Filtering NGS Variant Calls with NextGENe® Software

January 2012

Megan Manion, Kevin LeVan, John McGuigan, Shouyong Ni, CS Jonathan Liu

Introduction

Next Generation Sequencing technologies provide the capability to sequence large genomic regions, including large gene panels, whole exome analysis and whole genome sequencing. These analyses often result in an extensive list of detected variants that cannot realistically be manually reviewed. Because of this, filtering variant calls is an essential tool to allow researchers to narrow down a large variant list to those of most importance for their study.

NextGENe includes several options for filtering variants. Filtering options include only showing mutations in coding regions, hiding synonymous mutations and filtering mutations with low variant confidence scores. Users can also choose to load a BED file, specifying regions of interest to output mutation calls for those regions only.

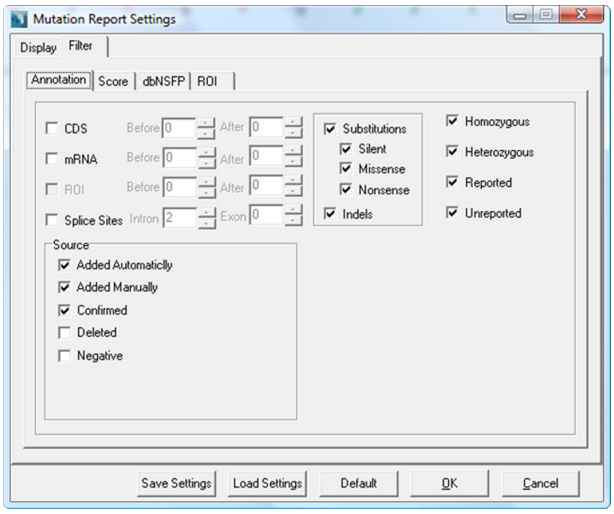
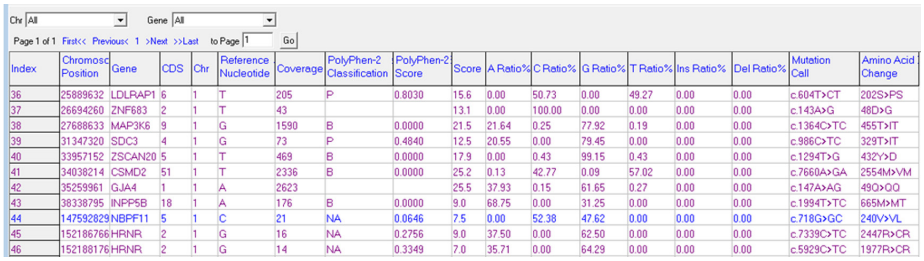


Figure 1: Mutation Filtering Options based on annotation

Procedure

1. Align Sample Data to Reference Sequence
   a. Load sample file in fasta format
   b. Load reference file as gbk, fasta, or preloaded index file provided by SoftGenetics. For whole large genomes, such as the human genome, indexed reference files are required.
   c. Adjust alignment settings as needed
2. Alignment results are automatically displayed in the NextGENe Viewer.
   Note: when using batch processing options, results are not displayed automatically. To manually load results, open the NextGENe Viewer by going to File > Open NextGENe Viewer. Then in the viewer go to File > Load Project to load results.
3 Click on the icon to open the Mutation Report Settings
4. Clicking on the "Filter" tab at the top will bring up several sub-tabs with filtering options.
5. Select appropriate options
6. Use the Show/Hide Mutation Report dropdown menu to view the filtered list of variant calls.



Chr All Gene All

Page 1 of 1 First<< Previous< 1 >Next >>Last to Page 1 Go

Index	Chromosc Position	Gene	CDS	Chr	Reference Nucleotide	Coverage	PolyPhen-2 Classification	PolyPhen-2 Score	Score	A Ratio%	C Ratio%	G Ratio%	T Ratio%	Ins Ratio%	Del Ratio%	Mutation Call	Amino Acid Change
36	25889632	LDLRAP1	6	1	T	205	P	0.8030	15.6	0.00	50.73	0.00	49.27	0.00	0.00	c.604T>CT	202S>PS
37	26694260	ZNF683	2	1	T	43			13.1	0.00	100.00	0.00	0.00	0.00	0.00	c.143A>G	48D>G
38	27688633	MAP3K6	9	1	G	1590	B	0.0000	21.5	21.64	0.25	77.92	0.19	0.00	0.00	c.1364C>TC	455T>IT
39	31347320	SDC3	4	1	G	73	P	0.4840	12.5	20.55	0.00	79.45	0.00	0.00	0.00	c.986C>TC	329T>IT
40	33957152	ZSCAN20	5	1	T	469	B	0.0000	17.9	0.00	0.43	99.15	0.43	0.00	0.00	c.1294T>G	432Y>D
41	34038214	CSMD2	51	1	T	2336	B	0.0000	25.2	0.13	42.77	0.09	57.02	0.00	0.00	c.7660A>GA	2554M>VM
42	35259961	GJA4	1	1	A	2623			25.5	37.93	0.15	61.65	0.27	0.00	0.00	c.147A>AG	49Q>QQ
43	38338795	INPP5B	18	1	A	176	B	0.0000	9.0	68.75	0.00	31.25	0.00	0.00	0.00	c.1994T>TC	665M>MT
44	147592829	NBPF11	5	1	C	21	NA	0.0646	7.5	0.00	52.38	47.62	0.00	0.00	0.00	c.718G>GC	240V>VL
45	152186766	HRNR	2	1	G	16	NA	0.2756	9.0	37.50	0.00	62.50	0.00	0.00	0.00	c.7339C>TC	2447R>CR
46	152188176	HRNR	2	1	G	14	NA	0.3349	7.0	35.71	0.00	64.29	0.00	0.00	0.00	c.5929C>TC	1977R>CR

Figure 2: Mutation Report

Discussion

NextGENe provides a variety of options for filtering variant calls. Annotation information can be used to show only mutations found in certain regions such as in coding regions, mRNA regions or splice sites. Users can select to show only SNPs or only Indels, or for SNP calls, to filter based on whether the mutation is silent, missense or nonsense. Also, reported variants, variants that are included in the dbSNP database, can be hidden. See Figure 1 above.

A mutation confidence score is provided for all mutation calls. Users can also choose to filter based on these scores to remove less confident calls. The overall confidence score is based on several sub-scores so users can filter based on the overall score and/or based on individual sub-scores. (See Variant Call Confidence Scoring.)

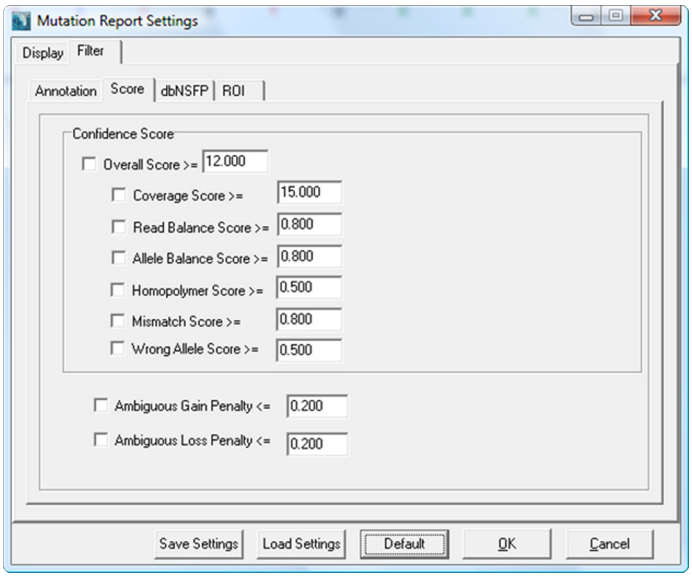


Figure 3: Mutation Filtering Options based on mutation confidence scores

Information from the dbNSFP database can be imported to NextGENe to provide mutation prediction scores and 1000 Genomes frequencies (1). The database information is imported using the Convert dbNSFP Files Tool . Users can then use the dbNSFP scores to filter for only mutations with specified scores.

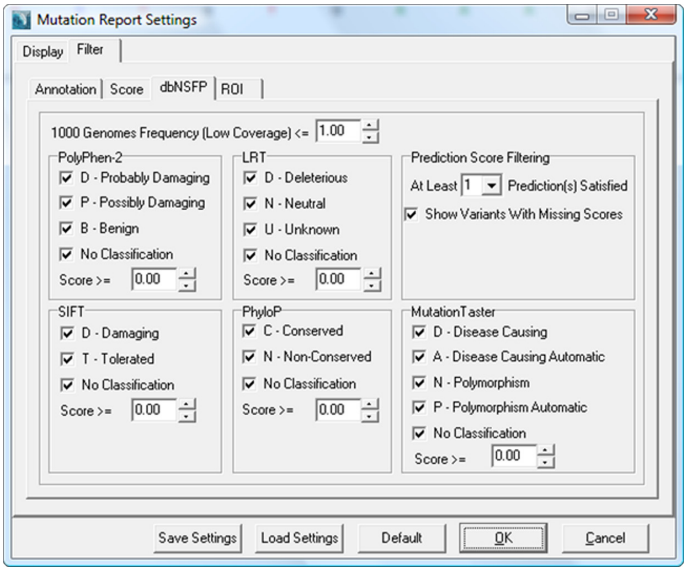


Figure 4: Mutation Filtering Options based on dbNSFP scores

Lastly, a file in BED or text format, which indicates specific regions of interest can be uploaded to include only mutations found in the designated regions. Users can also choose to upload a BED or text format file listing specific regions to be excluded from the analysis.

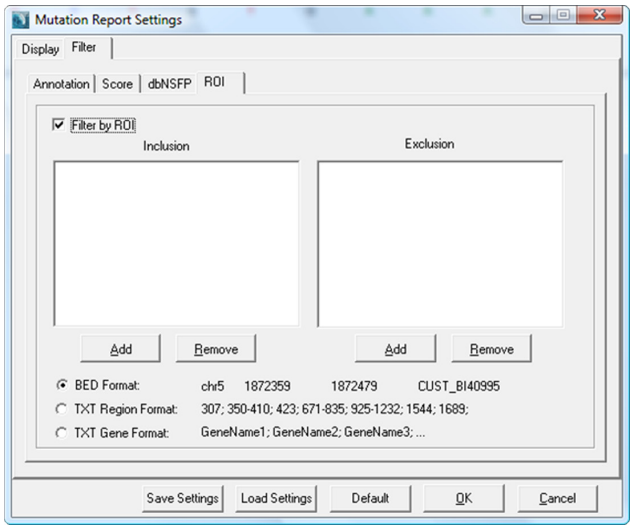


Figure 5: Mutation Filtering Options based on ROI files

NextGENe can be used for analysis of DNA sequencing data from Illumina GA, MiSeq & HiSeq systems, Roche/ 454 GS FLX, FLX Titanium and Junior, Applied Biosystems' SOLiD System, Ion PGM and PacBio platforms. Modules for various applications with unique technologies and tools are all included in a single package. Results are displayed in the NextGENe Viewer, providing a high level of visualization not available in other commercial programs like Lasergene's SeqMan Pro, CLC Bio & DNASTAR's NGEN or in open-source tools like TopHat, Bowtie & BWA.

References

1. Liu, X., Jian, X. and Boerwinkle, E. (2011), dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Human Mutation, 32: 894–899. doi: 10.1002/humu.21517