.a....Aaa
AaaA..AAAA,,.A.,a.a.,Aa,a.A.aAAAAaAAAA.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

VCF header

```
##format=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Annotations:

- Mandatory header lines:** ##format=VCFv4.0, ##fileDate=20100707, ##source=VCFtools, ##reference=NCBI36
- Optional header lines (meta-data about the annotations in the VCF body):** ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">, ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">, ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">, ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">, ##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">, ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">, ##ALT=<ID=DEL,Description="Deletion">, ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">, ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
- Reference alleles (GT=0):** A, T, G, T
- Alternate alleles (GT>0 is an index to the ALT column):** AT, CT, G,
- Deletion:**
- SNP:** C to T
- Large SV:** A to AT
- Insertion:** T to CT
- Other event:** G to G
- Phased data (G and C above are on the same chromosome):** 0|1:100

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

		PREDICTED	
		positive	negative
TRUE	positive	2016	1159
	negative	26605	2849010

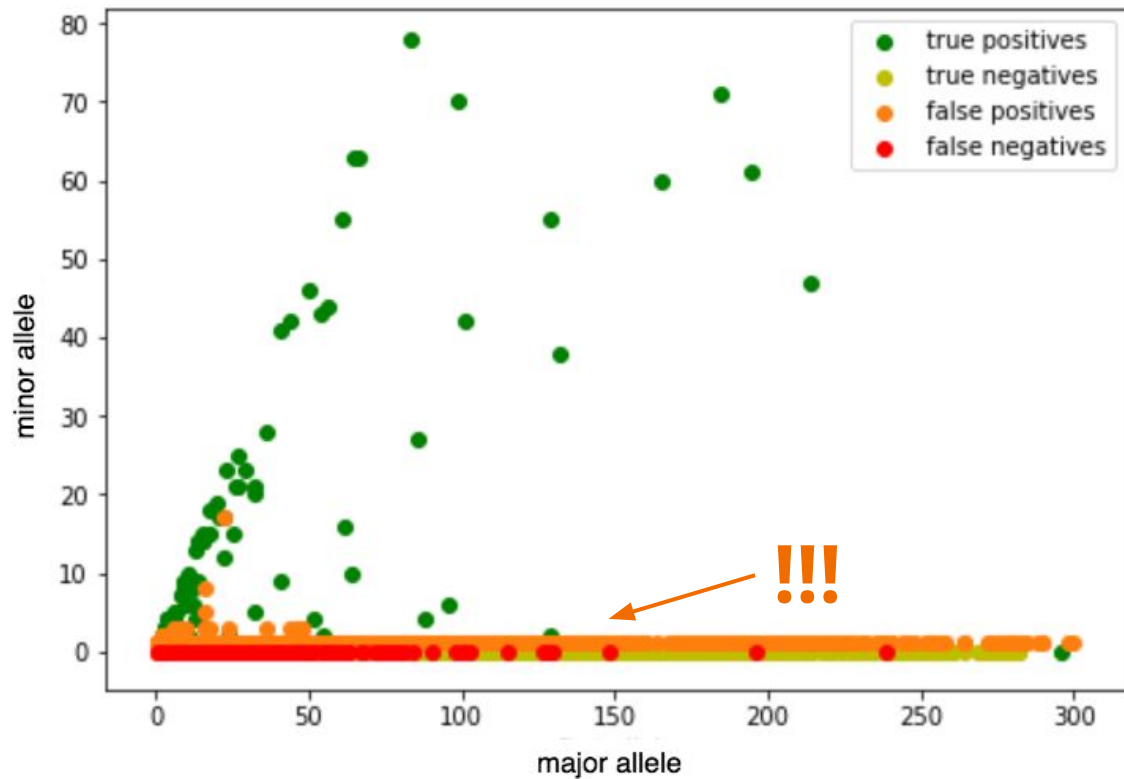
- bad results
- unacceptably low precision

precision: 7.0437 %

recall: 63.4940 %

F1 score: 12.5994 %

$p = 0.9$



p = 0.8

		PREDICTED	
		positive	negative
TRUE	positive	2390	785
	negative	1172	2874443

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

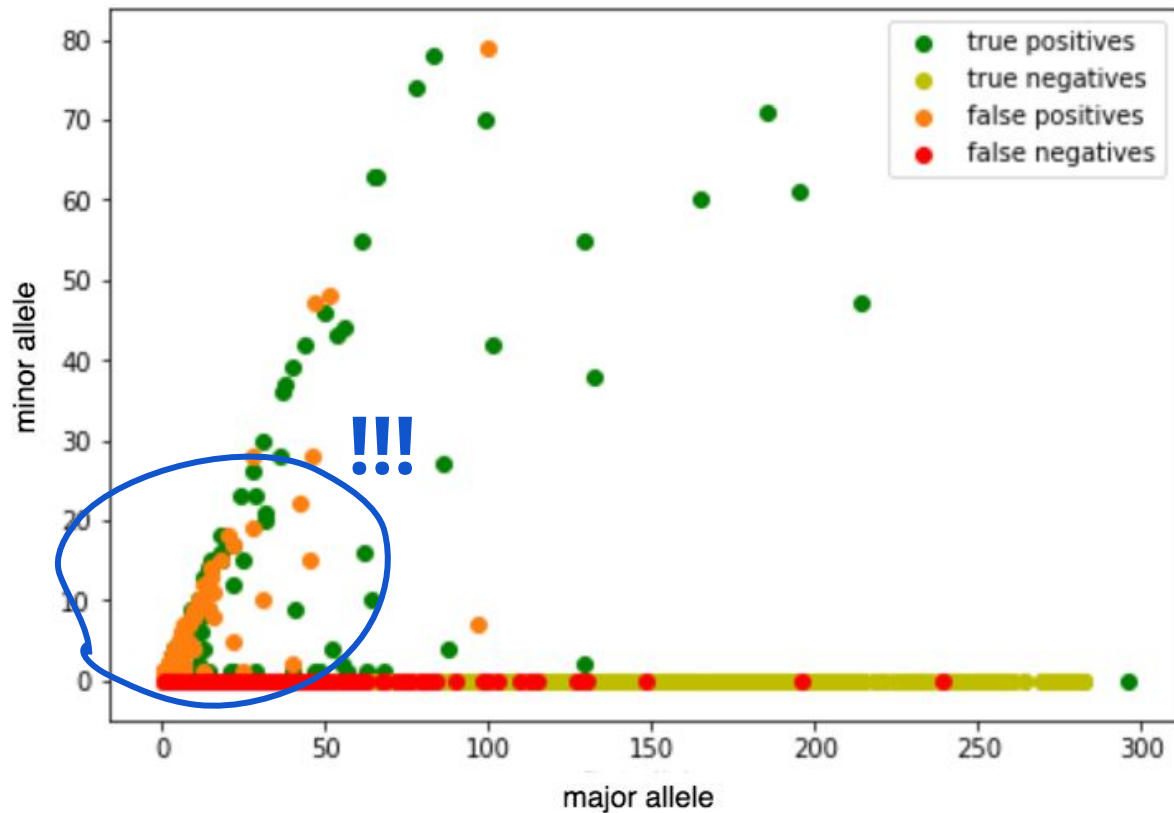
precision: 67.0971 %

recall: 75.2756 %

F1 score: 70.9515 %

F2 score: 73.4842 %

$p = 0.8$



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

		PREDICTED	
		positive	negative
TRUE	positive	3031	144
	negative	1272	2874343

- 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 %

