strandngs

SNP Detection and Prioritization

Tutorial

Analyze | Visualize | Annotate | Discover

strand
New Generation Healthcare

SNPs are the most common genetic variation in human genome. We humans have at least one SNP every 300 base pairs in DNA. SNPs are useful in tracking the inheritance of diseases within families. The presence of SNPs affect gene function if they are in the regulatory regions, exons, and splice sites of the gene and can be used to predict risk of developing a particular disease and assess response and susceptibility to certain toxins or drugs or used as biological biomarkers.

The **SNP Analysis feature** in Strand NGS identifies the variants (SNPs/ MNPs/ InDels) in the sample by comparing the aligned reads against the reference genome. The SNP Detection algorithm in Strand NGS is based on MAQ (refer Strand NGS manual). The algorithm compares the bases present on aligned reads against the reference, at each position. To make a call, SNP caller takes into account the bases that are sufficiently covered, have good base quality and good mapping quality.

The SNP Analysis workflow in Strand NGS consists of running the **pre-processing steps** of split read realignment, local realignment, and base quality recalibration, and then filtering the reads to retain only those with high mapping and base qualities. The SNPs are called using **SNP Detection** option in the workflow that outputs a master list SNP multi sample report and two lists called Single Base Variant **(SBV)** and a Multi Base Variant **(MBV)** list. One can then find significant SNPs from the multi sample report and carry out **SNP Effect Analysis** on the filtered SNPs.

The SNPs can be visually analysed in the Genome Browser, or by using Variant Support View **(VSV)** feature in the tool. One can also run clustering to see if among the significant SNPs there are any patterns correlating the clusters with the experimental grouping information.

This document illustrates the key analysis steps for detection of SNPs in Strand NGS:

1. Display and enumerate the SNP calling steps
2. SNP Detection output
3. Visual verification of the SNPs/InDels
4. Prioritizing the SNPs/ InDels

Display and enumerate SNP calling steps

The SNP Detection feature can be invoked on data from exome, whole genome, targeted experiments, and data from transcriptomic studies. The data from exome, whole genome, and targeted experiments can be loaded into Strand NGS as a DNA-Seq experiment where as data from transcriptome samples is loaded into Strand NGS as an RNA-Seq experiment.

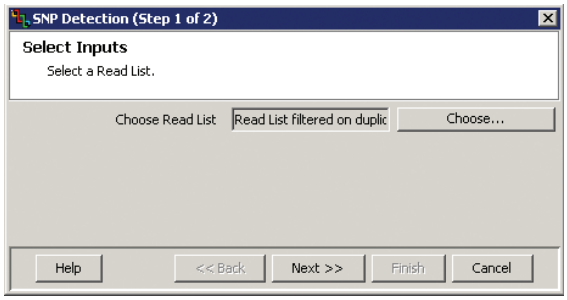
The **Analysis** section in the Workflow pane contains the option for carrying out SNP Detection. Clicking on this workflow step will enable the user to detect SNPs in a chosen read list. The chosen read list could be the list of all aligned reads, or any other read list generated after filtering or SNP pre processing steps.

SNP Detection wizard in Strand NGS is multi-tabbed. The parameters for SNP Detection, SNP Filtering and some advanced setting to be used in a given SNP Detection run are present in the SNP Detection wizard.

The steps for SNP Detection are as follows:

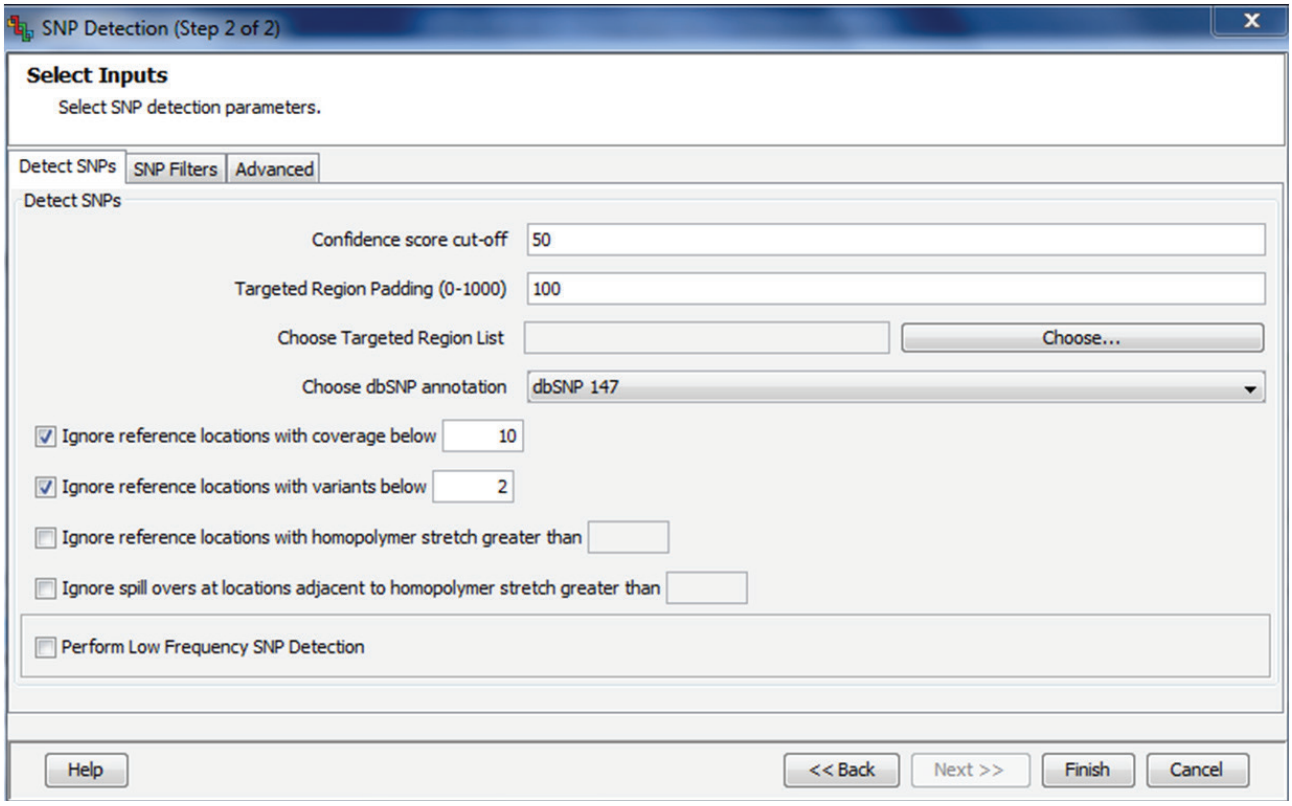
Step 1

SNP Detection input steps.

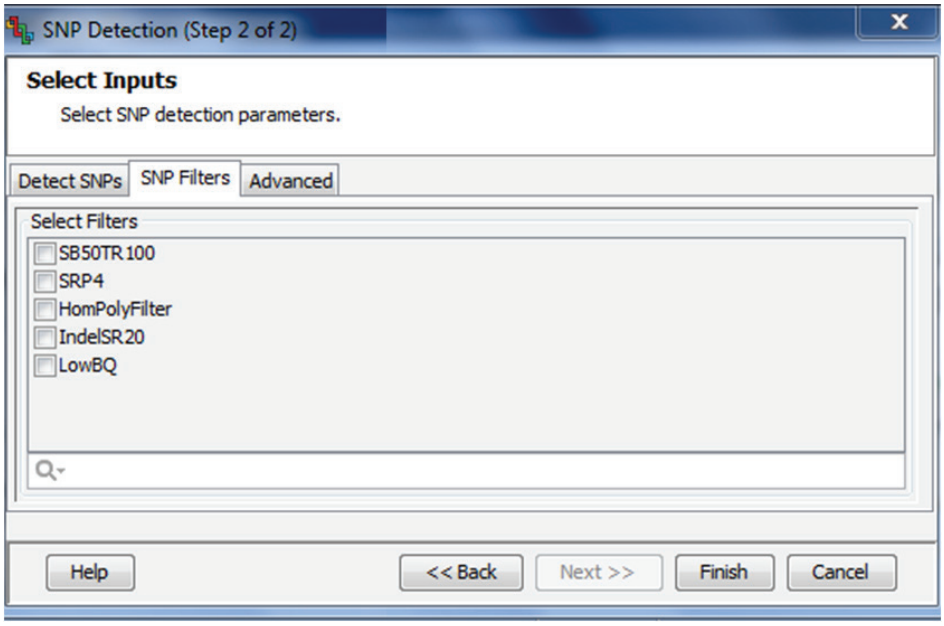


Step 2

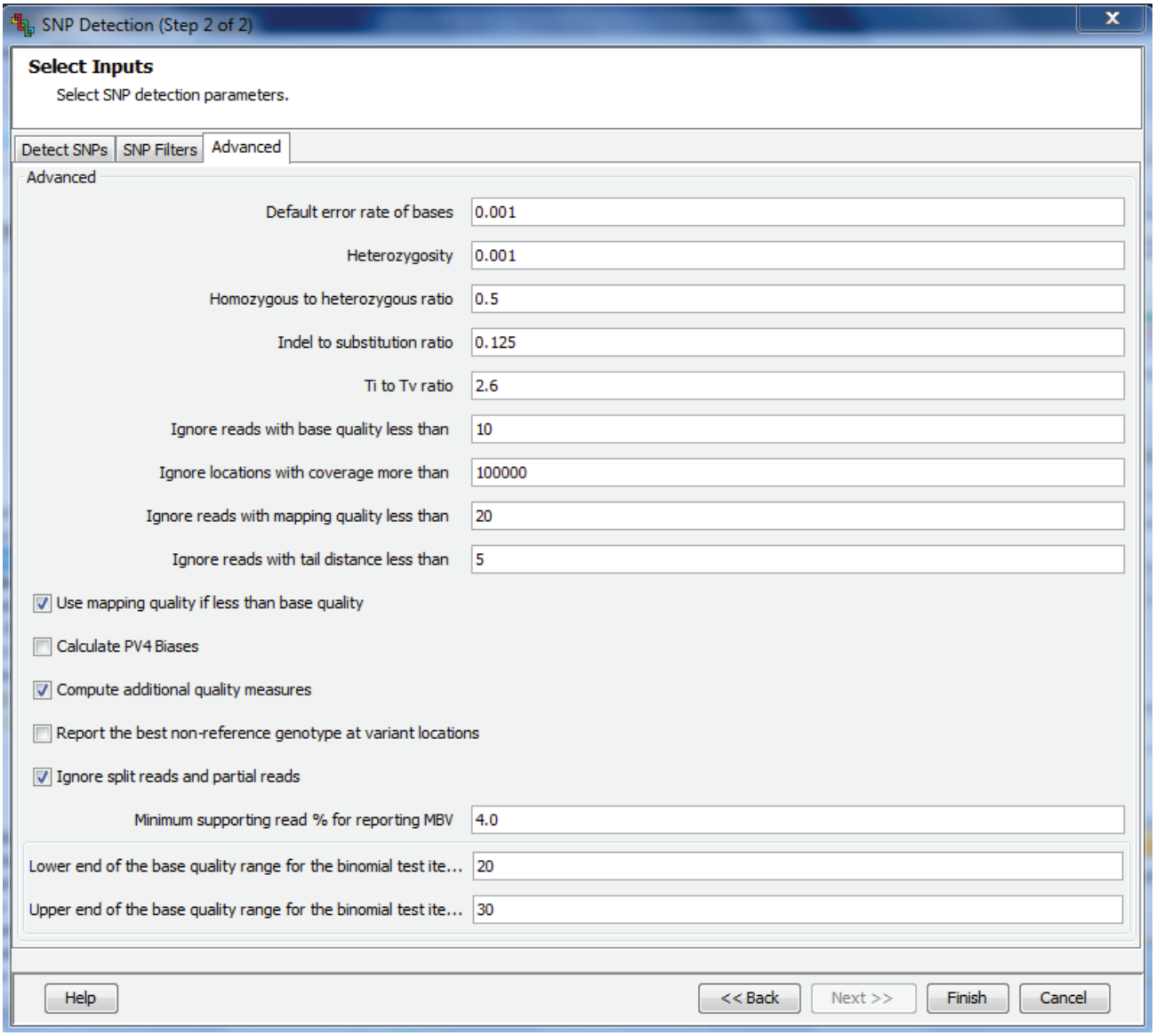
a. **Detect SNPs- SNP Detection input steps.**



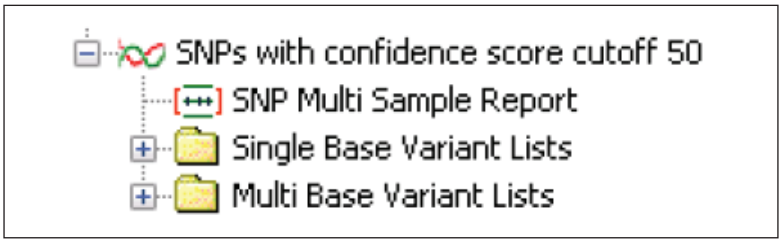
b. **SNP Filters- SNP Detection input Steps.**



c. **SNP Detection input steps.**



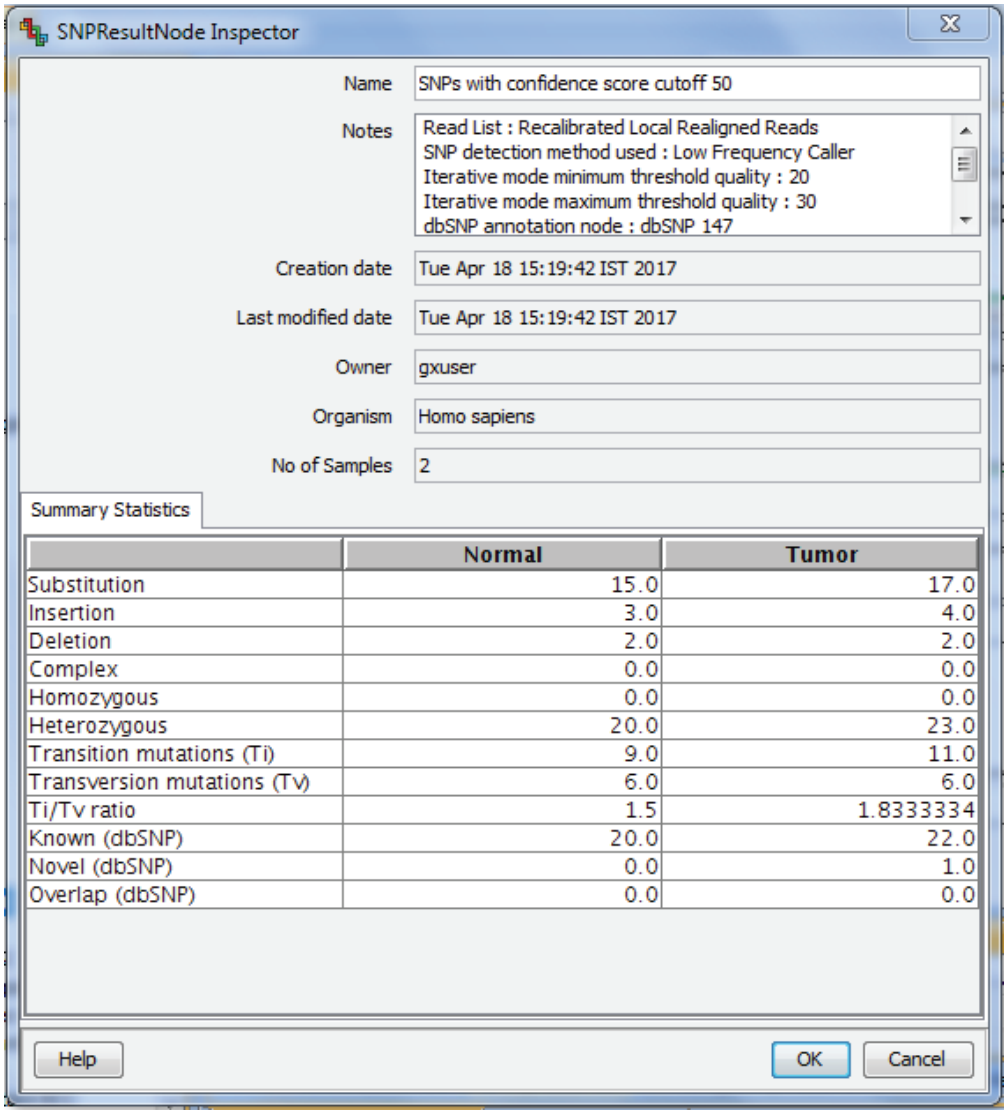
SNP Detection output: Once the SNP Detection step is over, an output list like the one shown below appears in the navigator pane.



Output objects in the navigator after SNP Detection

It consists of a main SNP object (SNPs list with confidence score cut-off 50), SNP Multi Sample Report, and two folders named Single Base Variant (SBV) Lists and Multi Base Variant (MBV) lists; each of which includes one region list per sample.

a. **Inspect SNP results:** One can right click on the SNPs list with confidence score cut-off 50 and have a glance at the summary statistics of number of substitutions, Insertions, deletions, Ti/Tv ratio etc. as shown below.



	Normal	Tumor
Substitution	15.0	17.0
Insertion	3.0	4.0
Deletion	2.0	2.0
Complex	0.0	0.0
Homozygous	0.0	0.0
Heterozygous	20.0	23.0
Transition mutations (Ti)	9.0	11.0
Transversion mutations (Tv)	6.0	6.0
Ti/Tv ratio	1.5	1.8333334
Known (dbSNP)	20.0	22.0
Novel (dbSNP)	0.0	1.0
Overlap (dbSNP)	0.0	0.0

SNP Result Node Inspector

b. **SNP Multi-Sample Report or Multi-Sample Variant Allele List (MSVAL):** Is a combined report of SNPs detected in all the samples in the experiment. Each row in the report contains information about variant allele occurring in one or more samples at that particular location and the sample count column shows the number of samples in which the variant allele is present in the SNP call. SNP Multi-Sample Report is used for all the subsequent analysis steps such as find Significant SNPs, SNP Effect Analysis, and Cluster Regions.

RegionListNode Inspector

Name: SNP Multi Sample Report
Organism: Homo sapiens
Notes:
Creation date: Sun Jun 25 21:58:43 IST 2017
Last modified date: Sun Jun 25 21:58:43 IST 2017
Owner: gxuser
Number of Regions: 4,386

Regions | Histogram | Summary Statistics | Annotations

Chromos...	Start	End	Reference	Variant A...	Variant T...	Sample ...	Experime...	Supporti...	Variant R...	Total Re...	Percent S...	SNP Call ...	Score (P0...	Zygosity (P0...	Supporti...	Variant R...	Total Re...	Percent S
chr9	98011602	98011602	T	G	Substitution	2	0.083333	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98209509	98209509	C	T	Substitution	1	0.041667	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98209594	98209594	G	A	Substitution	9	0.583333	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98211572	98211572	T	A	Substitution	2	0.083333	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98215664	98215664	G	T	Substitution	1	0.041667	24.00000	28.00000	25	12.00000	ref			19.44444	25.00000	108	41.5343
chr9	98215671	98215671	T	C	Substitution	2	0.083333	43.65079	53.17460	126	4.992785	C/T	1000.000	Heterozygous	31.32530	44.17671	249	26.4133
chr9	98215678	98215678	C	T	Substitution	1	0.041667	5.440414	8.031088	386	3.010116	C/T	146.9523	Heterozygous	2.322738	3.789731	818	8.9177
chr9	98215692	98215692	G	A	Substitution	2	0.083333	20.79208	24.25743	404	6.435644	A/G	1000.000	Heterozygous	15.44910	18.68263	835	7.2487
chr9	98215697	98215697	C	T	Substitution	1	0.041667	5.319149	6.117021	376	5.319149	C/T	160.3773	Heterozygous	3.283174	4.651163	731	4.7879
chr9	98215701	98215701	T	C	Substitution	2	0.083333	32.97587	35.92493	373	2.223239	C/T	1000.000	Heterozygous	35.77465	39.71831	710	2.8612
chr9	98215704	98215704	A	G	Substitution	2	0.083333	24.39678	24.93298	373	6.192735	A/G	1000.000	Heterozygous	22.03148	23.46209	699	10.928
chr9	98215708	98215708	G	A	Substitution	2	0.083333	3.927492	4.229608	331	12.22403	A/G	145.2626	Heterozygous	6.500000	6.500000	600	26.410
chr9	98215715	98215715	T	A	Substitution	1	0.083333	100.0000	100.0000	7	0.000000	lc			92.59259	92.59259	27	7.4074
chr9	98218474	98218474	G	GC	Insertion	3	0.166667	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98218624	98218624	G	T	Substitution	1	0.041667	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98220322	98220322	A	C	Substitution	2	0.083333	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98221861	98221861	T	C	Substitution	3	0.166667	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98224360	98224360	C	G	Substitution	3	0.166667	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000
chr9	98229389	98229389	C	G	Substitution	3	0.166667	0.000000	0.000000	0	0.000000	lc			0.000000	0.000000	0	0.0000

c. **Single Base Variant Lists:** This folder contains one region list per sample giving the information about location and type of variant called along with the score associated with the variant. The supporting attributes are listed in the report; like percent of supporting reads for the variant, ATGC composition, total reads overlapping the variant and strand bias etc. Some of these attributes are assigned by the SNP caller like Zygosity, SNP called and score while other attributes in the list are the qualifying attributes like strand bias, total reads, percent supporting reads, and percent variant reads.

RegionListNode Inspector

Name: SBV_P000102-S131400060-EYE_S4
Organism: Homo sapiens
Notes:
Creation date: Sun Jun 25 21:58:48 IST 2017
Last modified date: Sun Jun 25 21:58:48 IST 2017
Owner: gxuser
Number of Regions: 301

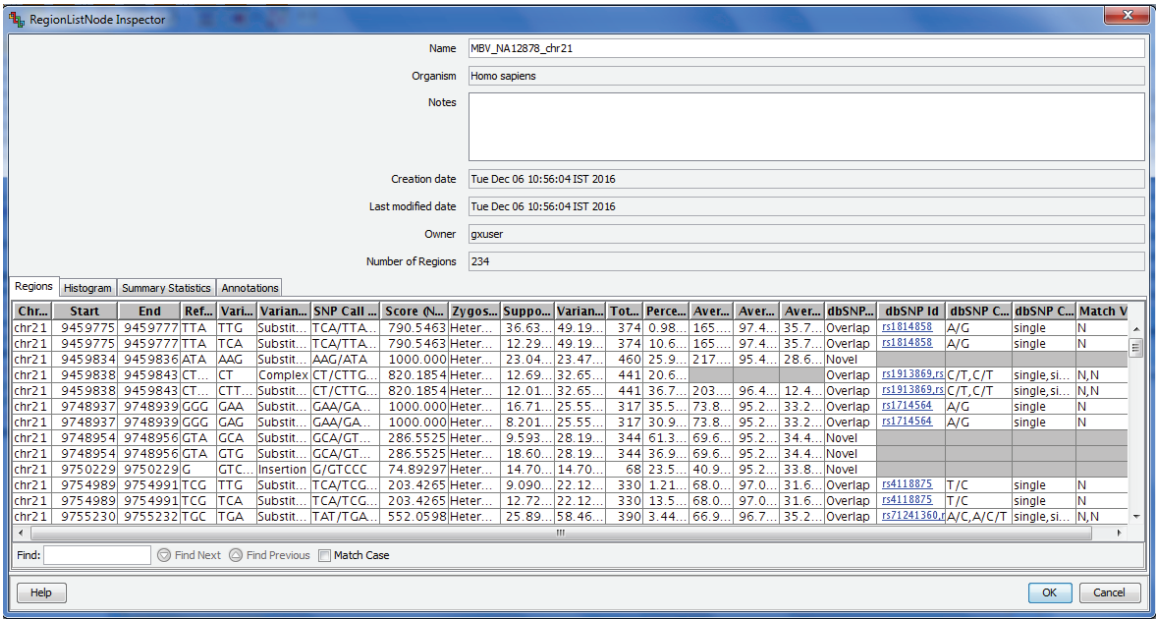
Regions | Histogram | Summary Statistics | Annotations

Chromos...	Start	End	Reference	Variant A...	Variant T...	SNP Call ...	Score (P0...	Zygosity...	Supporti...	Variant R...	Total Re...	Percent S...	As (P000...	Cs (P000...	Gs (P000...	Ts (P000...	=s (P000...	Part Of ...
chr7	6022378	6022378	T	G	Substitution	A/G	441.6179	Heterozy...	50.00000	88.88889	18	11.11111	6	0	9	2	1	Yes
chr7	6022386	6022386	T	A	Substitution	A/G	292.3228	Heterozy...	31.25000	81.25000	16	22.50000	5	0	8	3	0	No
chr7	6022386	6022386	T	G	Substitution	A/G	292.3228	Heterozy...	50.00000	81.25000	16	37.50000	5	0	8	3	0	No
chr7	55209988	55209988	G	A	Substitution	A	519.5365	Homozyg...	100.0000	100.0000	13	0.000000	13	0	0	0	0	No
chr7	55209992	55209992	G	-	Deletion	-	510.9733	Homozyg...	100.0000	100.0000	13	0.000000	0	0	0	0	13	No
chr7	148512...	148512...	C	-	Deletion	-/C	207.1682	Heterozy...	36.84211	36.84211	19	63.15789	0	12	0	0	7	No
chr7	148512...	148512...	T	A	Substitution	A/T	209.5816	Heterozy...	36.84211	36.84211	19	63.15789	7	0	0	12	0	No
chr8	30954249	30954249	T	C	Substitution	C/T	671.1508	Heterozy...	72.00000	72.00000	25	12.00000	0	18	0	7	0	No
chr8	30954270	30954270	G	T	Substitution	T	738.2117	Homozyg...	95.00000	95.00000	20	4.736842	0	0	1	19	0	No
chr8	31007828	31007828	G	A	Substitution	A/G	184.0016	Heterozy...	40.00000	40.00000	15	0.000000	6	0	9	0	0	No
chr8	31007837	31007838	-	C	Insertion	-/C	211.1817	Heterozy...	46.66667	46.66667	15	0.000000	0	7	0	0	8	No
chr8	90976627	90976627	G	-	Deletion	-	390.6703	Homozyg...	100.0000	100.0000	10	0.000000	0	0	0	0	10	Yes
chr8	90976628	90976628	C	-	Deletion	-	390.6703	Homozyg...	100.0000	100.0000	10	0.000000	0	0	0	0	10	Yes
chr8	90976629	90976629	T	-	Deletion	-	390.6703	Homozyg...	100.0000	100.0000	10	0.000000	0	0	0	0	10	Yes
chr8	90976630	90976630	T	-	Deletion	-	390.6703	Homozyg...	100.0000	100.0000	10	0.000000	0	0	0	0	10	Yes
chr8	90976631	90976631	C	-	Deletion	-	390.6703	Homozyg...	100.0000	100.0000	10	0.000000	0	0	0	0	10	Yes

Find: | Find Next | Find Previous | Match Case

Single Base Variant Region List

d. **Multi Base Variant Lists:** This folder contains one region list per sample giving the details of MNPs, multi-base insertions and deletions.



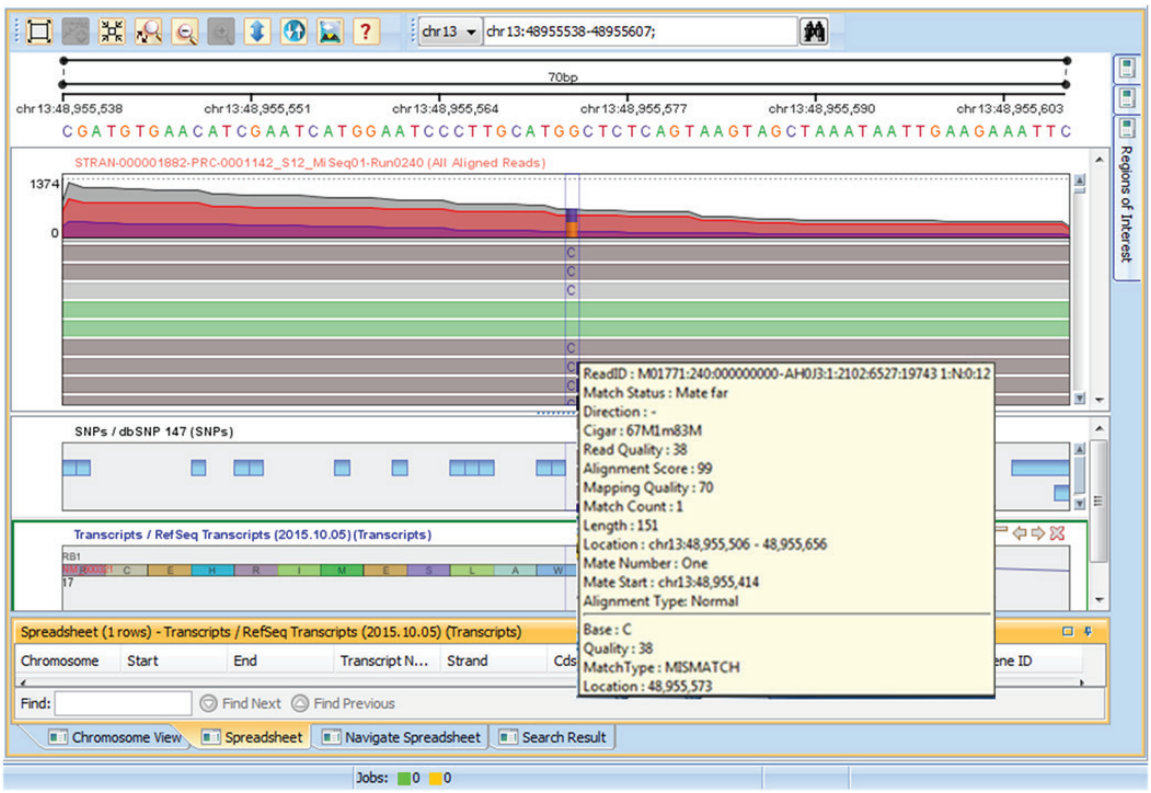
Chr...	Start	End	Ref...	Vari...	Varian...	SNP Call ...	Score (N...	Zygos...	Suppo...	Varian...	Tot...	Perce...	Aver...	Aver...	Aver...	dbSNP...	dbSNP Id	dbSNP C...	dbSNP C...	Match V
chr21	9459775	9459777	TTA	TTG	Substit...	TCA/TTA...	790.5463	Heter...	36.63...	49.19...	374	0.98...	165....	97.4...	35.7...	Overlap	rs1814858	A/G	single	N
chr21	9459775	9459777	TTA	TCA	Substit...	TCA/TTA...	790.5463	Heter...	12.29...	49.19...	374	10.6...	165....	97.4...	35.7...	Overlap	rs1814858	A/G	single	N
chr21	9459834	9459836	ATA	AAG	Substit...	AAG/ATA	1000.000	Heter...	23.04...	23.47...	460	25.9...	217....	95.4...	28.6...	Novel				
chr21	9459838	9459843	CT...	CT	Complex	CT/CTTG...	820.1854	Heter...	12.69...	32.65...	441	20.6...				Overlap	rs1913869,rs	C/T,C/T	single,si...	N,N
chr21	9459838	9459843	CT...	CTT...	Substit...	CT/CTTG...	820.1854	Heter...	12.01...	32.65...	441	36.7...	203....	96.4...	12.4...	Overlap	rs1913869,rs	C/T,C/T	single,si...	N,N
chr21	9748937	9748939	GGG	GAA	Substit...	GAA/GA...	1000.000	Heter...	16.71...	25.55...	317	35.5...	73.8...	95.2...	33.2...	Overlap	rs1714564	A/G	single	N
chr21	9748937	9748939	GGG	GAG	Substit...	GAA/GA...	1000.000	Heter...	8.201...	25.55...	317	30.9...	73.8...	95.2...	33.2...	Overlap	rs1714564	A/G	single	N
chr21	9748954	9748956	GTA	GCA	Substit...	GCA/GT...	286.5525	Heter...	9.593...	28.19...	344	61.3...	69.6...	95.2...	34.4...	Novel				
chr21	9748954	9748956	GTA	GTG	Substit...	GCA/GT...	286.5525	Heter...	18.60...	28.19...	344	36.9...	69.6...	95.2...	34.4...	Novel				
chr21	9750229	9750229	G	GTC...	Insertion	G/GTCCC	74.89297	Heter...	14.70...	14.70...	68	23.5...	40.9...	95.2...	33.8...	Novel				
chr21	9754989	9754991	TCG	TTG	Substit...	TCA/TCG...	203.4265	Heter...	9.090...	22.12...	330	1.21...	68.0...	97.0...	31.6...	Overlap	rs4118875	T/C	single	N
chr21	9754989	9754991	TCG	TCA	Substit...	TCA/TCG...	203.4265	Heter...	12.72...	22.12...	330	13.5...	68.0...	97.0...	31.6...	Overlap	rs4118875	T/C	single	N
chr21	9755230	9755232	TGC	TGA	Substit...	TAT/TGA...	552.0598	Heter...	25.89...	58.46...	390	3.44...	66.9...	96.7...	35.2...	Overlap	rs71241360,r	A/C,A/C/T	single,si...	N,N

Multi Base Variant Region List

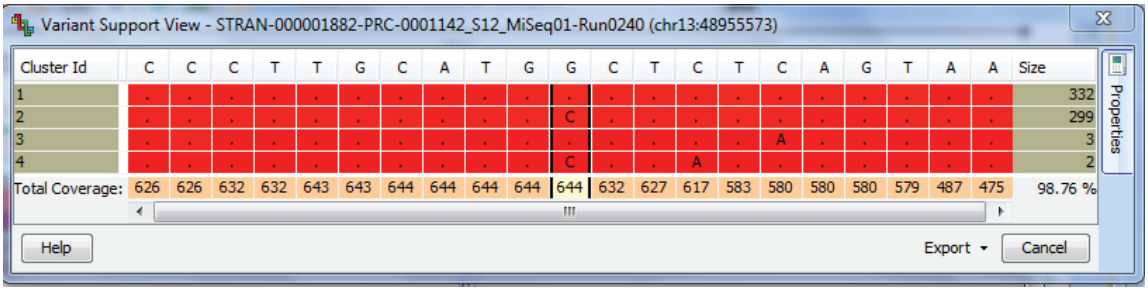
If multiple SNPs appear contiguously, they are split up as single SNPs in the single base variant reports, but are put together into a single MNP in the Multi-Base Variant Reports.

Visual verification of the SNPs/InDels

1. **SNP visualization in the Genome Browser:** Dragging and dropping either the SNP result object, or the individual sample SNP result region list into the Genome Browser will show all the SNPs and Indels in the Genome Browser. One can navigate from one SNP to next in the Genome Browser very easily using the navigator options. One can also look for options available on the track and the mismatch histogram shown in the coverage profile in the read list track. This can be used to qualify and identify loci having SNPs. This mismatch histogram will show what part of the total reads is reference and what part is variant. In order to look at the reads supporting variants one could look at the attributes of reads in the reads/ bases in the read track by holding the mouse over particular read/ base (refer below picture).



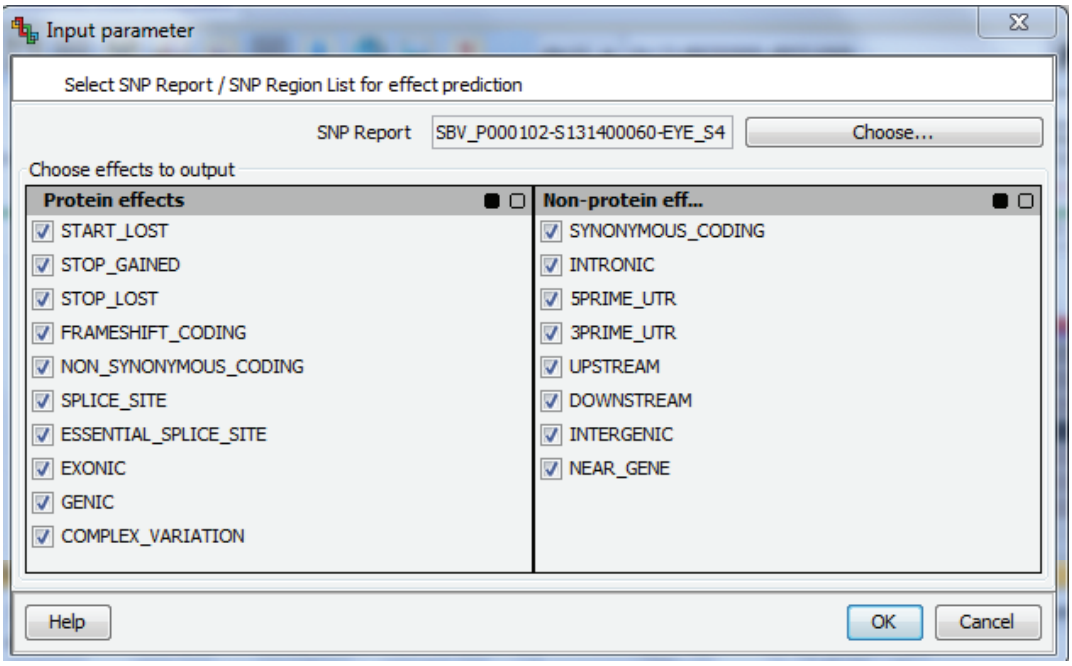
2. **Variant Support View (VSV):** The other visualization is called the "Variant Support View" and can be launched by right-clicking on the SNP results object (SBV, MBV, or Muti-Sample Report) in the navigator pane or from within the Genome Browser via the right-click menu on the read list track. The variant support view is very useful when the coverage for certain regions is very deep and it is not possible to look at all the reads in the Genome Browser. Variant support view takes all the reads and compresses the neighborhood region 10 bases on each side of the query base to a table containing few rows where one can see the neighboring bases and have a feel for overall quality of the region around the base. Therefore VSV is very useful for verifying heterozygous SNPs visually with more confidence.



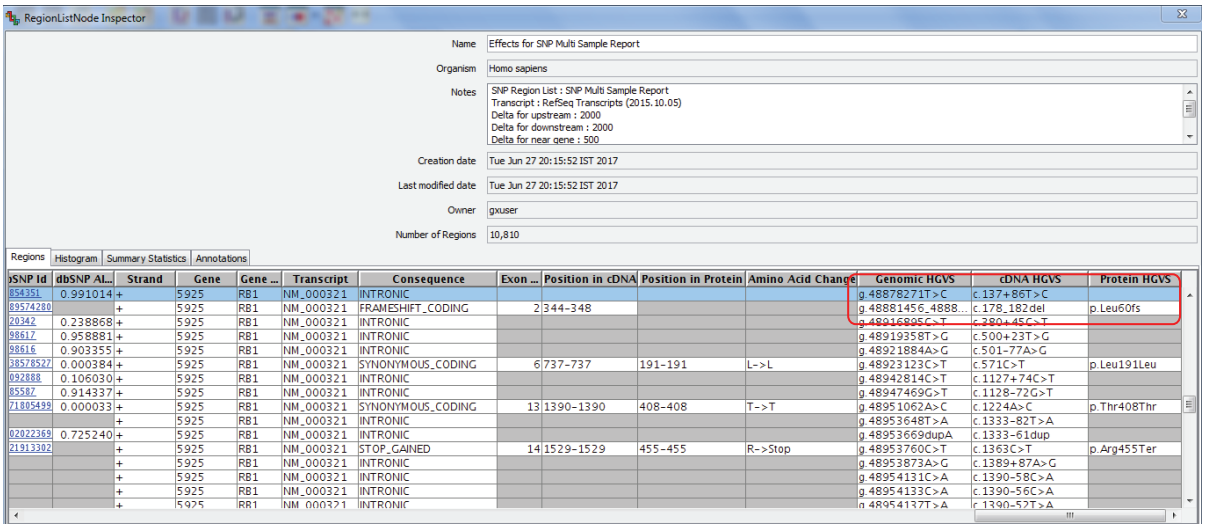
Prioritizing the SNPs/ InDels

The SNP Detection workflow may output hundreds and thousands of SNPs from a single sample and the numbers go up if there are many samples in an experiment. So ranking or rating these SNPs becomes very important. Described below are the various ways in which Strand NGS is able to prioritize the SNPs after detection and extract relevant SNPs.

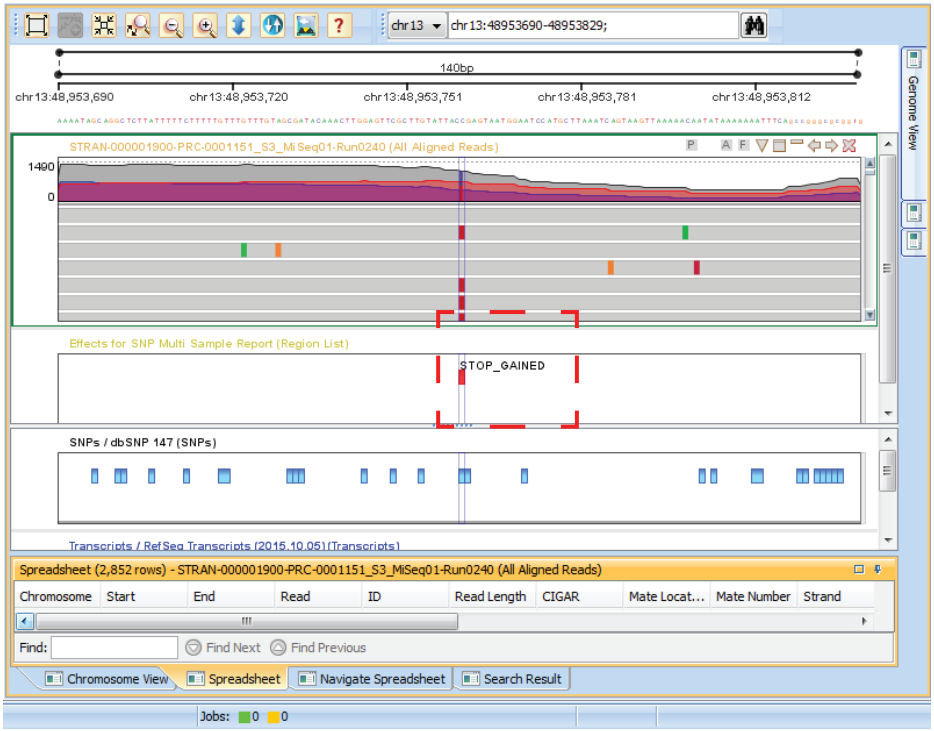
1. **SNP Effect Analysis:** This particular step in workflow helps find the SNPs which overlap a protein coding region of a gene and compute the effect it might have on the gene function. The SNP Effect Analysis finds not just the biological consequences of SNP but it can compute non protein coding effects too; like effect of presence of SNPs in 5' UTR etc. The list is shown below:



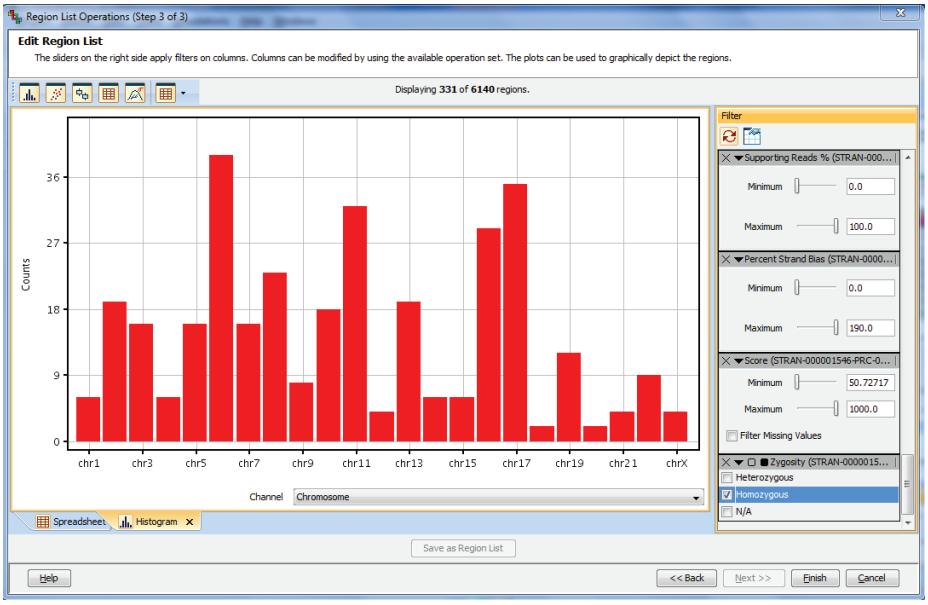
To execute this step, the desired transcript annotations should have been chosen during the experiment creation step (Ensembl, RefSeq or UCSC transcript annotations), which could be downloaded through Annotations Manager.
The results of the analysis are added as additional columns to the SNP Detection list as genes overlapping the SNPs, the transcript and the consequence column. The list also contains the columns indicating the consequences in standard HGVS nomenclature.



The resulting region list of SNP Effect Analysis can be dragged and dropped into the Genome Browser and one can view the results as shown in the figure below.



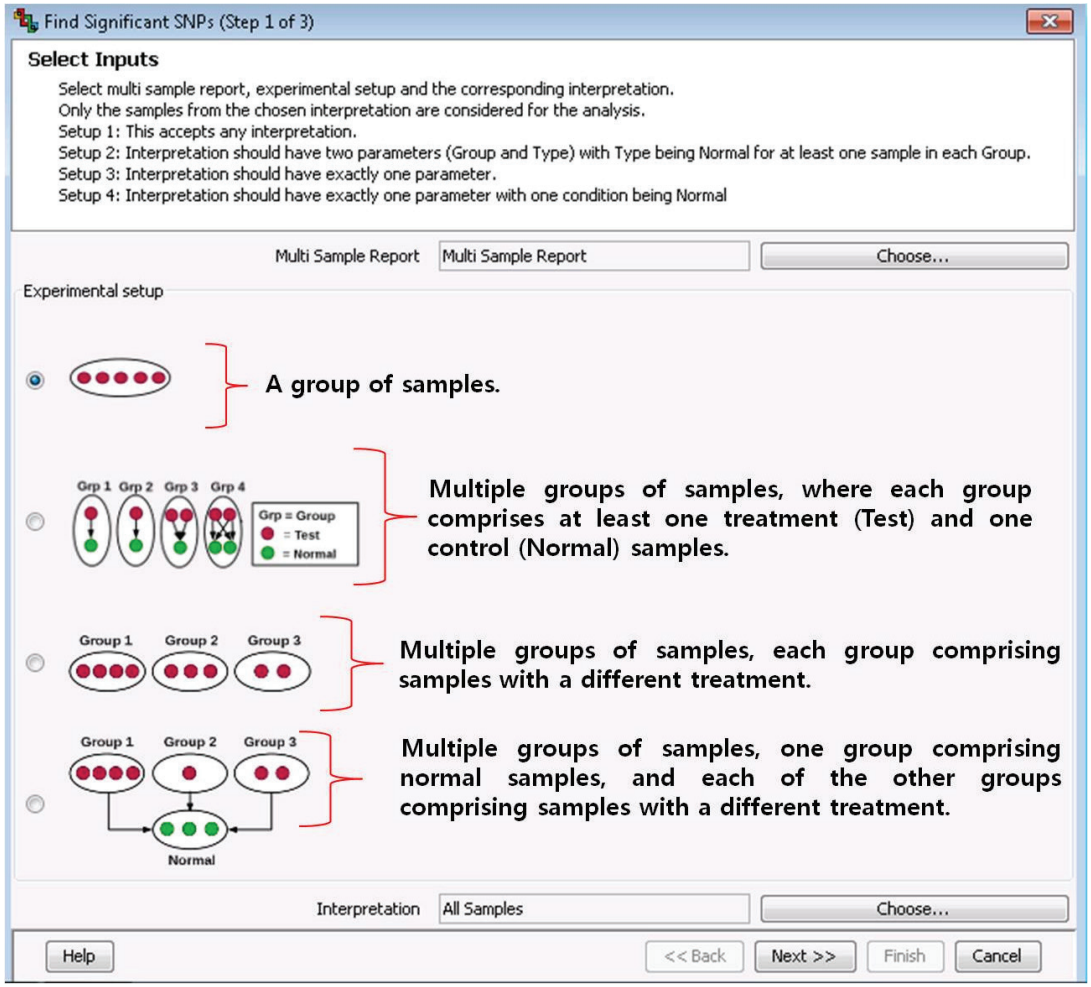
2. **Region List Operations:** The SNP list arising from SNP Detection can be filtered using region lists operations option in the tool. Even the SNP effect analysis report can further be filtered to like finding just stop gain or start loss events etc. The Region List Operations allows one to look at the data in the form of histograms or scatter plots to get a better sense of the SNPs or the data that we get from SNP Detection workflow. Using the Region List Operations one can select for the SNPs in the samples based on the type of consequence, % supporting reads, and coverage etc.

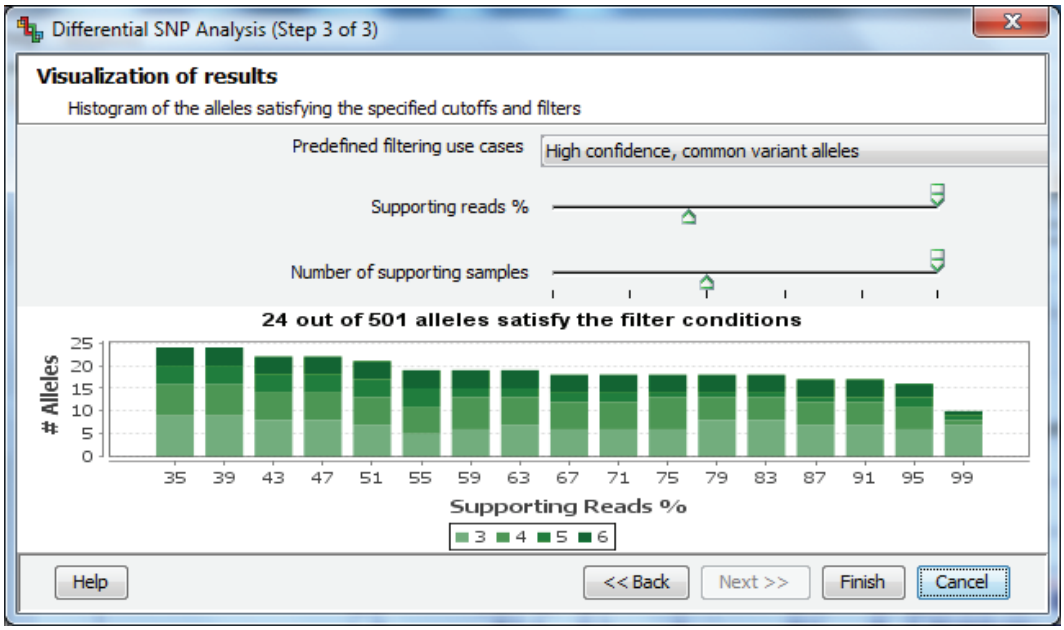


3. **Validation and Prioritization based on External Databases:** Another way of rating the SNPs is through external annotations like dbSNP. dbSNP is present in the annotations and can be downloaded. One can find the SNPs already listed in the database or look for SNPs that are novel i.e., not listed on the database.

4. **Find Significant SNPs:** is another way to prioritize the SNPs. It is most useful when one is dealing with multiple samples and multiple experimental setups. It can be used for quickly identifying population-specific variants, somatic mutations, and tumor specific markers by using filtering criteria based on attributes like total coverage and percent strand bias (both of which are fairly fixed) and supporting reads threshold (varies with respect to experiment or use case dependent). If we are looking at normal individuals the threshold could be as high as 35 to 50% but in cancer samples we may be looking for low frequency mutations then threshold could be as low as 5-10%.

The alleles are also filtered based on the number of samples/ groups where the allele is present. An allele present in large number of samples or groups is common allele and is a rare allele if it is present only in a small number of samples or groups. The exact specification of the confidence and the commonality criteria depends on the experimental design. The work has options to handle at least four different experimental designs or setups as depicted below:





The distribution of the alleles satisfying the specified filter conditions is shown as a histogram. The total number of alleles satisfying the filter condition is given on the top of the histogram.

A bar at a particular location in the histogram corresponds to the number of alleles satisfying the filter conditions whose supporting reads % is at least some value. The different shades in the bars correspond to different numbers of supporting samples. The histogram cannot show a large number of shades accurately. So when working with large number of samples, the samples are binned and instead of having one shade per sample, there will be one shade for a range of samples.

Clicking on `Finish' would save the filtered list of variant alleles as a child of the input MSVAL. The result of this analysis step is also an MSVAL; it will have no extra columns; only the rows that satisfy the chosen criteria will be present in the child MSVAL.

strandngs

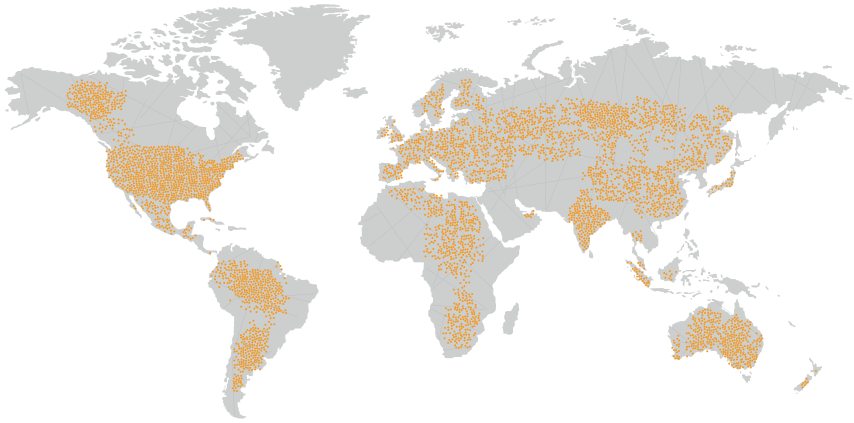
strand
New Generation Healthcare

About Strand

A History of Innovative Genomic Research

Strand Life Sciences is a global genomic profiling company and leader in precision medicine diagnostics, aimed at empowering cancer care and genetic testing for inherited diseases. Strand works with physicians and hospitals to enable faster clinical decision support for accurate molecular diagnosis, prognosis, therapy recommendations, and clinical trials. The Strand Center for Genomics & Personalized Medicine is India's 1st and only CAP & NABL accredited NGS laboratory.

www.strandls.com



A Trusted Partner to Companies Worldwide

For 15 years, our genomics products and solutions have facilitated the work of leading researchers and medical geneticists in over 2,000 laboratories and 100 hospitals around the world.

Strand Life Sciences Pvt. Ltd
5th Floor, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560024
Phone:+91-80-40 (787263) Fax: +91-80-4078-7299