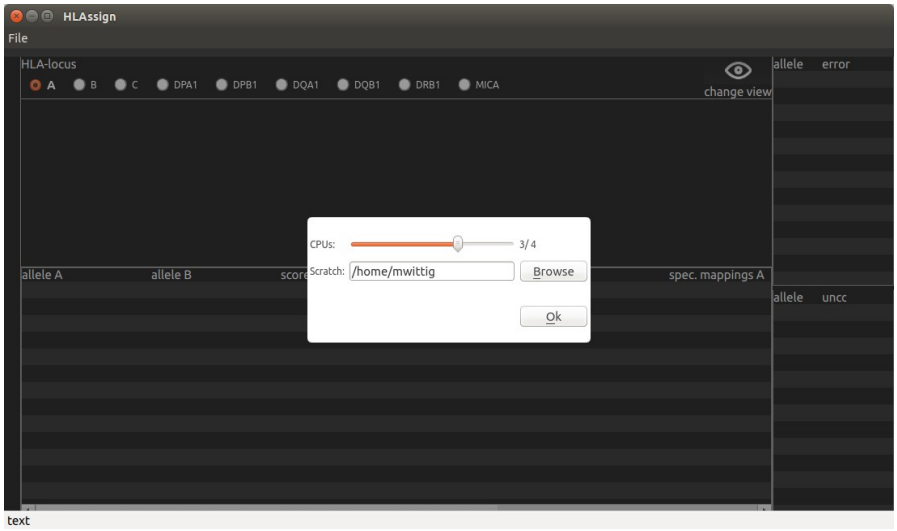
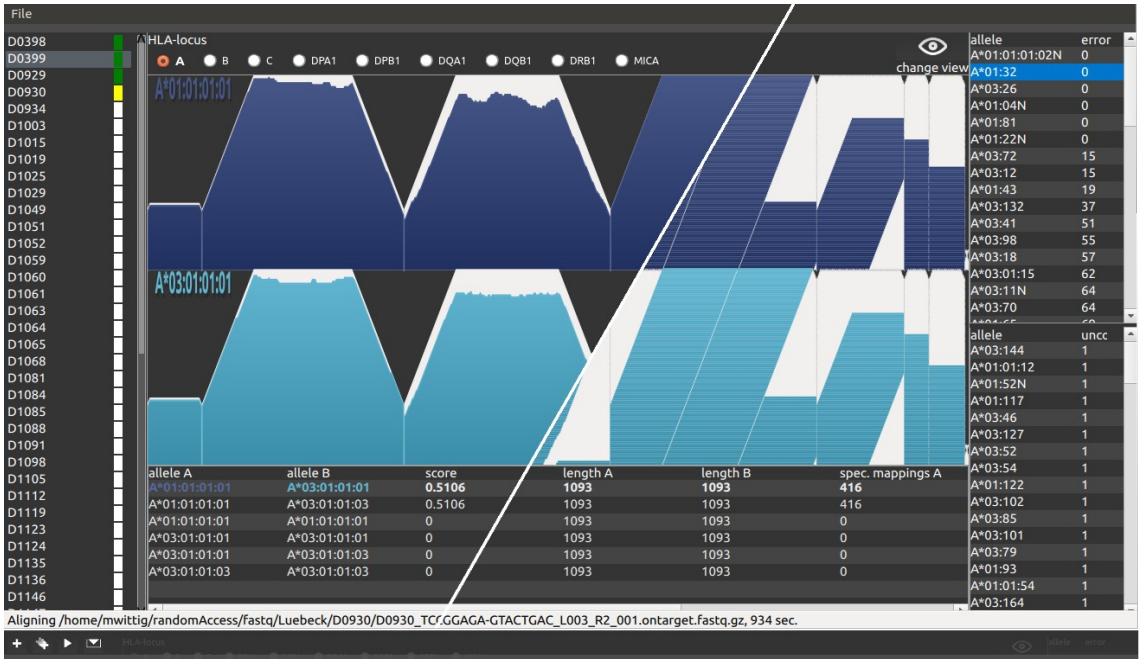
HLAssign

Starting the tool

On startup, the application opens a dialog where the user can enter the number of cores, that will be used for the analysis. The entry of the text field scratch points to the users home directory. This entry can be changed to point to an alternative folder, to which the user has write permission. The application will write some reference data to that folder.



The application window



For the above figure, we combined two screen shots of two different data views, separated by the white diagonal line. The left part of the graph shows the achieved single start point coverage (see

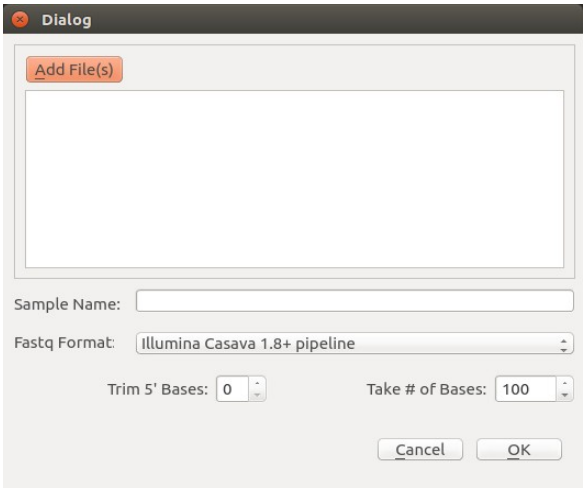
also Figure 1 for further details). The right part shows the corresponding single start point mappings of the same alignment. The different parts of the user interface are as follows: The list view on the left shows samples that are added for analysis. This sample list consists of the sample name followed by a colored rectangle. The colors are coding for the sample states, which are white (added for analysis), yellow (analysis running), green (analysis finished) and red (analysis failed). The middle part is divided into three sections. At the top, the user can select the locus for the selected sample. The “change view” button at the right switches between coverage and read view. Below, the user finds the NGS data visualization of the determined HLA type. The table at the bottom shows a sorted list of the different possible HLA types. The top most HLA type is the most likely determined by the algorithm. The user can change that order to manually correct possible errors. At the right side of the GUI the user finds two additional tables. The upper table contains a sorted list of alleles that failed the initial QC (see Methods). The column “error” shows the number of nucleotide positions, for which a QC failed value was calculated. The lower table shows the top 50 alleles that were not covered 100%. The table is sorted by the number of uncovered bases in ascending order. Alleles from these two tables can manually be included in the genotype calling. On the other hand, the user can move alleles from the allele calling to one of these tables. This allows for manual evaluation of the calling algorithm. Low covered alleles that failed the initial pre-filtering step can be added. Also degraded DNA, that is not 100% covered, can be analyzed. At the top of the application window, the user finds a small grey bar. When the mouse pointer hovers this area, a toolbox with buttons for sample adding, starting analysis and analysis report moves down (shown at the bottom, below the status bar).

Adding samples

Hover the gray bar at the upper window border to open the tool bar.

Add single sample

Push the + button to open the add sample dialog



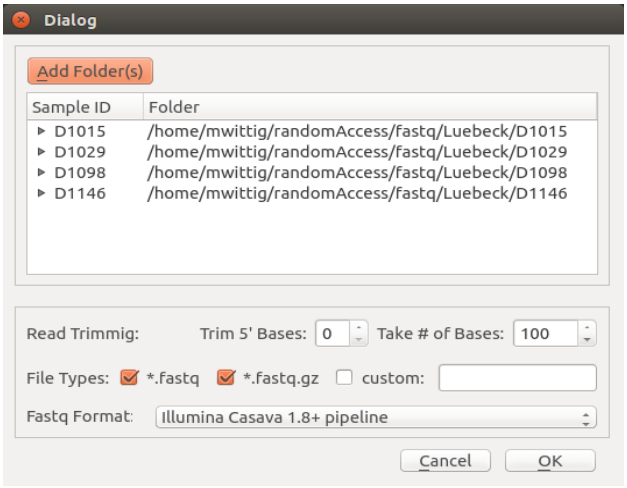
- Push the “Add File(s)” button to select the fastq files for that sample. You can add gzipped (*.fastq.gz) fastq files as well.
- Enter a sample name
- Select the fastq header format. The reference dataset is all in Casava 1.8+ format. If you are not sure which format your fastq files have, select unknown

- If necessary trim a defined number of bases at the 5' end of each read (default is 0)
- Select the number of bases of every read that should be taken for analysis (default is 100)
- Push “Ok”
- The sample should now appear at the left panel of the application window

We usually do not trim 5' bases. But if your sequencing run shows bad quality there (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) you maybe want to ignore a couple of bases. **Important**: The sum of “Trim 5' Bases” and “Take # of Bases” **must not** be longer than the actual read length really is. Please note that some of our reference data files are sequenced with read length 98. When analyzing these samples, set “Take # of Bases” to 98. Or if you would like to trim the first 5 bases of these reads set “Trim 5' Bases” to 5 and “Take # of Bases” to 93.

Add multiple samples

Push the “+++” button to add multiple samples for analysis.



- Push the “Add Folders(s)” button to select the Folders that contain the fastq files. The folder name is taken as the corresponding sample name
- If necessary trim a defined number of bases at the 5' end of each read (default is 0)
- Select the number of bases of every read that should be taken for analysis (default is 100)
- Select the file types of the fastq files that should be considered for analysis (do this before pushing the “Add Folder(s)” button)
- Select the fastq header format. The reference dataset is all in Casava 1.8+ format. If you are not sure which format your fastq files have, select unknown
- Push “Ok”
- The samples should now appear at the left panel of the application window

We usually do not trim 5' bases. But if your sequencing run shows bad quality there (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) you maybe want to ignore a couple of bases. **Important**: The sum of “Trim 5' Bases” and “Take # of Bases” **must not** be longer than the actual read length really is. Please note that some of our reference data files are sequenced with read length 98. When analyzing these samples, set “Take # of Bases” to 98. Or if you would like to trim the first 5 bases of these reads set “Trim 5' Bases” to 5 and “Take # of Bases” to 93. The supported file types are unpacked fastq files or gzipped fastq files. You can also set a custom file naming here,

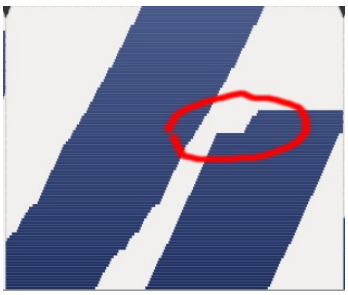
but the format of your files should be unpacked. Or, if packed, we do only support gzip and the fastq file names must have the .gz ending

Start analysis

Push the play button to start analysis. You can push it multiple times, but that will create a new alignment index (around 2.5GB) for every click. So this is only recommend for computers with big resources (e.g >= 32gb RAM and >= 8 cores). A typical system requirement for a “single click”-analysis is 4 cores and 8 GB RAM.
The sample panel at the left side of the application panel shows the samples that were loaded for analysis. Each entry consists of a sample name and a colored rectangle. The colors are coding for the sample states, which are white (added for analysis), yellow (analysis running), green (analysis finished) and red (analysis failed).

Screen results

The envelope icon ( ) at the tool bar opens a table with an entry for each locus of every sample. This table gives a result overview of the entire sample set. To screen the results and raw data in detail you have to go back to the detailed view (click the envelope icon again).
To be able to validate the results, it is necessary to understand the basics of the data analysis. It consists of two steps.
1. During the first step the alignments of each reference allele are verified and the allele is either filtered out or included in the second step of the analysis. If an allele is not covered 100% it is filtered out. We sort these alleles by the number of uncovered bases in an ascending order and list the top 50 of those at the lower table on the right site of the application window. In analogy to Wang *et al.* 2012 (PMID: 22589303) the algorithm also tries to identify alleles that have an unbalanced number of central/non-central reads for any nucleotide. The next three pictures show the single start point coverage of three exons where such an effect could be identified, so that these alleles are filtered out.



Alleles that are filtered out with respect to an unbalanced read coverage can be found at the upper right table.
2. The second step of the analysis generates hypothetical genotypes of all possible allele combinations with all alleles that passed the before described filtering step. A score is calculated and the highest score describes the most likely genotype.
The second step of the analysis usually works very well. The typical scenario of an manual correction is, that the user identifies an allele that passed the first filtering step even though it shows an unbalanced distribution of central and non-central reads.

A well defined heterozygous genotype has:
- a score close to 1, or at least much higher than the second best score
- a “read eq.” close to 1
- the sum of “spec. mappings A” and “spec. mappings B” should be as high as possible

A well defined homozygous genotype has:
- either all other alleles filtered out during step one
- or only heterozygous genotypes with a low “read eq.”

To get a better understanding how these criteria might be interpreted we recommend to analyze a couple of our reference samples and compare the results with the Supplementary Table S2 of our publication (PMID: UNDER_REVIEW).

Detailed data view

Double click a sample at the left panel that has the status “analysis finished” (green). At the upper part of the window select the locus you are interested in. The top-most entry of the center table is the determined genotype for that locus. Click the entry to visualize the raw data. Two data views are available. The coverage or, when clicking “change view”, the single start point mappings.
The user can remove alleles from the genotype table by right clicking the table cell with the corresponding allele entry. A context menu opens and the user can click “remove allele_name”. In the same way, the user can add samples from the two tables at the right side to the genotype table.