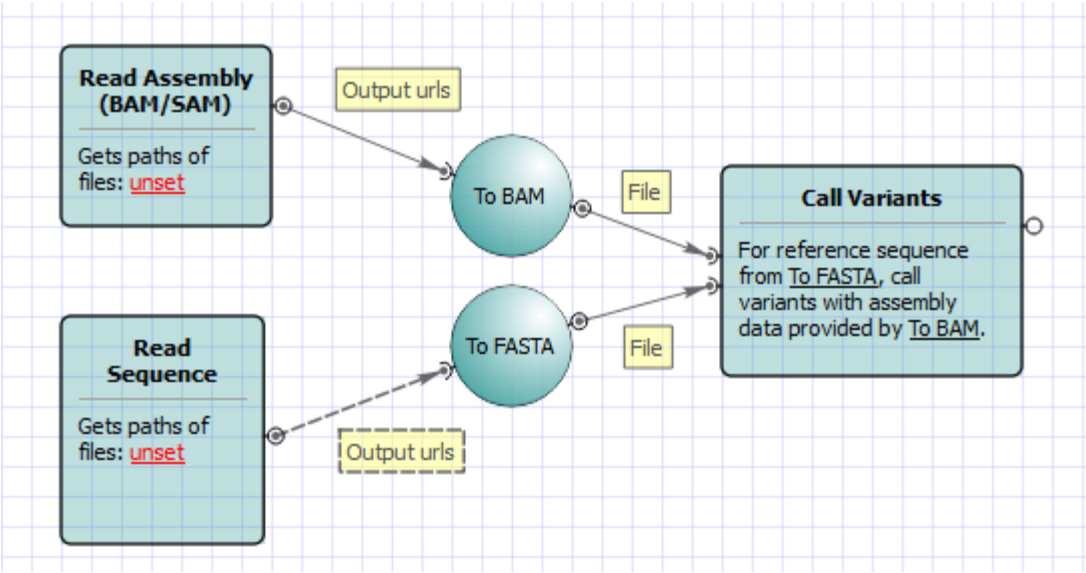
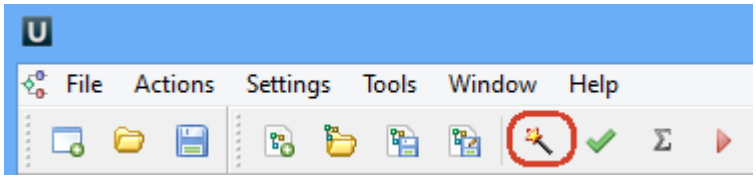
Call Variants with SAMtools

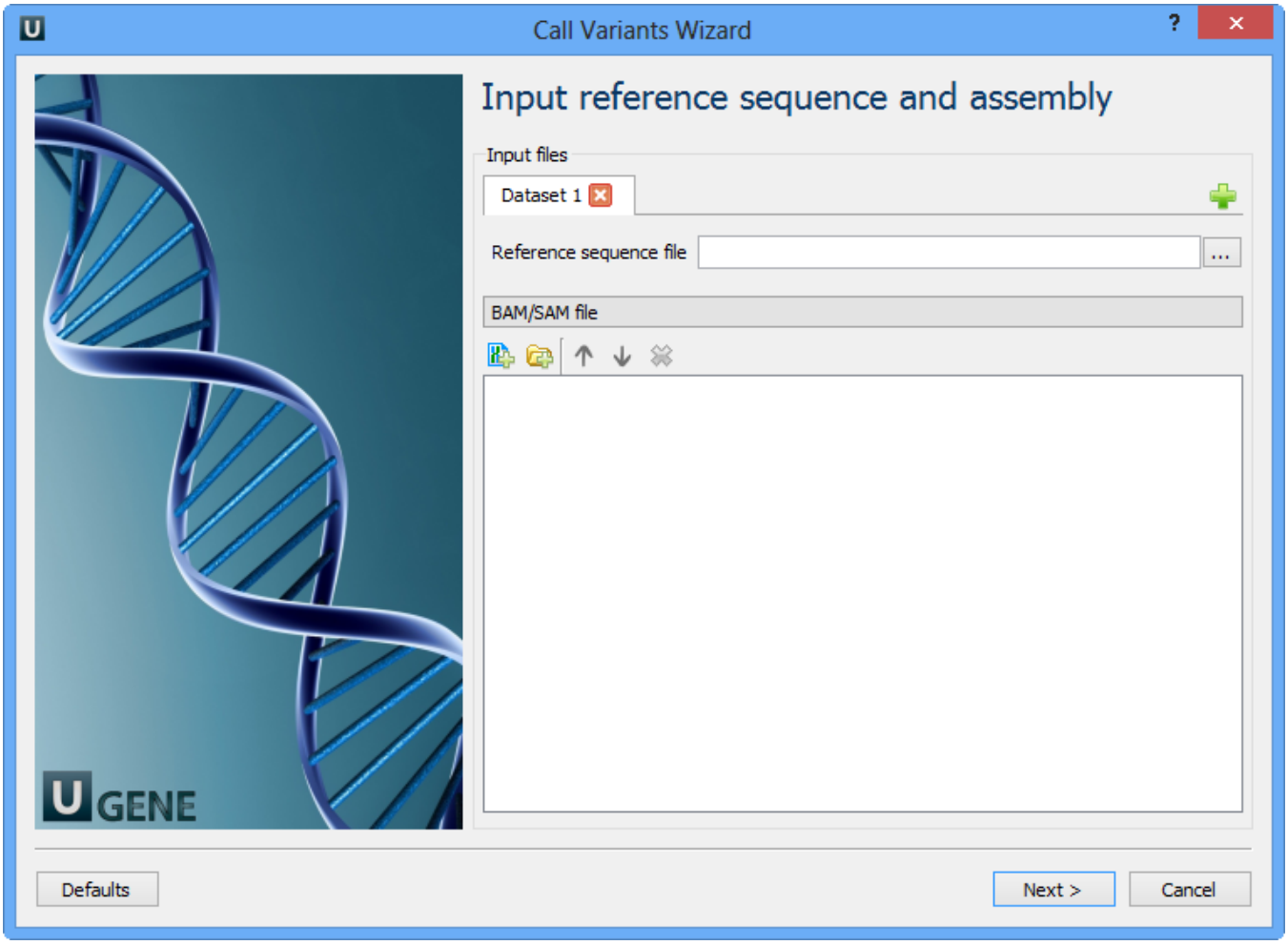
Call variants in UGENE can be done using SAMtools mpileup and bcftools view utilities. To read additional information about SAMtools and its utilities visit SAMtools homepage. Both utilities are embedded into UGENE and there is no need in additional configuration.



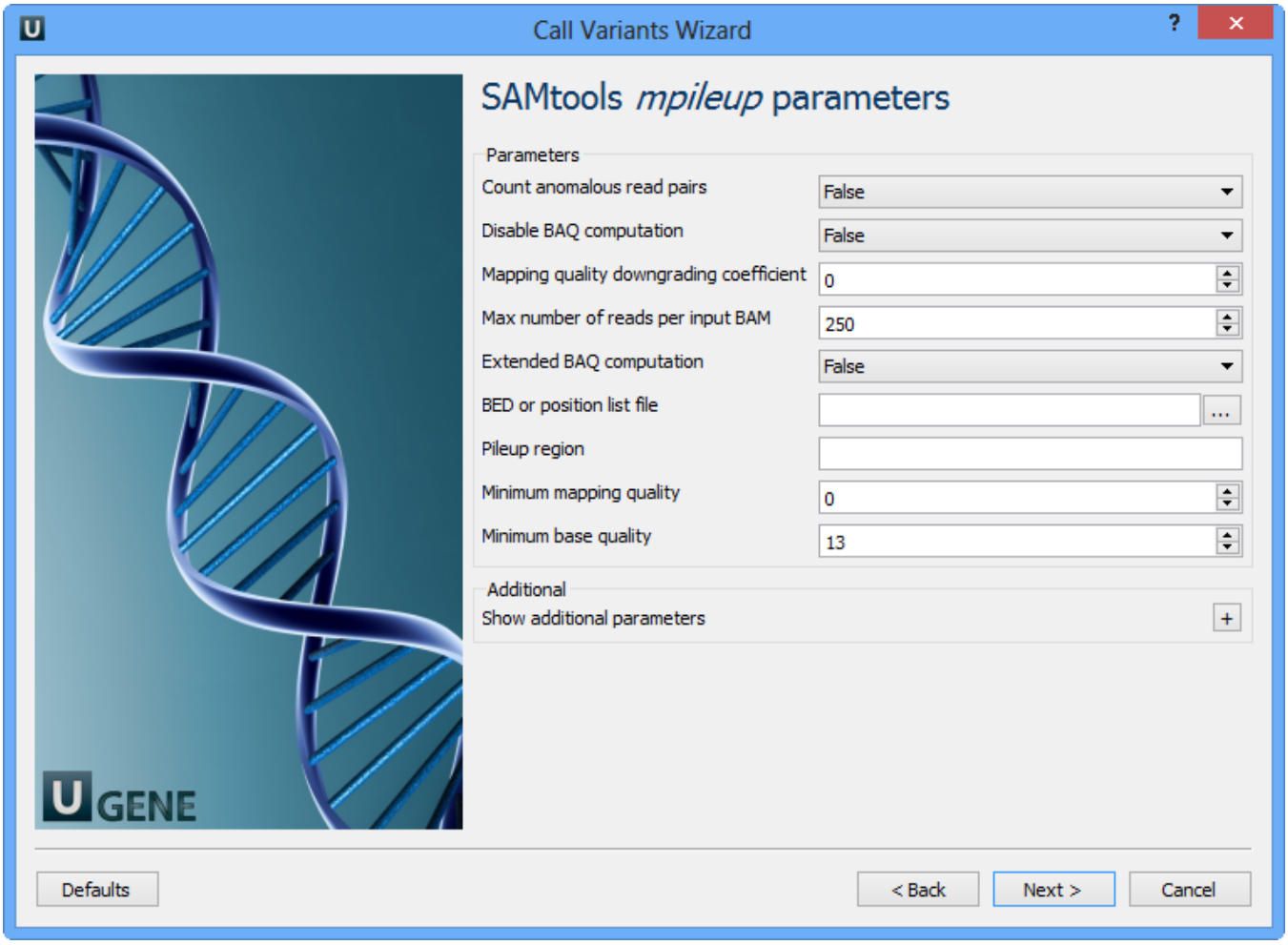
To run the worflow you need to select an input reference sequence, a BAM or SAM file and an output file with variations. Optionally, you can change other parameters, for example, set additional parameters of the SAMtools mpileup and bcftools view utilities. Use the workflow wizard to guide you through the parameters setup process. Click Show wizard button on the Workflow Designer toolbar to open it:



The first wizard page appears:



Here you need to input a file with a reference sequence and a sorted BAM or SAM file. Note that the input BAM or SAM file may be unsorted. Click the Next button. The next page appears:

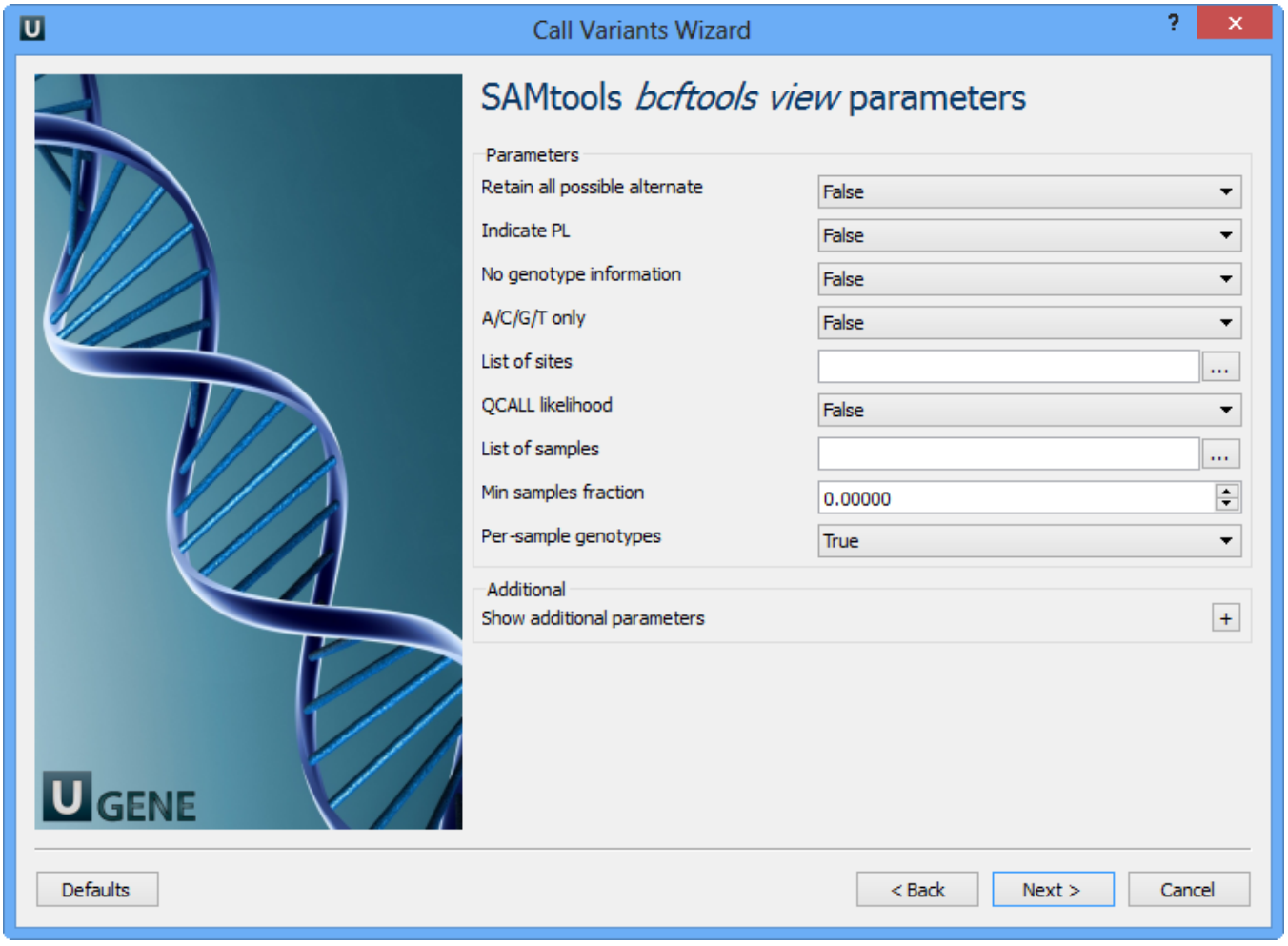


Here you can change default parameters of the SAMtools mpileup utility. To show additional parameters click the + button. The following parameters are available:

Count anomalous read pairs	Do not skip anomalous read pairs in variant calling.
Disable BAQ computation	Disable probabilistic realignment for the computation of base alignment quality (BAQ). BAQ is the Phred-scaled probability of a read base being misaligned. Applying this option greatly helps to reduce false SNPs caused by misalignments.
Mapping quality downgroading coefficient	Coefficient for downgrading mapping quality for reads containing excessive mismatches. Given a read with a phred-scaled probability q of being generated from the mapped position, the new mapping quality is about sqrt((INT-q)/INT)*INT. A zero value disables this functionality; if enabled, the recommended value for BWA is 50.
Max number of reads per input BAM	At a position, read maximally INT reads per input BAM.
Extended BAQ computation	Extended BAQ computation. This option helps sensitivity especially for MNPs, but may hurt specificity a little bit.
BED or position list file	BED or position list file containing a list of regions or sites where pileup or BCF should be generated.
Pileup region	Only generate pileup in region STR.
Minimum mapping quality	Minimum mapping quality for an alignment to be used.

Minimum base quality	Minimum base quality for a base to be considered.
Illumina-1.3 +encoding	Assume the quality is in the Illumina 1.3+ encoding.
Gap extension error	Phred-scaled gap extension sequencing error probability. Reducing INT leads to longer indels.
Homopolymer errors coefficient	Coefficient for modeling homopolymer errors. Given an l-long homopolymer run, the sequencing error of an indel of size s is modeled as INT*s/l.
No INDELs	Do not perform INDEL calling.
Max INDEL depth	Skip INDEL calling if the average per-sample depth is above INT.
Gap open error	Phred-scaled gap open sequencing error probability. Reducing INT leads to more indel calls.
List of platforms for indels	Comma dilimited list of platforms (determined by @RG-PL) from which indel candidates are obtained. It is recommended to collect indel candidates from sequencing technologies that have low indel error rate such as ILLUMINA.

Choose these parameters and click the Next button. The next page appears:

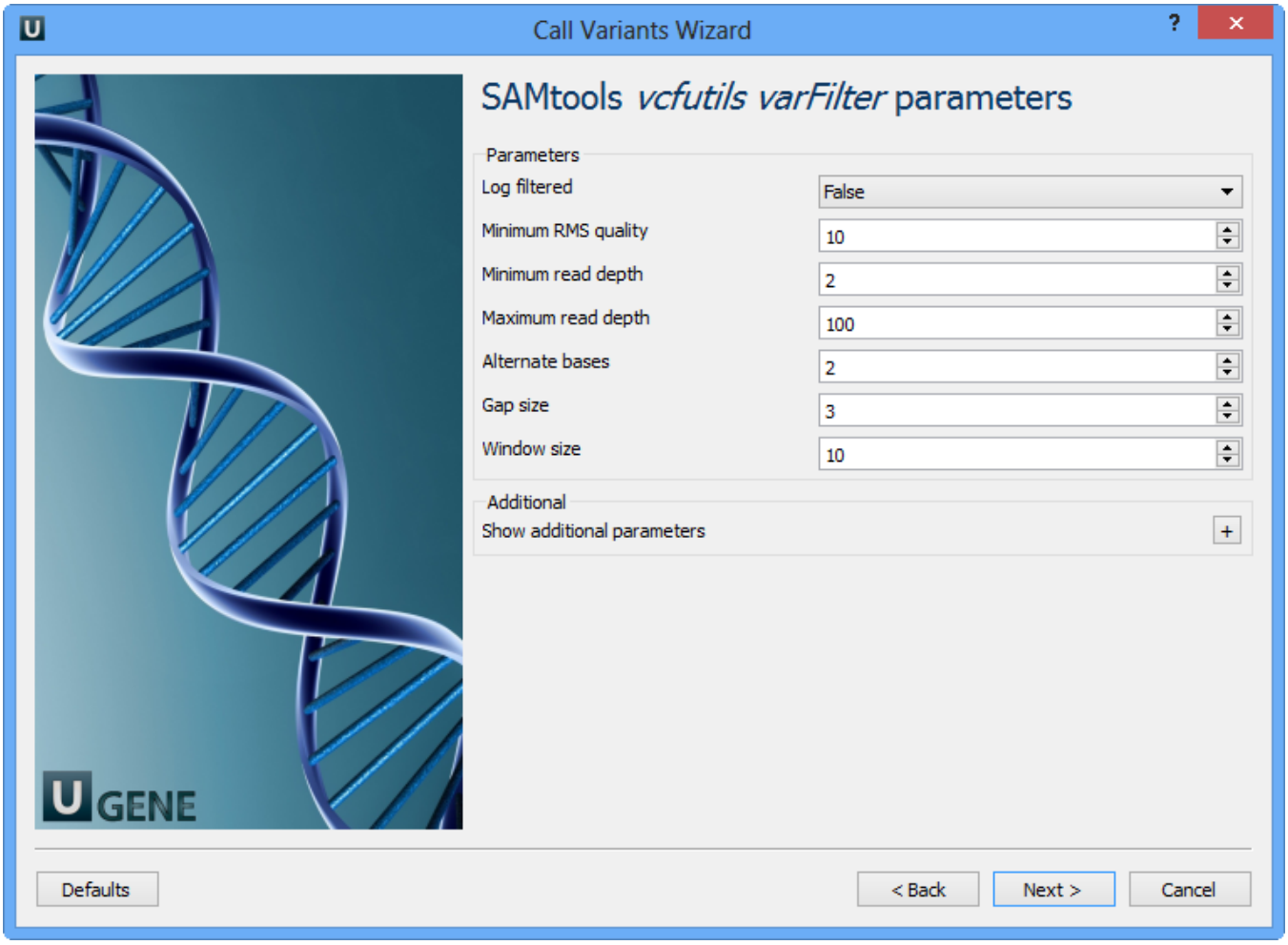


The next page allows one to configure SAMtools bcftools view utility parameters:

Retain all possible	Retain all possible alternate alleles at variant sites. By default, the view command discards unlikely alleles.

alternative	
Indicate PL	Indicate PL is generated by r921 or before (ordering is different).
No genotype information	Suppress all individual genotype information.
A/C/G/T only	Skip sites where the REF field is not A/C/G/T.
List of sites	List of sites at which information are outputted.
QCALL likelihood	Output the QCALL likelihood format.
List of samples	List of samples to use. The first column in the input gives the sample names and the second gives the ploidy, which can only be 1 or 2. When the 2nd column is absent, the sample ploidy is assumed to be 2. In the output, the ordering of samples will be identical to the one in FILE.
Min samples fraction	Skip loci where the fraction of samples covered by reads is below FLOAT.
Per-sample genotypes	Call per-sample genotypes at variant sites.
INDEL-to-SNP Ratio	Ratio of INDEL-to-SNP mutation rate.
Gap open error	Phred-scaled gap open sequencing error probability. Reducing INT leads to more indel calls.
Max P(ref|D)	A site is considered to be a variant if P(ref|D).
Pair/trio calling	Enable pair/trio calling. For trio calling, option -s is usually needed to be applied to configure the trio members and their ordering. In the file supplied to the option -s, the first sample must be the child, the second the father and the third the mother. The valid values of STR are “pair”, “trioauto”, “trioxd” and “trioxs”, where “pair” calls differences between two input samples, and “trioxd” (“trioxs”) specifies that the input is from the X chromosome non-PAR regions and the child is a female (male).
N group-1 samples	Number of group-1 samples. This option is used for dividing the samples into two groups for contrast SNP calling or association test. When this option is in use, the following VCF INFO will be outputted: PC2, PCHI2 and QCHI2.
N permutations	Number of permutations for association test (effective only with -1).
Max P(chi^2)	Only perform permutations for P(chi^2).

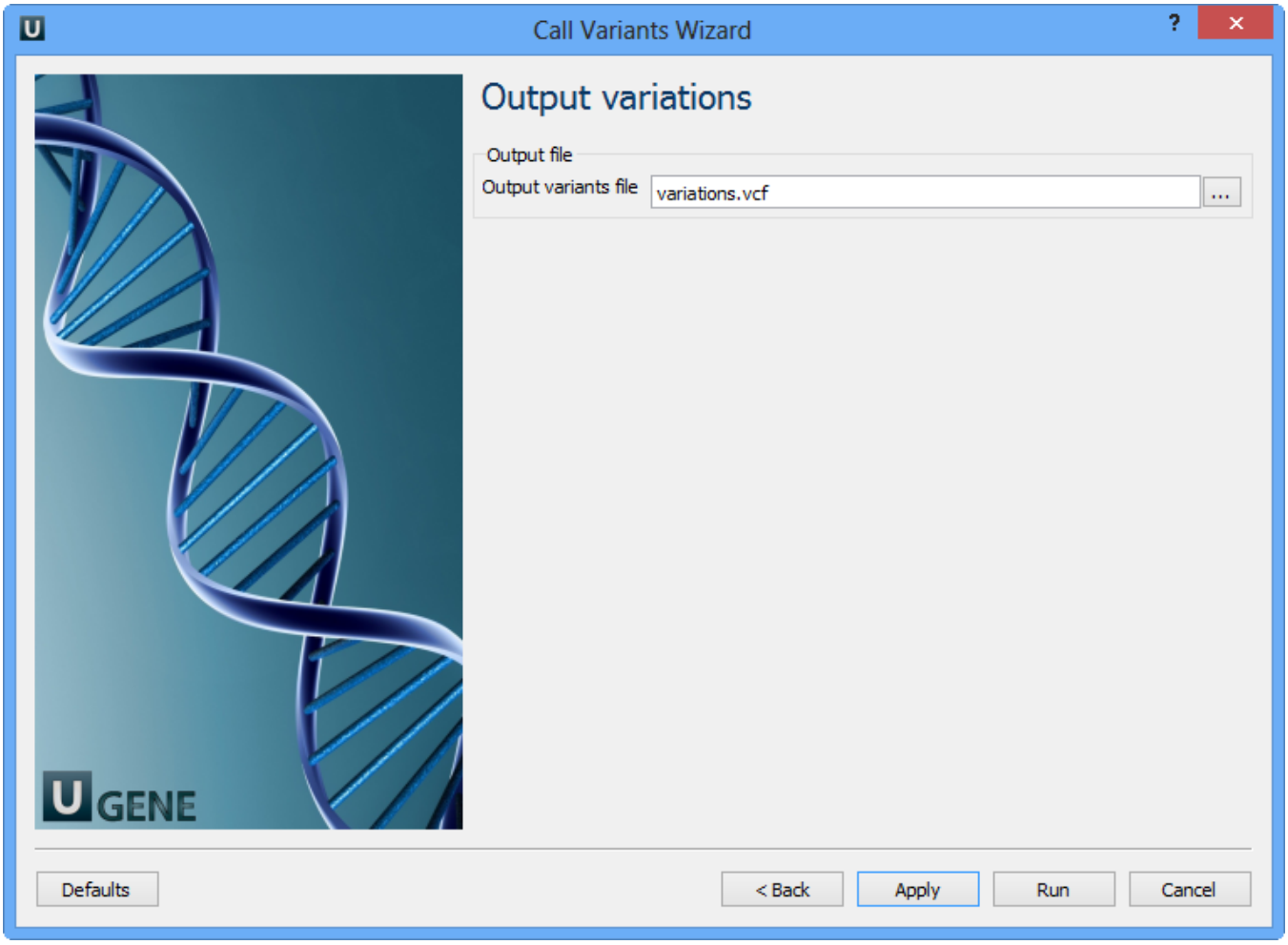
Choose these parameters and click the Next button. The next page of the wizard appears:



The next page allows one to configure SAMtools vcfutils parameters:

Log filtered	Print filtered variants into the log (varFilter) (-p).
Minimum RMS quality	Minimum RMS mapping quality for SNPs (varFilter) (-Q).
Minimum read depth	Minimum read depth (varFilter) (-d).
Maximum read depth	Maximum read depth (varFilter) (-D).
Alternate bases	Minimum number of alternate bases (varFilter) (-a).
Gap size	SNP within INT bp around a gap to be filtered (varFilter) (-w).
Window size	Window size for filtering adjacent gaps (varFilter) (-W).
Strand bias	Minimum P-value for strand bias (given PV4) (varFilter) (-1).
BaseQ bias	Minimum P-value for baseQ bias (varFilter) (-2).
MapQ bias	Minimum P-value for mapQ bias (varFilter) (-3).
End distance bias	Minimum P-value for end distance bias (varFilter) (-4).
HWE	Minimum P-value for HWE (plus F<0) (varFilter) (-e).

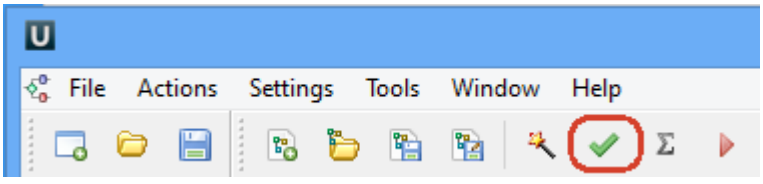
Choose these parameters and click the Next button. The last page of the wizard appears:



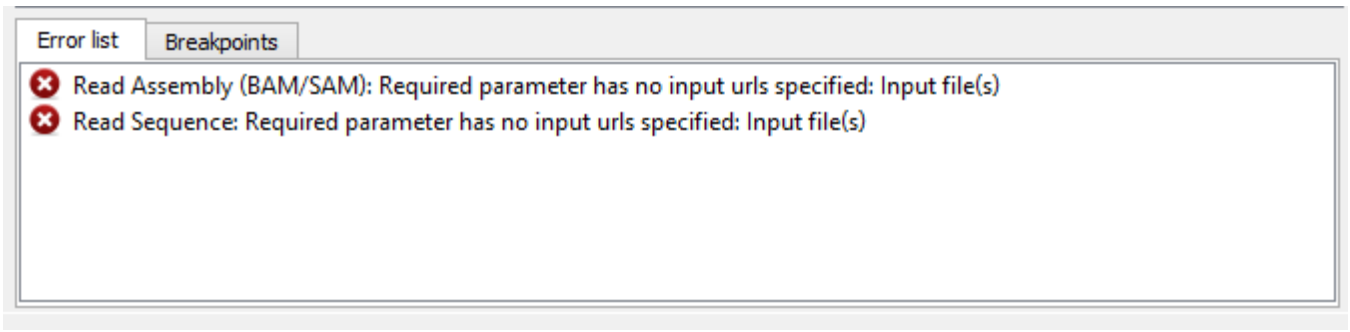
On this page you should select an output file. Set required parameters and click the Finish button.

Note that default button reverts all parameters to default settings.

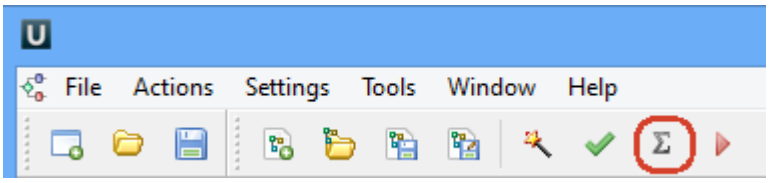
Now let's validate and run the workflow. To validate that the workflow is correct and all parameters are set properly click the Validate workflow button on the Workflow Designer toolbar:



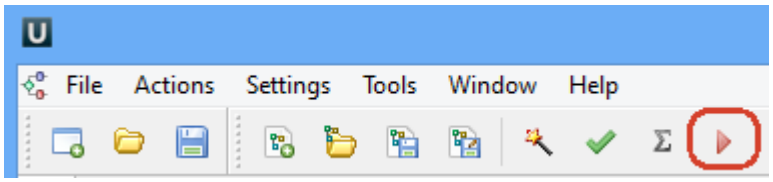
If there are some errors, they will be shown in the Error list at the bottom of the Workflow Designer window, for example:



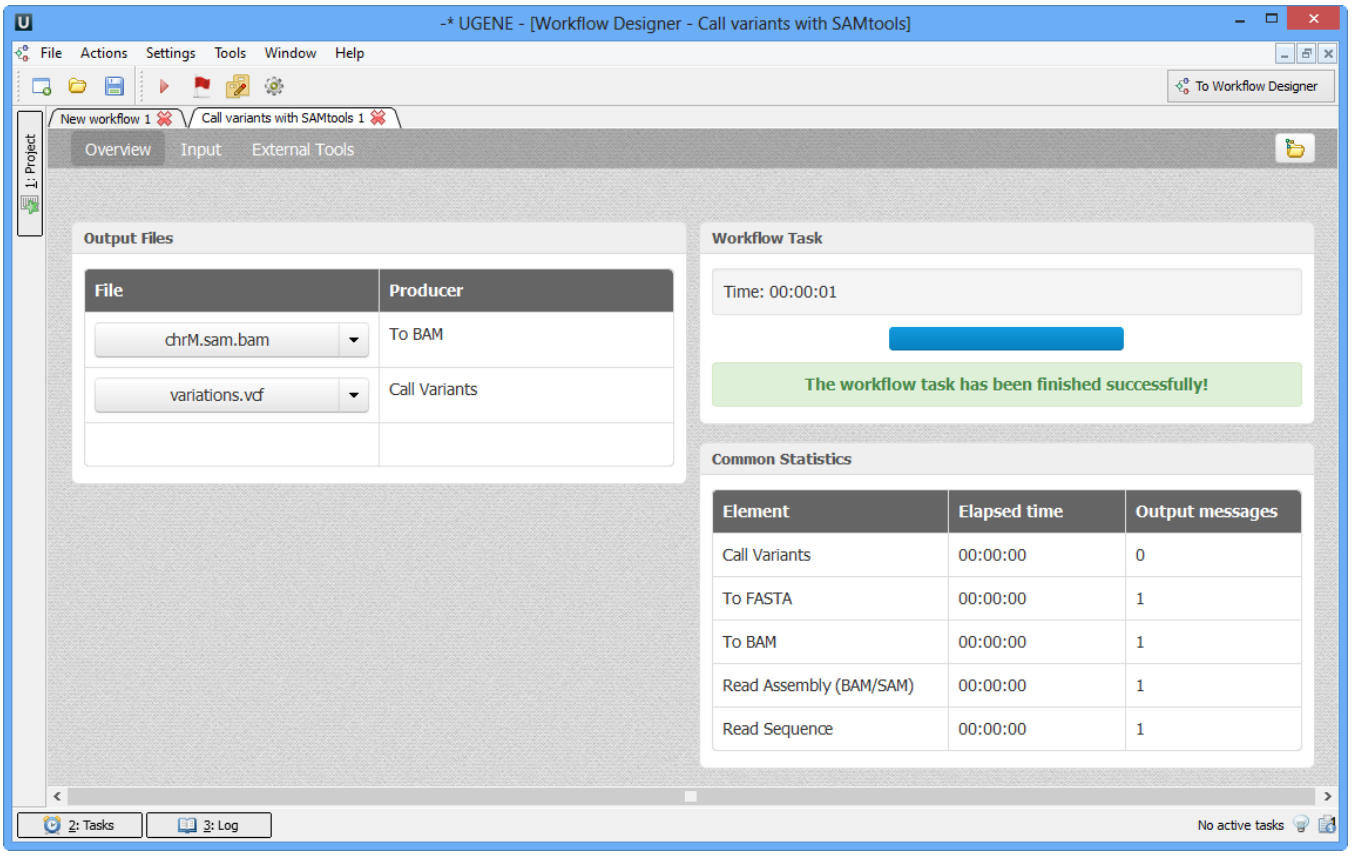
However, if you have set all the required parameters, then there shouldn't be errors. After that you can estimate the workflow. To run estimation click the *Es timate workflow* button:



To run a valid workflow, click the *Run workflow* button on the *Workflow Designer* toolbar:



As soon as the variants calling task is finished, a notification and dashboard will appear.



The dashboard will contain information about workflow: input and output files, all information about task. .

The work on this pipeline was supported by grant RUB1-31097-NO-12 from NIAID.