SNP Detection and Prioritization Tutorial
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/ MNP/ 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 whereas 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.

![SNP Detection (Step 1 of 2)](image)
Step 2

a. Detect SNPs- SNP Detection input steps.

![Detect SNPs window](image1)

b. SNP Filters- SNP Detection input Steps.

![SNP Filters window](image2)
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.
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.

![SNP Result Node Inspector]

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

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.
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