DNA-Seq Analysis

Introduction

This tutorial describes the DNA-Seq analysis workflow in Avadis NGS using an amplicon sequencing dataset. It assumes that Avadis NGS is installed and the steps listed in the 'Getting Started' tutorial have been completed.

Dataset

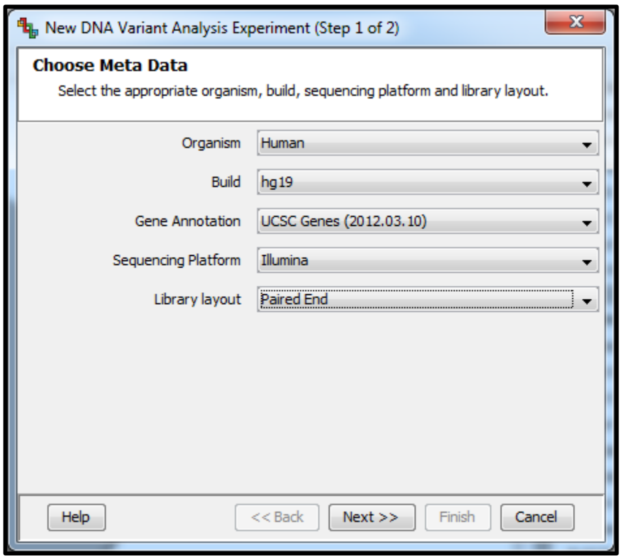
This is a paired-end amplicon sequencing data generated from 96 Yoruba HapMap samples. For the purpose of this tutorial, we will consider three samples from the same family (father, mother and child) out of the original 96. The dataset zip file (`dnaseq-small-dataset-illumina.zip`) contains the BAM files from these three samples. The reads were aligned against the hg19 assembly.

Goals

In this tutorial, we will check the quality of the amplicon resequencing visually and quantitatively, perform SNP detection and visualize the results, and priortize the list of identified SNPs for further analysis.

Sample Import

For this tutorial, we need to create a DNA Variant Analysis experiment with hg19 assembly and UCSC annotations. Choose "Illumina" as the sequencing platform and "Paired End" as the library layout.



New DNA Variant Analysis Experiment (Step 1 of 2)

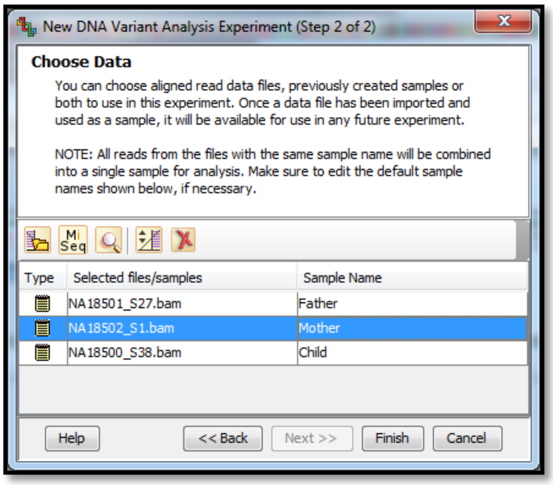
Choose Meta Data

Select the appropriate organism, build, sequencing platform and library layout.

Organism: Human
Build: hg19
Gene Annotation: UCSC Genes (2012.03.10)
Sequencing Platform: Illumina
Library layout: Paired End

Figure 1: DNA-Seq experiment creation

While loading the BAM files present in the zipped file `dnaseq-small-dataset-illumina.zip` (using the 1st icon from the left), we have an option of renaming the samples to make them easily distinguishable.



New DNA Variant Analysis Experiment (Step 2 of 2)

Choose Data

You can choose aligned read data files, previously created samples or both to use in this experiment. Once a data file has been imported and used as a sample, it will be available for use in any future experiment.

NOTE: All reads from the files with the same sample name will be combined into a single sample for analysis. Make sure to edit the default sample names shown below, if necessary.

Type	Selected files/samples	Sample Name
▤	NA18501_S27.bam	Father
▤	NA18502_S1.bam	Mother
▤	NA18500_S38.bam	Child

Figure 2: Renaming samples

Once the samples get loaded, the genome browser is launched showing the chr1 of each sample. The fact that similar peaks are present in each sample reflects the fact that this is indeed amplicon data.

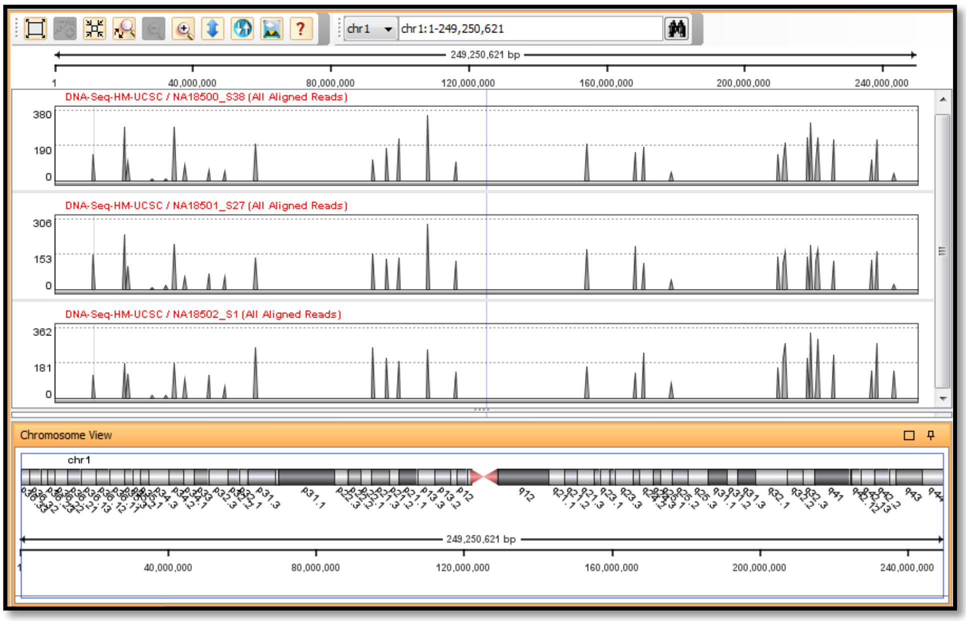


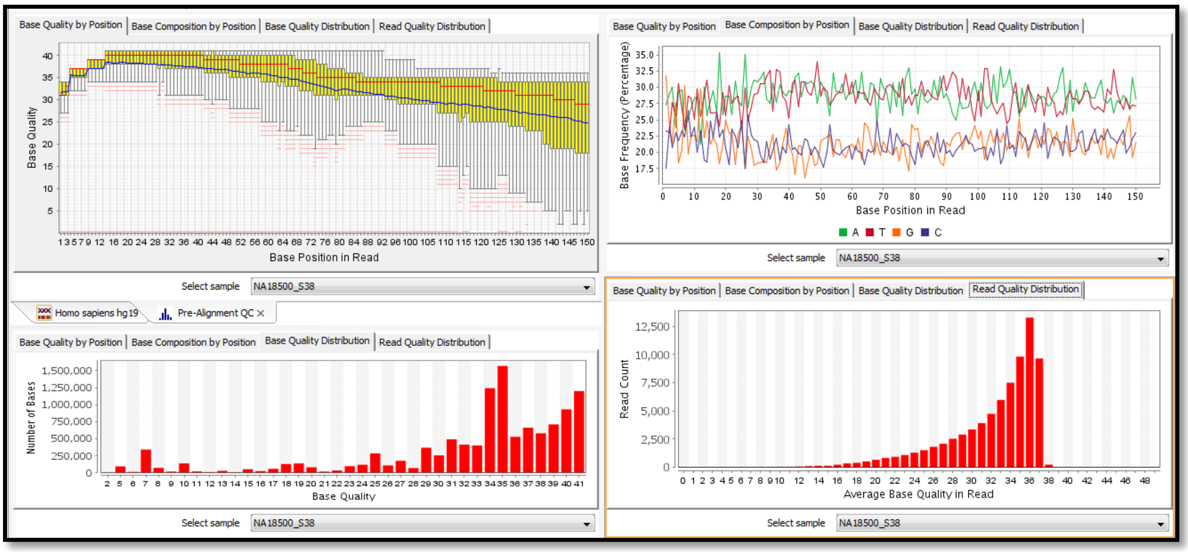
Figure 3: Genome browser tracks on initial launch

Quality Control

Since we are planning on running SNP analysis, the quality control metrics need to be carefully considered using the quality inspection in the workflow navigator (using the Quality Inspection section in the workflow browser).

Pre-Alignment QC Plots

Using these plots, we can discover the base quality by its position in the read, the approximate GC% of the amplicon regions, and the average quality of reads and bases in each sample.



Select sample NA18500_S38

Figure 4: Pre-Alignment QC plots

We can also take a look at QC plots specific to Illumina data (using Quality Inspection → Base Quality by Tile workflow step). The plot shows average base qualities rendererd as a heatmap with user controls for choosing the sample and lane of interest. Rows in the heatmap correspond to tiles of the lane under consideration, while columns represent flow cycles. Each cell is colored according to the average quality of all the bases that fall in the specific cycle of the specific tile.

The below figure shows the Base Quality QC plot for this dataset in which the qualities are good in the earlier flow cycles and degrade marginally towards the end.

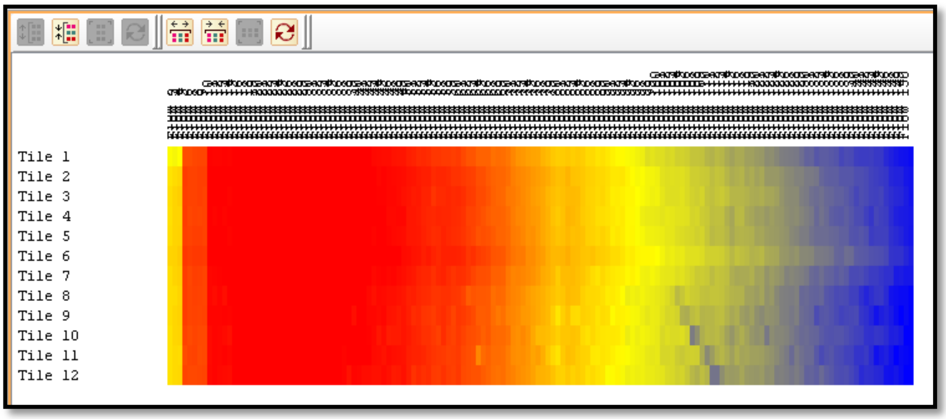


Figure 5: Base quality by tile

Post Alignment QC

A paired end library of a normal sample (i.e. not a cancer sample) is considered good quality if the majority of pairs align in the expected orientation and distance from each other. We can check the 'Match status' pie-charts (using the Quality Inspection → Alignment QC Plots workflow step), to verify if the paired end libraries are good.

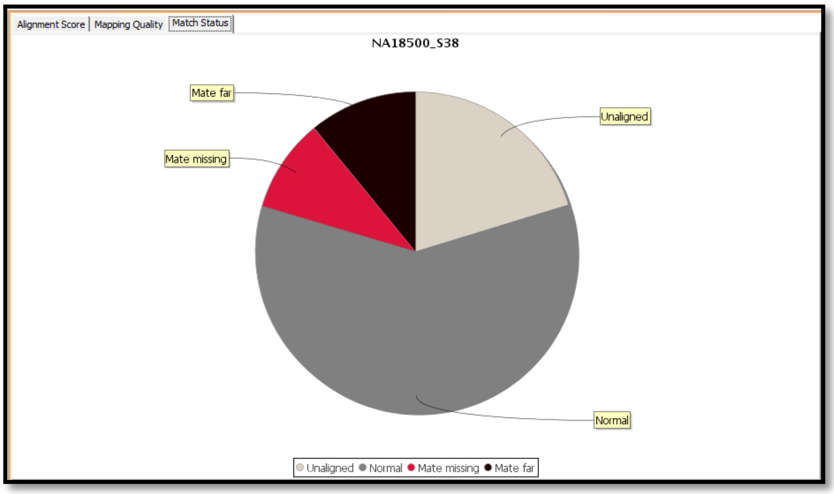
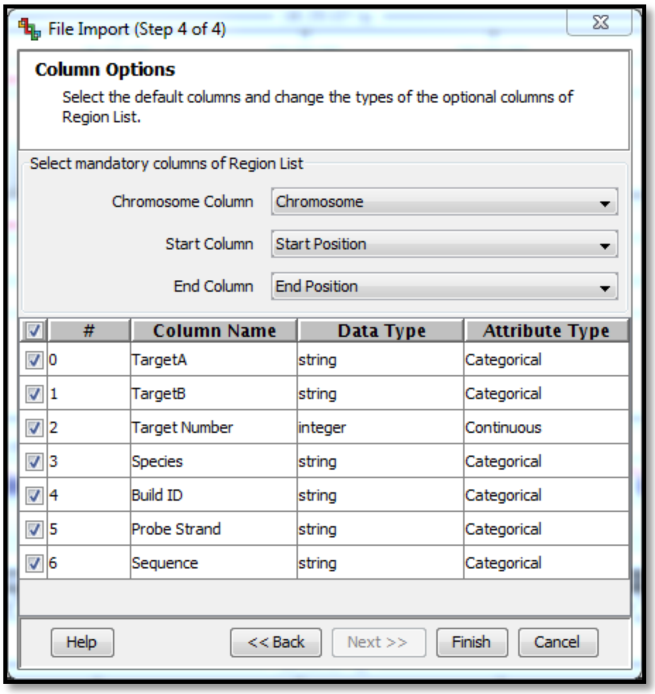


Figure 6: Match status of paired end reads

Targeted Region QC

One of the most important QC steps in amplicon analysis is to determine the efficacy of targeted resequencing. For this, we need to import the file containing information about the target region first (using the Utilities→ Import Region List workflow step). The Region list is a simple list of chromosome, start, end, strand and other annotation information, such as zygocity, score, type etc. For this tutorial, use the `Target-Regions.tsv` file provided with the dataset. (Note that currently, this is obtained by extracting the information from the [Targets] section of the Manifest file present in the MiSeq folder). After giving the file location in the first step, we can proceed with the defaults till step 4 and then mark the mandatory columns as shown below.



Column Options

Select the default columns and change the types of the optional columns of Region List.

Select mandatory columns of Region List

Chromosome Column: Chromosome

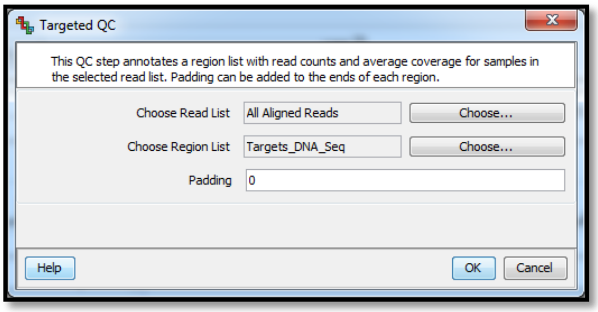
Start Column: Start Position

End Column: End Position

☑	#	Column Name	Data Type	Attribute Type
☑	0	TargetA	string	Categorical
☑	1	TargetB	string	Categorical
☑	2	Target Number	integer	Continuous
☑	3	Species	string	Categorical
☑	4	Build ID	string	Categorical
☑	5	Probe Strand	string	Categorical
☑	6	Sequence	string	Categorical

Figure 7: Column selection

After the region list gets imported, we can use this as input for the Quality Inspection → Targeted Region QC workflow step and run this step with default parameters.



This QC step annotates a region list with read counts and average coverage for samples in the selected read list. Padding can be added to the ends of each region.

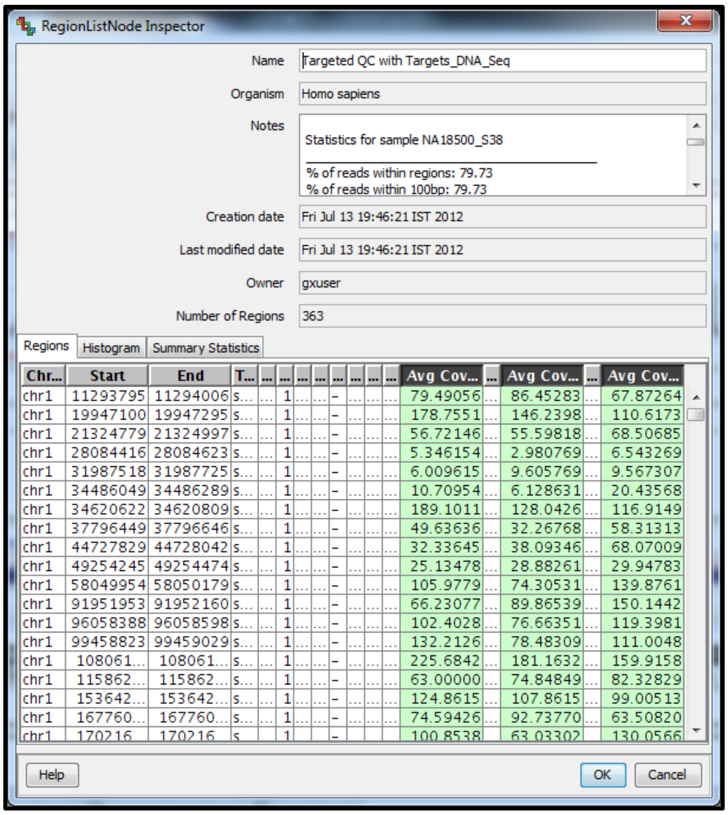
Choose Read List: All Aligned Reads

Choose Region List: Targets_DNA_Seq

Padding: 0

Figure 8: Target region QC

For this tutorial dataset, the coverage across the target regions is very high with all three samples having coverage of above 75%. This can be determined by going through the 'Notes' section in the inspector of the output of this analysis step.



Name: Targeted QC with Targets_DNA_Seq

Organism: Homo sapiens

Notes:
Statistics for sample NA18500_S38
% of reads within regions: 79.73
% of reads within 100bp: 79.73

Creation date: Fri Jul 13 19:46:21 IST 2012

Last modified date: Fri Jul 13 19:46:21 IST 2012

Owner: gxuser

Number of Regions: 363

Chr...	Start	End	T...	...	...	...	...	...	...	...	...	Avg Cov...	...	Avg Cov...	...	Avg Cov...
chr1	11293795	11294006	s...	...	1	...	...	-	...	...	...	79.49056	...	86.45283	...	67.87264
chr1	19947100	19947295	s...	...	1	...	...	-	...	...	...	178.7551	...	146.2398	...	110.6173
chr1	21324779	21324997	s...	...	1	...	...	-	...	...	...	56.72146	...	55.59818	...	68.50685
chr1	28084416	28084623	s...	...	1	...	...	-	...	...	...	5.346154	...	2.980769	...	6.543269
chr1	31987518	31987725	s...	...	1	...	...	-	...	...	...	6.009615	...	9.605769	...	9.567307
chr1	34486049	34486289	s...	...	1	...	...	-	...	...	...	10.70954	...	6.128631	...	20.43568
chr1	34620622	34620809	s...	...	1	...	...	-	...	...	...	189.1011	...	128.0426	...	116.9149
chr1	37796449	37796646	s...	...	1	...	...	-	...	...	...	49.63636	...	32.26768	...	58.31313
chr1	44727829	44728042	s...	...	1	...	...	-	...	...	...	32.33645	...	38.09346	...	68.07009
chr1	49254245	49254474	s...	...	1	...	...	-	...	...	...	25.13478	...	28.88261	...	29.94783
chr1	58049954	58050179	s...	...	1	...	...	-	...	...	...	105.9779	...	74.30531	...	139.8761
chr1	91951953	91952160	s...	...	1	...	...	-	...	...	...	66.23077	...	89.86539	...	150.1442
chr1	96058388	96058598	s...	...	1	...	...	-	...	...	...	102.4028	...	76.66351	...	119.3981
chr1	99458823	99459029	s...	...	1	...	...	-	...	...	...	132.2126	...	78.48309	...	111.0048
chr1	108061...	108061...	s...	...	1	...	...	-	...	...	...	225.6842	...	181.1632	...	159.9158
chr1	115862...	115862...	s...	...	1	...	...	-	...	...	...	63.00000	...	74.84849	...	82.32829
chr1	153642...	153642...	s...	...	1	...	...	-	...	...	...	124.8615	...	107.8615	...	99.00513
chr1	167760...	167760...	s...	...	1	...	...	-	...	...	...	74.59426	...	92.73770	...	63.50820
chr1	170216	170216	s		1			-				100.8538		63.03302		130.0566

Another important QC step is to determine if there are any regions with uniformly low coverage across all the samples. To do this visually, we can perform clustering on both samples and regions based on the average coverage (using Utilites→ Cluster Regions workflow step). This would give us the below image wherein we can see certain regions having uniform low coverage (blue) across the samples.

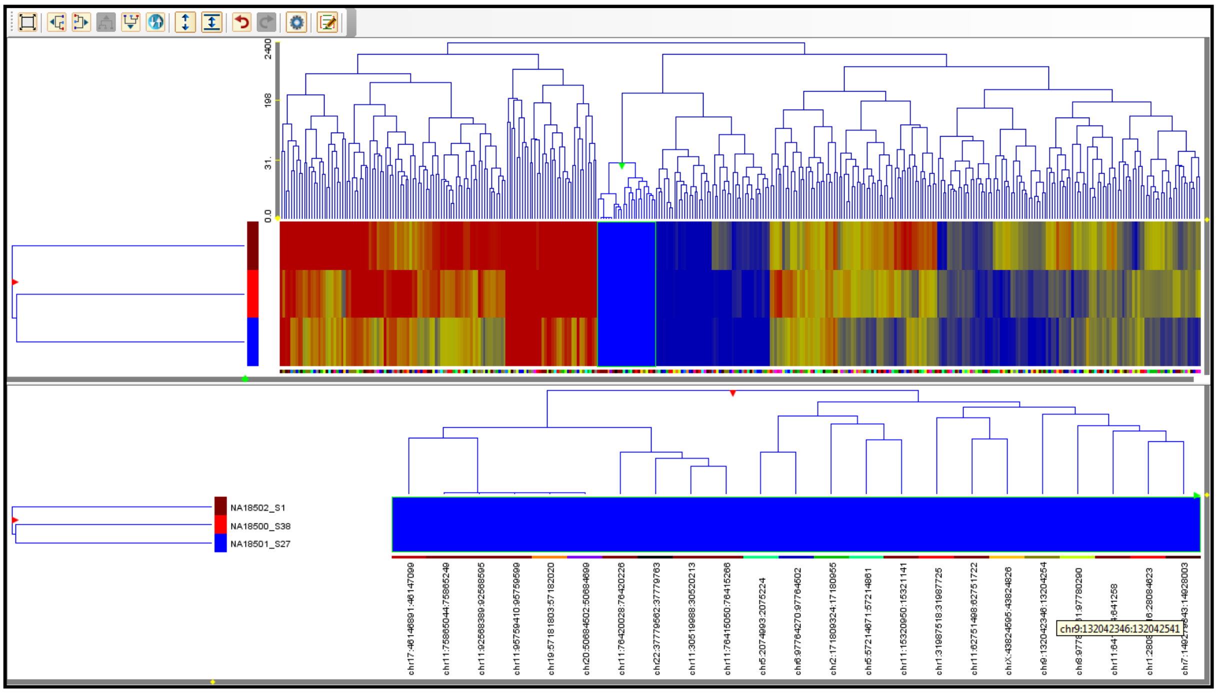
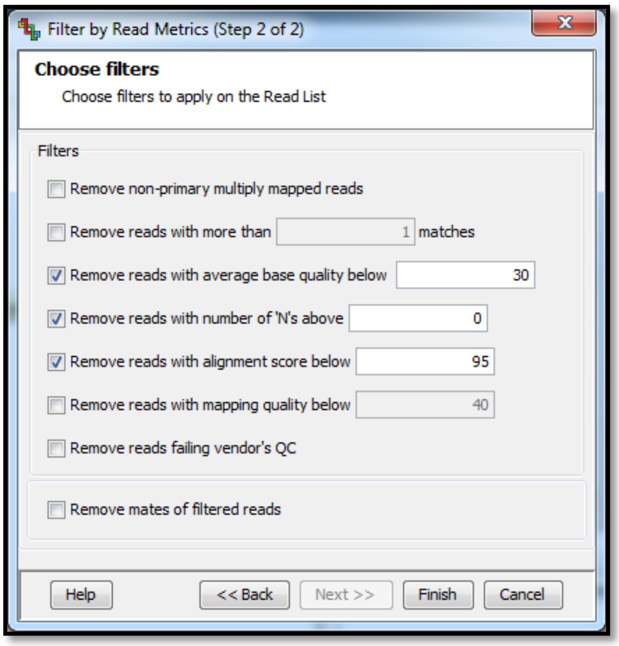


Figure 9: Cluster on target region QC list

Filter

Before proceeding with SNP Detection, we can filter out some of the low quality reads (using Filter → Filter by Read Metrics workflow step) with criteria as shown below. This would leave us with a total of 156,512 reads from the original total of 229,262 reads.



Filter by Read Metrics (Step 2 of 2)

Choose filters

Choose filters to apply on the Read List

Filters

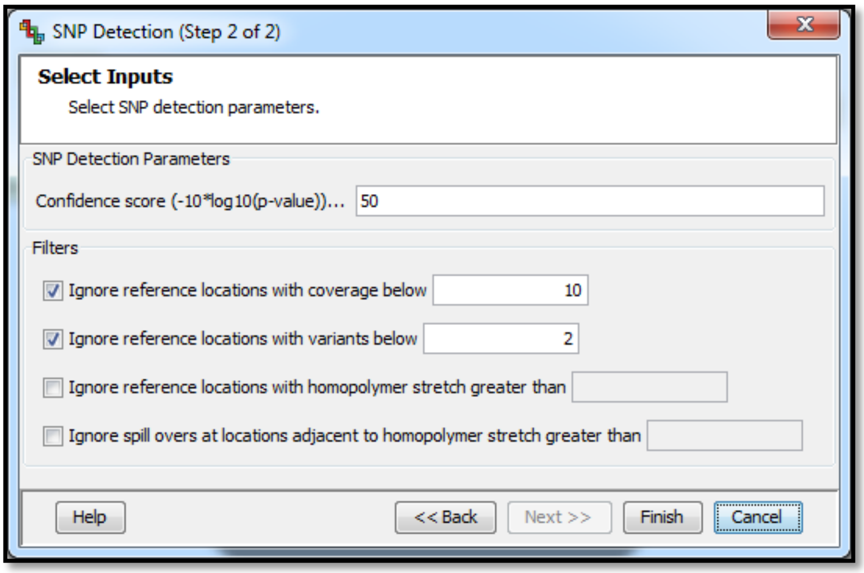
- □ Remove non-primary multiply mapped reads
- □ Remove reads with more than 1 matches
- ☑ Remove reads with average base quality below 30
- ☑ Remove reads with number of 'N's above 0
- ☑ Remove reads with alignment score below 95
- □ Remove reads with mapping quality below 40
- □ Remove reads failing vendor's QC

□ Remove mates of filtered reads

Figure 10: Filter criteria

SNP Detection

After QC and filtering, we are ready to proceed with the variant analysis. We can run SNP Detection on the filtered read list using the Analysis→ SNP Detection workflow step. We can choose a dbSNP annotation database. This needs to be downloaded from the 'Annotations Manager'. Please note that this takes a significant amount of time so we can choose the option as 'None' if we want to skip that part.



SNP Detection (Step 2 of 2)

Select Inputs

Select SNP detection parameters.

SNP Detection Parameters

Confidence score (-10*log10(p-value))... 50

Filters

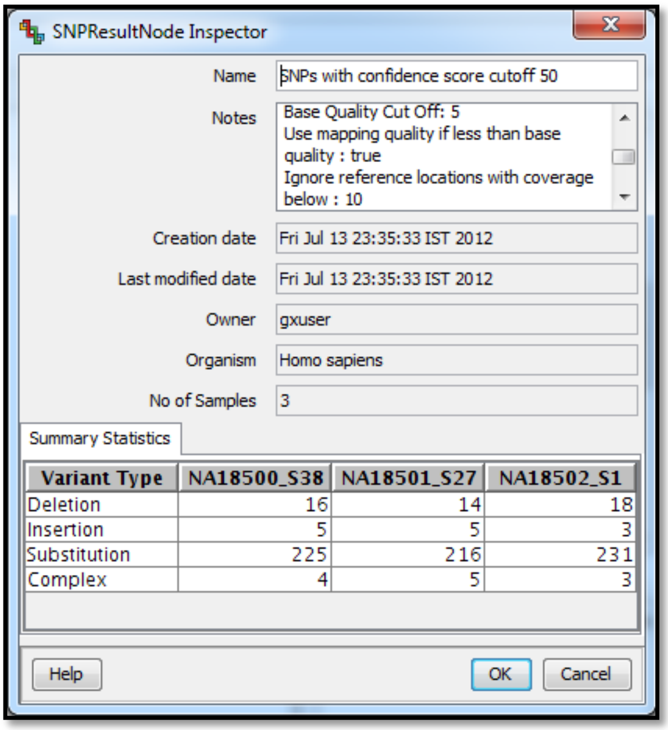
☑ Ignore reference locations with coverage below 10

☑ Ignore reference locations with variants below 2

□ Ignore reference locations with homopolymer stretch greater than

□ Ignore spill overs at locations adjacent to homopolymer stretch greater than

Figure 11: SNP Detection Parameters



SNPResultNode Inspector

Name: SNPs with confidence score cutoff 50

Notes: Base Quality Cut Off: 5
Use mapping quality if less than base quality : true
Ignore reference locations with coverage below : 10

Creation date: Fri Jul 13 23:35:33 IST 2012

Last modified date: Fri Jul 13 23:35:33 IST 2012

Owner: gxuser

Organism: Homo sapiens

No of Samples: 3

Summary Statistics

Variant Type	NA18500_S38	NA18501_S27	NA18502_S1
Deletion	16	14	18
Insertion	5	5	3
Substitution	225	216	231
Complex	4	5	3

Figure 12: SNP results

The SNP results are divided into single-base variant lists and multi-base variant lists. The SNPs of each sample are listed independently. In addition a multi-sample report is created which is the union of all variants found across all the samples. We can double-click and open each of these outputs to get an idea of their contents.

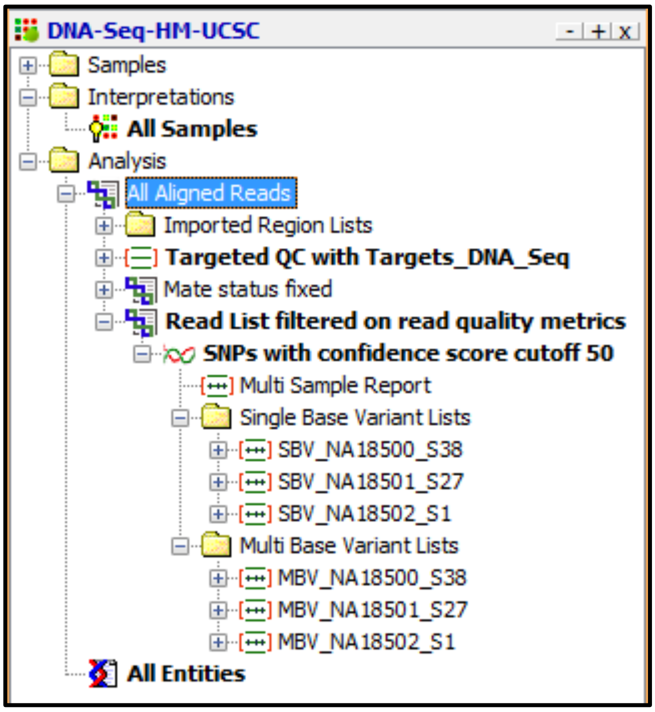


Figure 13: Experiment navigator after SNP calling

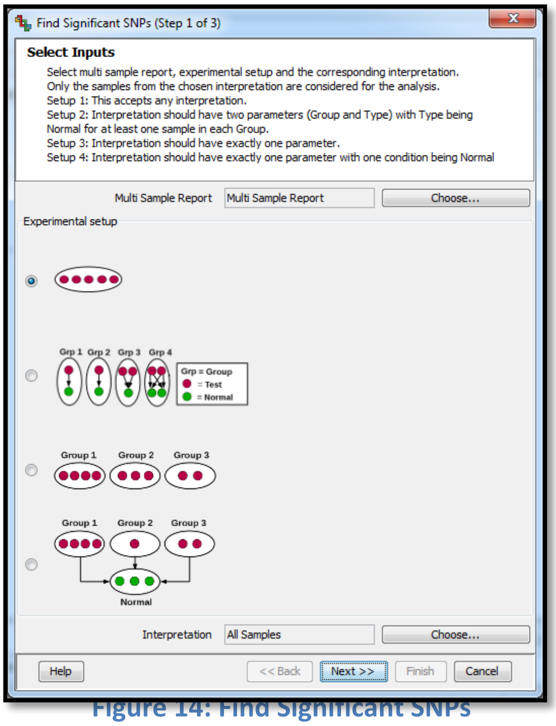
The multi-sample report has one variant allele per row. For each variant allele it has six sample specific columns:

- total reads
- % reads supporting variant allele
- % reads which are different from the reference
- Strand bias
- SNP call at this location (could be ref)
- Score of the SNP call (empty if no SNP)

Since we have 3 samples, we will have 3x6 plus the additional allele specific columns in the multi-sample report. The next step would be to narrow down this list of SNPs to the most interesting ones according to the use case.

Find Significant SNPs

In this step, we need to specify the multiple sample report and an interpretation depending on the experimental setup. The below figure shows the various experimental setups that are supported in this analysis step.



Find Significant SNPs (Step 1 of 3)

Select Inputs

Select multi sample report, experimental setup and the corresponding interpretation.
Only the samples from the chosen interpretation are considered for the analysis.
Setup 1: This accepts any interpretation.
Setup 2: Interpretation should have two parameters (Group and Type) with Type being Normal for at least one sample in each Group.
Setup 3: Interpretation should have exactly one parameter.
Setup 4: Interpretation should have exactly one parameter with one condition being Normal

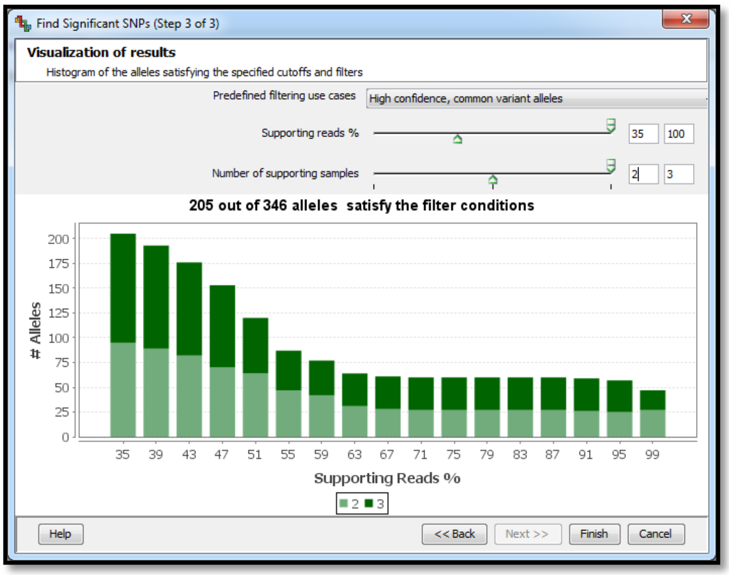
Multi Sample Report: Multi Sample Report

Experimental setup

Interpretation: All Samples

Figure 14: Find Significant SNPs

On selecting the first setup, 119 of the 346 alleles listed in the multi-sample report are present in 3 out of 3 samples with reasonable number of supporting reads (35%). The number of supporting samples can be changed using the slider.



Find Significant SNPs (Step 3 of 3)

Visualization of results

Histogram of the alleles satisfying the specified cutoffs and filters

Predefined filtering use cases: High confidence, common variant alleles

Supporting reads %: 35 100

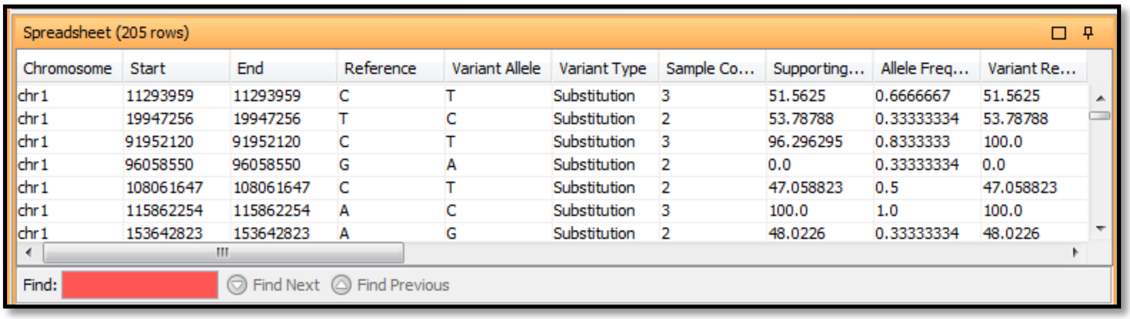
Number of supporting samples: 2 3

205 out of 346 alleles satisfy the filter conditions

Figure 15: Visualization of results

Visualization of Results

To quickly browse through the 119 alleles that were saved, we can right-click on output in the experiment navigator and select the option to navigate in the Genome Browser. The Navigate Spreadsheet tab in the Genome Browser will show the contents of the region list.



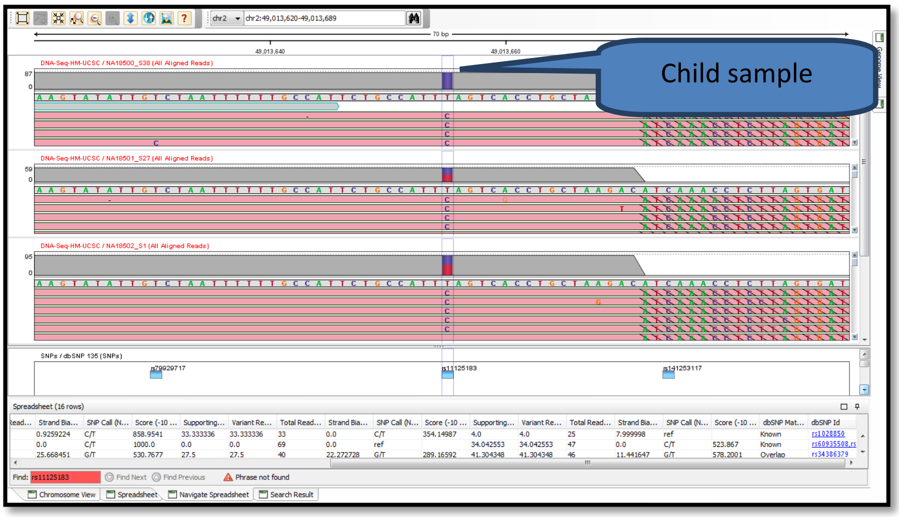
Spreadsheet (205 rows)

Chromosome	Start	End	Reference	Variant Allele	Variant Type	Sample Co...	Supporting...	Allele Freq...	Variant Re...
chr1	11293959	11293959	C	T	Substitution	3	51.5625	0.6666667	51.5625
chr1	19947256	19947256	T	C	Substitution	2	53.78788	0.33333334	53.78788
chr1	91952120	91952120	C	T	Substitution	3	96.296295	0.8333333	100.0
chr1	96058550	96058550	G	A	Substitution	2	0.0	0.33333334	0.0
chr1	108061647	108061647	C	T	Substitution	2	47.058823	0.5	47.058823
chr1	115862254	115862254	A	C	Substitution	3	100.0	1.0	100.0
chr1	153642823	153642823	A	G	Substitution	2	48.0226	0.33333334	48.0226

Find: ○ Find Next ○ Find Previous

Figure 16: Navigating the region list in the Genome Brower

When we double click on a specific region as shown in the below image, the view zooms to that location. We can also look for a particular region of interest using the search functionality. Alternatively we can also scan through the browser looking for interesting SNPs. In the below figure, for this particular variant, the parents are heterozygous whereas the child is homozygous indicating a loss of heterozygosity.



chr2
chr2:49,013,620-49,013,689
70 bp
49,013,640
49,013,660
DNA-Seq-HM-UCSC / NA18500_S38 (All Aligned Reads)
87
0
Child sample
DNA-Seq-HM-UCSC / NA18501_S27 (All Aligned Reads)
59
0
DNA-Seq-HM-UCSC / NA18502_S1 (All Aligned Reads)
95
0
SNPs / dbSNP 135 (SNPs)
rs79929717
rs11125183
rs141253117
Genome View

Spreadsheet (16 rows)

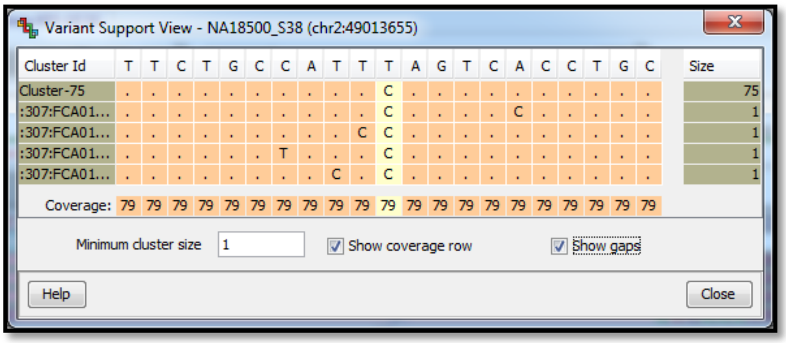
Read...	Strand Bia...	SNP Call (N...	Score (-10 ...	Supporting...	Variant Re...	Total Read...	Strand Bia...	SNP Call (N...	Score (-10 ...	Supporting...	Variant Re...	Total Read...	Strand Bia...	SNP Call (N...	Score (-10 ...	dbSNP Mat...	dbSNP Id
	0.9259224	C/T	858.9541	33.333336	33.333336	33	0.0	C/T	354.14987	4.0	4.0	25	7.999998	ref		Known	rs1028850
	0.0	C/T	1000.0	0.0	0.0	69	0.0	ref		34.042553	34.042553	47	0.0	C/T	523.867	Known	rs60935508,rs
	25.668451	G/T	530.7677	27.5	27.5	40	22.272728	G/T	289.16592	41.304348	41.304348	46	11.441647	G/T	578.2001	Overlap	rs34386379

Find: rs11125183 ○ Find Next ○ Find Previous ⚠ Phrase not found

Chromosome View | Spreadsheet | Navigate Spreadsheet | Search Result

Figure 17: GB view of the SNP

From the mismatch histogram alone it might be hard to glean information about the alleles present at this location. For this, we can right-click on the genome browser track at the location of interest and launch the variant support view. It confirms that in the present case it is a homozygous variant in the off-spring.



Variant Support View - NA18500_S38 (chr2:49013655)

Cluster Id	T	T	C	T	G	C	C	A	T	T	T	A	G	T	C	A	C	C	T	G	C	Size
Cluster-75	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	75
:307:FCA01...	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	C	.	.	.	.	.	1
:307:FCA01...	.	.	.	.	.	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	1
:307:FCA01...	.	.	.	.	.	.	T	.	.	.	C	.	.	.	.	.	.	.	.	.	.	1
:307:FCA01...	.	.	.	.	.	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	1
Coverage:	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	

Minimum cluster size 1

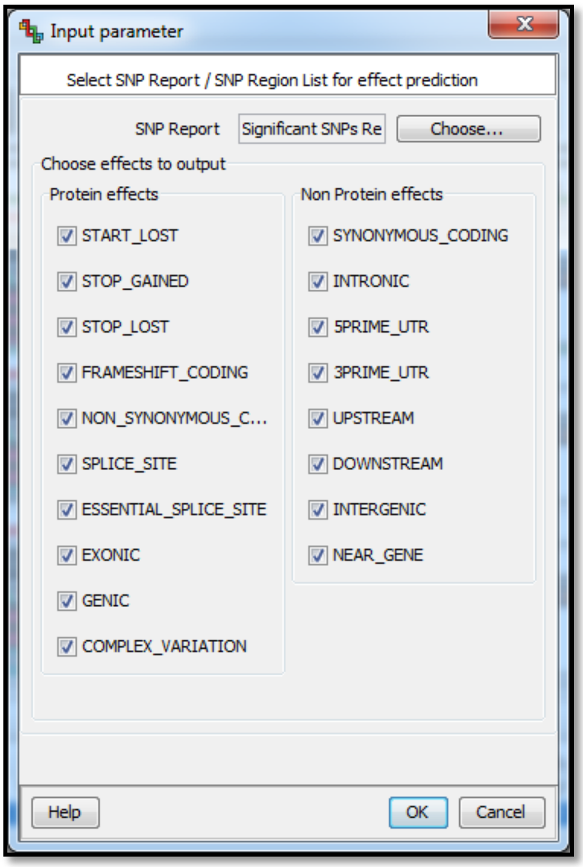
☑ Show coverage row

☑ Show gaps

Figure 18: Variant Support View

SNP Effect Analysis

We can see the effect of these 119 alleles by using the SNP effect analysis. Please select all the effects and run the analysis. This would output a region list as well as a gene list. In the present case, we would see the presence of a SNP in an exonic region and the resultant affected gene(MPZL1) in the gene list . This functionality is similar to the Ensembl SNP Effect predictor tool.



Input parameter

Select SNP Report / SNP Region List for effect prediction

SNP Report: Significant SNPs Re

Choose effects to output

Protein effects

- ☑ START_LOST
- ☑ STOP_GAINED
- ☑ STOP_LOST
- ☑ FRAMESHIFT_CODING
- ☑ NON_SYNONYMOUS_C...
- ☑ SPLICE_SITE
- ☑ ESSENTIAL_SPLICE_SITE
- ☑ EXONIC
- ☑ GENIC
- ☑ COMPLEX_VARIATION

Non Protein effects

- ☑ SYNONYMOUS_CODING
- ☑ INTRONIC
- ☑ 5PRIME_UTR
- ☑ 3PRIME_UTR
- ☑ UPSTREAM
- ☑ DOWNSTREAM
- ☑ INTERGENIC
- ☑ NEAR_GENE

Figure 19: SNP effect analysis

If you have access to the whole dataset containing the MiSeq run folder for all the 96 samples, you can load that in and repeat this analysis with the larger dataset.

This concludes our DNA-Seq tutorial. This is a very brief overview of the DNA-Seq experiment workflow in Avadis NGS. For more details or clarifications, please revert back (sales@avadisngs.com or support@avadisngs.com) and we will address your queries.