Improved Mutation Scoring with NextGENe® Software

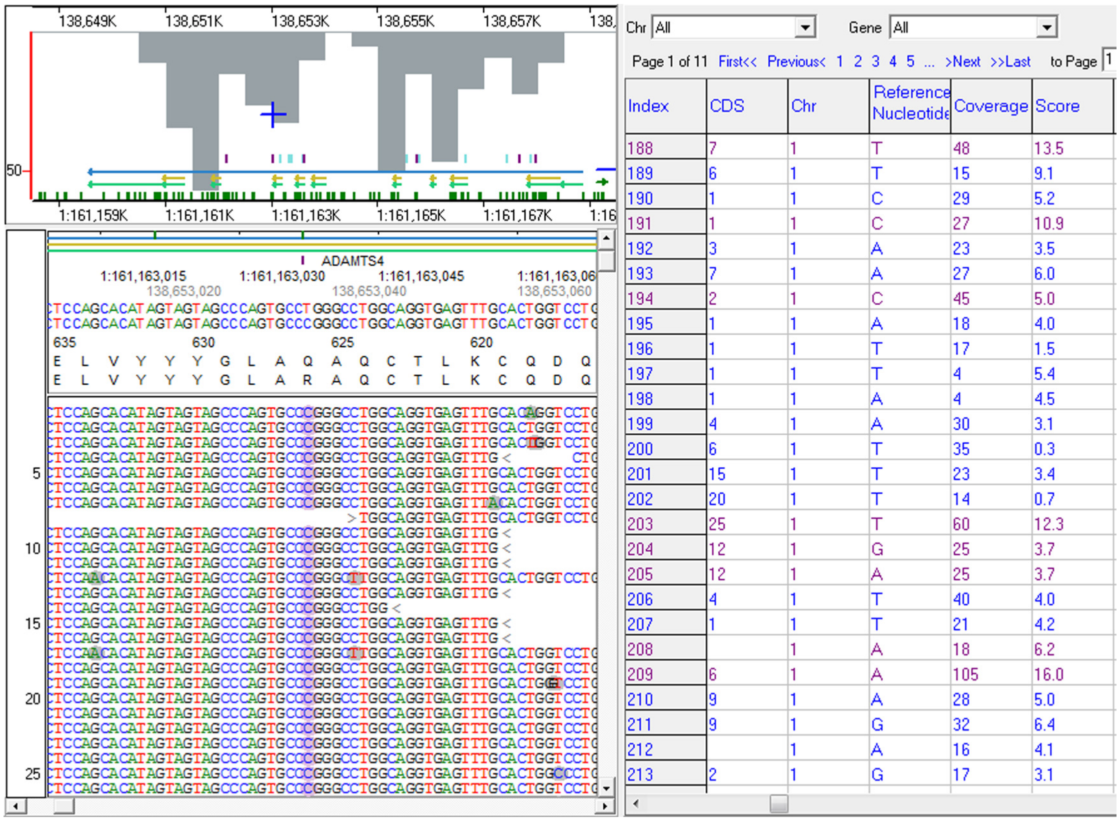
December 2011

John McGuigan, Megan Manion, CS Jonathan Liu

Introduction

When analyzing next-generation sequencing data there are many variables that affect the accuracy of mutation calls. The level of coverage, the fraction of reads with the mutation, and the fraction of reads aligned in the forward or reverse direction are always important. The possibility of misalignment as well as potential homopolymer errors also need to be considered. This can be overwhelming for large alignments that may generate thousands of potential mutations. Considering multiple sources of error at once and generating an overall quality measurement is more useful than considering several aspects individually in an all-or-nothing mutation filter. NextGENe now includes an updated mutation confidence scoring system that is compatible with all supported data types including Roche GS FLX™ and FLX Titanium, Illumina Genome Analyzers, Applied Biosystems SOLiD™ Systems, and the Ion PGM™ system.

A mutation confidence score is assigned to every mutation call, allowing the user to quickly identify the mutations that have the most supporting evidence. The software takes several variables into account in order to make an empirical estimate of the evidence that any one mutation call is a true SNP or indel. A low score indicates only that a potential mutation cannot be confidently distinguished from error based on the available data- additional sequencing may be required to confidently make a mutation call at that position. The aligned reads can be examined by double-clicking the listing in the mutation report (figure 1). The overall mutation confidence scores are not calibrated to match specific error rates. The penalty sub-scores are always directly comparable but the overall score can only be compared for mutations with similar levels of coverage.



Chr All Gene All

Page 1 of 11 First<< Previous< 1 2 3 4 5 ... >Next >>Last to Page 1

Index	CDS	Chr	Reference Nucleotide	Coverage	Score
188	7	1	T	48	13.5
189	6	1	T	15	9.1
190	1	1	C	29	5.2
191	1	1	C	27	10.9
192	3	1	A	23	3.5
193	7	1	A	27	6.0
194	2	1	C	45	5.0
195	1	1	A	18	4.0
196	1	1	T	17	1.5
197	1	1	T	4	5.4
198	1	1	A	4	4.5
199	4	1	A	30	3.1
200	6	1	T	35	0.3
201	15	1	T	23	3.4
202	20	1	T	14	0.7
203	25	1	T	60	12.3
204	12	1	G	25	3.7
205	12	1	A	25	3.7
206	4	1	T	40	4.0
207	1	1	T	21	4.2
208		1	A	18	6.2
209	6	1	A	105	16.0
210	9	1	A	28	5.0
211	9	1	G	32	6.4
212		1	A	16	4.1
213	2	1	G	17	3.1

Figure 1 – Double-clicking a cell in the mutation report will show the reads aligned at that position.

A score distribution histogram is available in order to assist with determining the best filtering settings (figure 2). Any of the penalty scores can be ignored by adjusting the mutation report settings. Analysis of sanger-verified data has shown that this scoring system can reliably increase specificity while maintaining sensitivity. As a general guideline- if coverage is high (500 to several thousand) and the data is bi-directional, mutations with scores of 5 and lower are most likely false while mutations with scores of 25 and higher are most likely true.

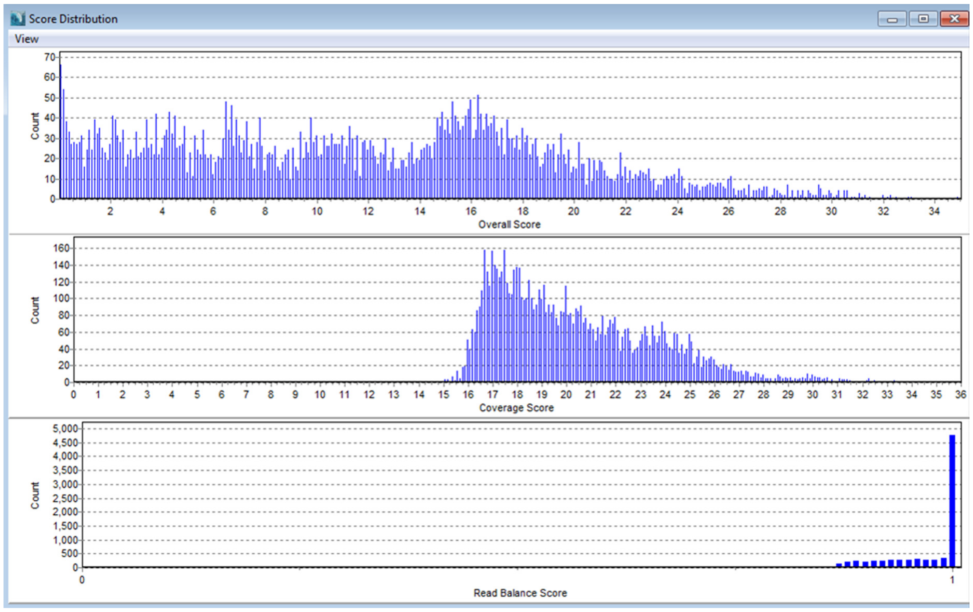


Figure 2- Score Distribution Histogram.

The mutation confidence scores are based on a log-scaled coverage score and several penalty scores, each with its own distribution. The overall score is the product of the coverage score (figure 3) and optional penalty scores which are set to 1 if there is no penalty or if they were ignored in the mutation report settings.

- The coverage score (usually between 4 and 32, but without specified limits) is based on the adjusted coverage number. Coverage from the 5' end of reads is counted as slightly better evidence than coverage from the 3' (lower basecalling quality) end of reads. If no penalties are applied the overall score is equal to the coverage score.
- The read balance score (0.3 to 1) is based on the relative number of reads aligned in each direction. Directionally biased coverage may indicate misalignment due to ambiguous sequence. This score should be disabled if the data is expected to be directionally biased, such as with certain targeted sequencing technologies.
- The allele balance score (0 to 1) is based on similar principles. The relative directional bias is considered for reads with and reads without the mutation. If coverage is present in both directions but the mutation mostly occurs in one direction it may be due to low quality sequence at the 3' end of a read or misalignment due to ambiguous sequence.
- The homopolymer score (0 to 1) penalizes indels occurring in homopolymer stretches. It is only used for data from Roche sequencers and the Ion PGM™.
- The mismatch score (0 to 1) penalizes called mutations if several occur in a small area. This does not include multi-base indels. The presence of several mismatches close together usually indicates misalignment, untrimmed adapter sequence, or poor quality data.
- The wrong allele score (0 to 1) penalizes called mutations if there are more alleles than the called mutation and the reference. This penalty score is very useful for identifying false positives due to nonspecific amplification (from PCR) or nonspecific sequence capture. If there are multiple potential mutation alleles occurring at similar frequencies, there is a larger penalty because it is more difficult to determine which is the true mutation allele.

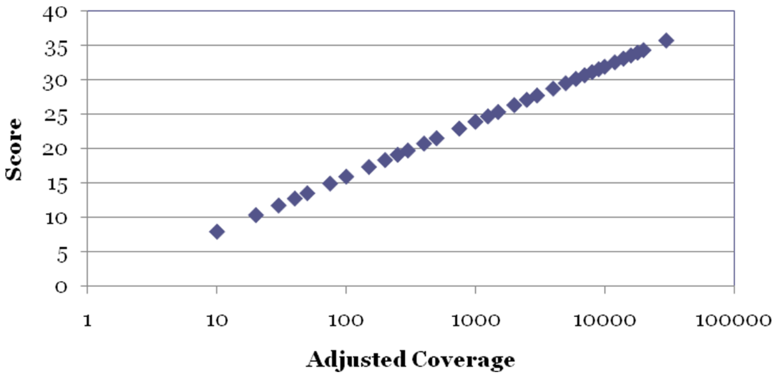
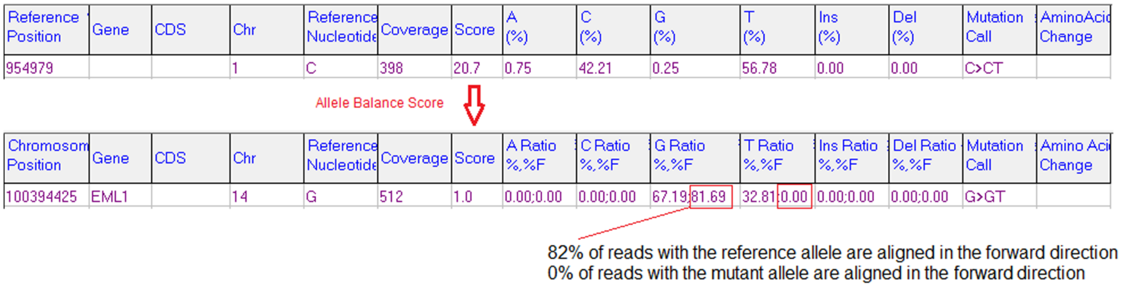


Figure 3 – Coverage score distribution.

Examples

Figure 4 shows an example of a mutation with a large allele balance penalty (low allele balance score). The top row shows the overall mutation score ignoring all penalties while the bottom row shows the same score with the allele balance penalty applied. The mutation appears to be well-supported by the evidence in the top row with over 40% of reads showing a mutation. However, if the fraction of forward reads is examined it is clear that this mutation only occurs in reads in the reverse direction.



Reference Position	Gene	CDS	Chr	Reference Nucleotide	Coverage	Score	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	Mutation Call	AminoAcid Change
954979			1	C	398	20.7	0.75	42.21	0.25	56.78	0.00	0.00	C>CT	

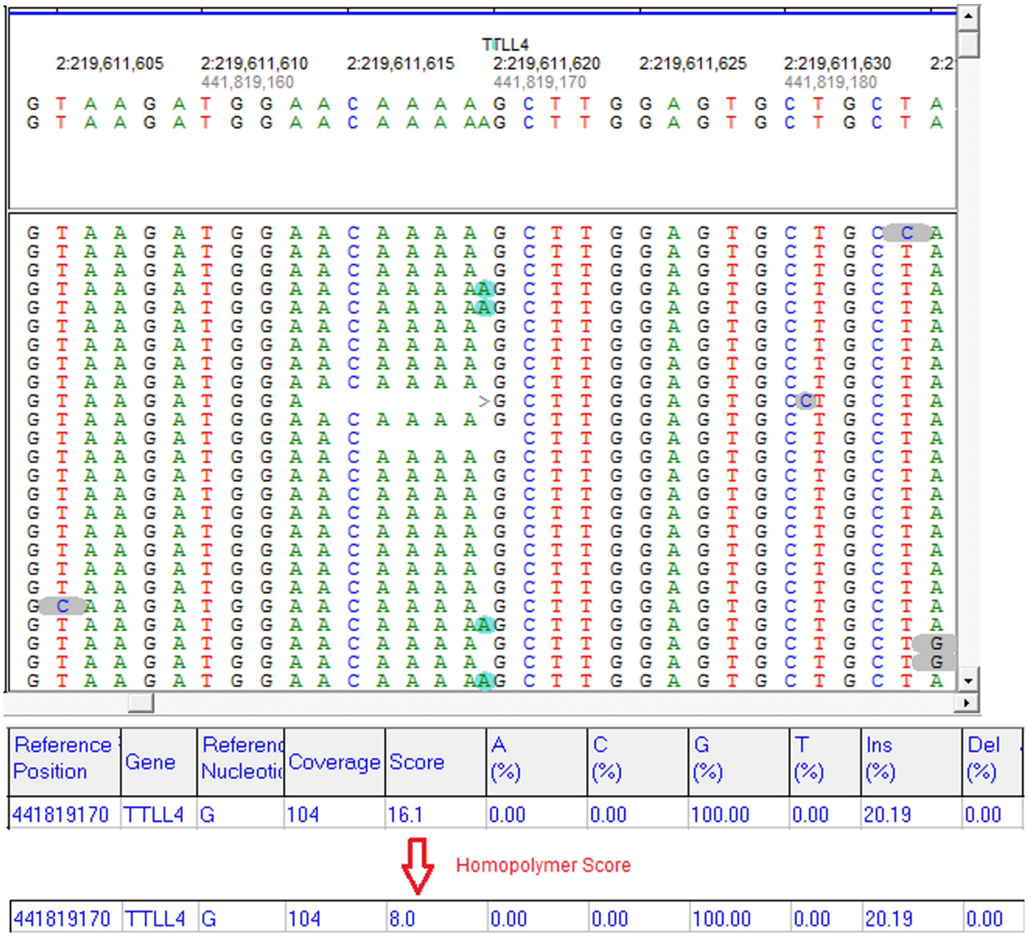
Allele Balance Score

Chromosom Position	Gene	CDS	Chr	Reference Nucleotide	Coverage	Score	A Ratio %,%F	C Ratio %,%F	G Ratio %,%F	T Ratio %,%F	Ins Ratio %,%F	Del Ratio %,%F	Mutation Call	Amino Aci Change
100394425	EML1		14	G	512	1.0	0.00;0.00	0.00;0.00	67.19;81.69	32.81;0.00	0.00;0.00	0.00;0.00	G>GT	

82% of reads with the reference allele are aligned in the forward direction
0% of reads with the mutant allele are aligned in the forward direction

Figure 4

Figure 5 shows an example of a mutation in pyrosequencing data that has been penalized by the homopolymer score. Just over 20% of reads aligned at this position showed an insertion which increased the homopolymer from 4 bases to 5 bases. This gave a homopolymer score of about 0.5, which reduced the overall score from 16.1 to 8.0.



```
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A AAG C T T G G A G T G C T G C T A

G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C C A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A AAG C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A AAG C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A            >G C T T G G A G T G CCT G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C           C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G C A A G A T G G A A C A A A A G C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A AAG C T T G G A G T G C T G C T A
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T G
G T A A G A T G G A A C A A A A G C T T G G A G T G C T G C T G
G T A A G A T G G A A C A A A AAG C T T G G A G T G C T G C T A
```

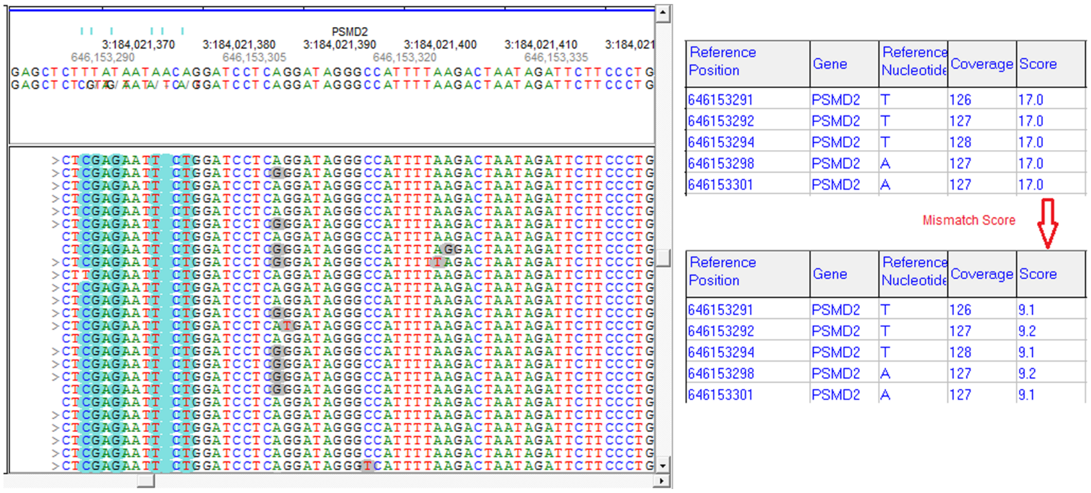
Reference Position	Gene	Referenc Nucleoti	Coverage	Score	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)
441819170	TTLL4	G	104	16.1	0.00	0.00	100.00	0.00	20.19	0.00

Homopolymer Score

441819170	TTLL4	G	104	8.0	0.00	0.00	100.00	0.00	20.19	0.00

Figure 5

Figure 6 shows an untrimmed adapter sequence that was detected at a high enough frequency to be called as a mutation. The mismatch score penalized these mutations because it is very rare for so many real mutations to occur so close together. The overall scores in this region fell from 17 to about 9.1.



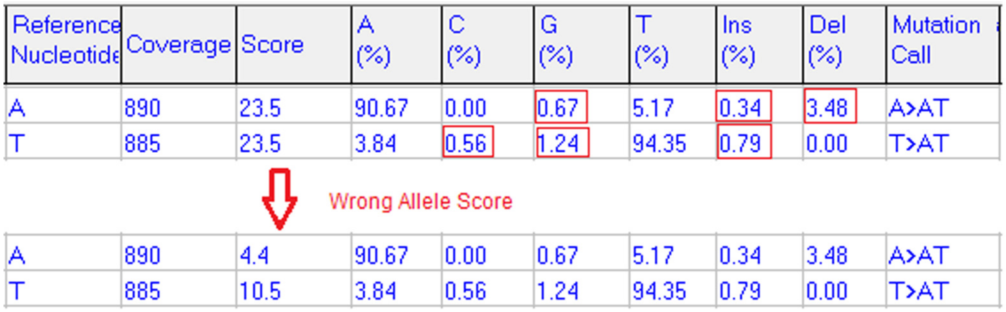
```
GAGCTCTTTATAATAACAGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
GAGCTCTCGTAG/ AATA/ TCA/ GGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG

>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
 CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
 CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTAGGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTTAGACTAATAGATTCTTCCCTG
>CTTGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCATGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
 CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
 CTCGAGAATT CTGGATCCTCGGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
 CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGCCATTTTAAGACTAATAGATTCTTCCCTG
>CTCGAGAATT CTGGATCCTCAGGATAGGGTCATTTTAAGACTAATAGATTCTTCCCTG
```

Reference Position	Gene	Reference Nucleotide	Coverage	Score
646153291	PSMD2	T	126	17.0
646153292	PSMD2	T	127	17.0
646153294	PSMD2	T	128	17.0
646153298	PSMD2	A	127	17.0
646153301	PSMD2	A	127	17.0

Reference Position	Gene	Reference Nucleotide	Coverage	Score
646153291	PSMD2	T	126	9.1
646153292	PSMD2	T	127	9.2
646153294	PSMD2	T	128	9.1
646153298	PSMD2	A	127	9.2
646153301	PSMD2	A	127	9.1

Figure 6

Figure 7 shows two mutations that were penalized for having additional alleles at a frequency close to the called mutation allele. If two non-reference alleles have similar frequencies it is difficult to call one of them as correct. Mutations that involve a double substitution (such as A>TC) are not penalized because the reference allele isn't detected.


Wrong Allele Score

Reference Nucleotide	Coverage	Score	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	Mutation Call
A	890	23.5	90.67	0.00	0.67	5.17	0.34	3.48	A>AT
T	885	23.5	3.84	0.56	1.24	94.35	0.79	0.00	T>AT

Reference Nucleotide	Coverage	Score	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	Mutation Call
A	890	4.4	90.67	0.00	0.67	5.17	0.34	3.48	A>AT
T	885	10.5	3.84	0.56	1.24	94.35	0.79	0.00	T>AT

Figure 7

Conclusion

NextGENe's updated mutation scoring system still makes analysis of large projects faster and easier by identifying the mutation calls with the most supporting evidence, but now with improved specificity. It is flexible enough to account for several different sources of error and to ignore some of them if necessary. The filtering tool allows for easy adjustment of sensitivity (allow lower scores) and specificity (allow only high scores) in mutation detection projects.