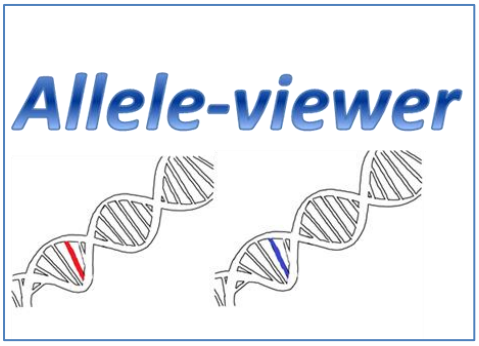
„Allele Viewer“ app



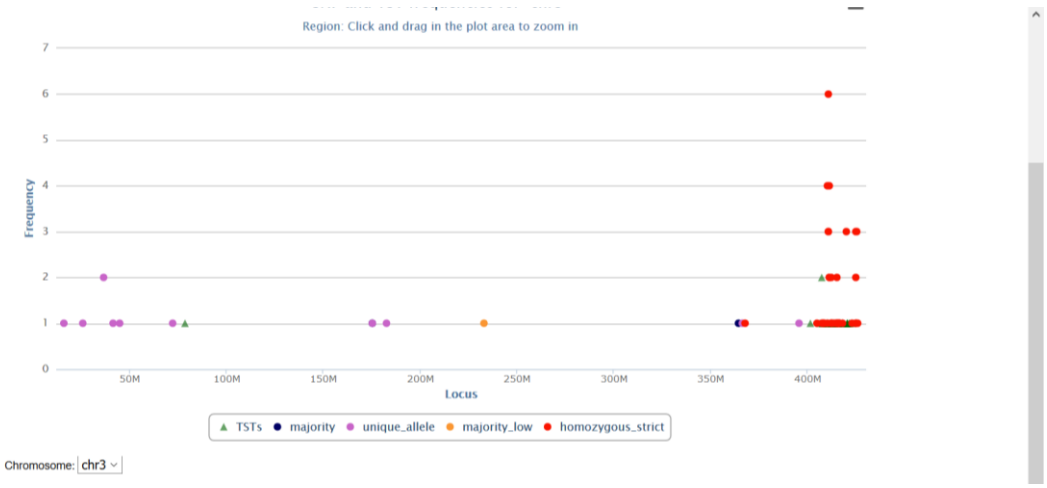
The “Allele Viewer” app is focused on SNPs and transcripts that are exclusively present in one of the two samples in a comparison. The rationale behind this app is to identify genetic region(s) that carry the genetic information which is responsible for the specific traits of the sample. The app combines SNP information and unique transcripts and maps them on the genome. We use the terms “Trait-specific SNPs” (TSS) and “Trait specific transcripts” (TSTs) for SNPs and transcripts that are found exclusively or to a higher degree in one of the two samples that are compared (for more information about the SNP- filtering, please see section” SNP Filter” below).

The SNP-containing MACE- contigs, RNA-Seq reads or genomic-reads are mapped to either the genome of the sample or to the genome of related organisms. Each chromosome is divided into 10kb bins and the amount of TSSs and TSTs are quantified for each bin. The result is displayed in a table showing detailed information about the transcript or genomic region that contains the SNPs.

Getting started:

After login and starting of the app, your different projects and comparisons can be chosen to open the results table which contains the mapping information, information form the SNV-mix output and annotation if available. For a more detailed information about the results table, please see section “TSS and TST results table” below. The results are based on the tables **“SNP_Result_annotated.csv” and the MACE results (“all_comparisons_merged.xls”)** that are available in the “**data center**” app. The different SNP-types are explained in the Section “SNP-filtering from SNV-mix output” below.

Chromosome Chart



The “toggle chart mode” button opens the chart mode of the tool. Again, the different pairwise comparisons can be chosen here.

The chart shows an overview of a chromosome and quantifies the amount TSTs (in red) and TSSs (in blue) on the chromosome. In the lower left part of the screen, the chromosomes can be chosen.

Zoom:

By holding down the left mouse button and scrolling over the chart, the region of interest can be enlarged.

SNP-filtering from SNV-mix output:

1) SNP_type_homozygous: All loci must have at least 5 reads in both samples. Alleles must be distributed in an AA vs BB fashion. Because of possible sequencing errors and barcode misinterpretation, we allow one false allele read in 100, e.g. 100xA and 1XB in sample 1 and 100xB in sample 2 would still be considered as homozygous.

2) SNP_type_unique_allele: All loci must be covered by at least 5 reads per sample. One Allele must be unique, hence AB vs. AA or AB vs. BB.

3) SNP_type_majority_strict: Alleles that are at least 10x overrepresented within one sample sample and at least 10x underrepresented in the other sample.

4) SNP_type_majority: All loci must be covered by at least 10 reads per sample. Alleles that are at least 2x overrepresented when compared in one sample and at least 2x underrepresented in to other sample.

5) SNP_type_majority_low: All loci must be covered by at least 20 reads per sample. Alleles that are at least 1.25 x overrepresented when compared in one sample and at least 1.25 x underrepresented in to other sample.

SNV-results: Short Description of the Excel-File

The results are provided in tab-separated format. The table should be opened with Excel 2007 or higher, otherwise the amount of rows is limited to 65,536.

Numbers are provided with points as separators for decimals.
Excel uses decimal separators according to the regional settings. If in your region commas are used as separators for decimals, please follow the instructions described under the following link to change this temporarily: http://office.microsoft.com/en-us/excel-help/change-the-separator-for-thousands-and-decimals-HP003089676.aspx?CTT=1

The following information is provided in separated columns:

Column	Explanation	remark
chrom	**Chromosome were SNV is located**	
position	**Position on chromosome**	
ref_base	**Reference base**	
var_base	**Variant base (base found in sample)**	
Sample_1_counts_ref	**number of evidences, were the ref base was identified at the position in sample 1**	
Sample_1_counts_var	**number of evidences, were the var base was identified at the position in sample1**	
Sample_2_counts_ref	**number of evidences, were the ref base was identified at the position in sample 2**	

Sample_2_counts_var	**number of evidences, were the var base was identified at the position in sample 2**	
p_AA_AA	**p-value for allele combination AA_AA**	
p_AA_AB	**p-value for allele combination AA_AB**	
p_AA_BB	**p-value for allele combination AA_BB**	
p_AB_AA	**p-value for allele combination AB_AA**	
p_AB_AB	**p-value for allele combination AB_AB**	
p_AB_BB	**p-value for allele combination AB_BB**	
p_BB_AA	**p-value for allele combination BB_AA**	
p_BB_AB	**p-value for allele combination BB_AB**	
p_BB_BB	**p-value for allele combination BB_BB**	
Query	**query name**	**BLAST**
SNP-type	**See explanation below(1)**	
Query length	**query length**	**BLAST**
Subject	**subject name**	**BLAST**
Subject length	**subject length**	**BLAST**
% id	**% of identical base-pairs**	**BLAST**
alignment length	**length of aligned stretch of nucleotides**	**BLAST**
mistmatches	**number of mismatches**	**BLAST**
gap openings	**number of gap-openings**	**BLAST**
q.start	**query sequence start position**	**BLAST**
q.end	**query sequence end position**	**BLAST**
s.start	**subject sequence start position**	**BLAST**
s.end	**subject sequence start end**	**BLAST**
e-value	**BLAST e-value**	**BLAST**
bit score	**BLAST bit-score**	**BLAST**
Region (+10000bp)	**Genomic region + 10,000 bps**	**The chromosome is distributed into 10,000 bp subregions. This information is used to quantify the amount of SNPs in the region of 10 000 bps**
100 bp sequence 5 prime of SNV	**100 bps 5' (upstream) of the SNV**	
RefBase/SNV (var base)	**The reference base and the variant base**	
100 bp sequence 3 prime of SNV	**100 bps 3' (downstream) of the SNV**	

[1] Based on the SNV-mix analysis, the following SNPs were chosen:

6) SNP_type_homozygous: All loci must have at least 5 reads in both samples. Alleles must be distributed in an AA vs BB fashion. Because of possible sequencing errors and barcode misinterpretation, we allow one false allele read in 100, e.g. 100xA and 1XB in sample 1 and 100xB in sample 2 would still be considered as homozygous.

7) SNP_type_unique_allel: All loci must be covered by at least 5 reads per sample. One Allele must be unique, hence AB vs. AA or AB vs. BB.

8) SNP_type_majority_strict: Alleles that are at least 10x overrepresented within one sample sample and at least 10x underrepresented in the other sample.

9) SNP_type_majority: All loci must be covered by at least 10 reads per sample. Alleles that are at least 2x overrepresented when compared in one sample and at least 2x underrepresented in to other sample.

10) SNP_type_majority_low: All loci must be covered by at least 20 reads per sample. Alleles that are at least 1.25 x overrepresented when compared in one sample and at least 1.25 x underrepresented in to other sample.

Filter:
The data can be conveniently analyzed by using the "filter" option of Excel, to filter for example for keywords e.g. "Transcription". The filters can be combined.
For more information please visit http://office.microsoft.com/en-us/excel-help/filter-data-in-a-range-or-table-HP010073941.aspx

Literature:
SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Goya et al. - Bioinformatics. 2010 March 15; 26(6): 730–736.