



Omixon handbook

Omixon HLA Twin CE 4.2.2

07/31/2020

1	Company information.....	8
2	Genotyping dashboard	9
2.1	General overview	9
2.2	Top panel.....	9
2.2.1	Top row.....	9
	Memory Usage Widget Panel	9
	Typer Scheduler	9
	Event Log	9
2.2.2	Middle row	10
2.2.3	Bottom row	10
2.3	Genotyping and other functions	10
2.3.1	Typing and data analysis options	10
2.3.2	Typing and data analysis options	11
2.3.3	File browser functions	11
2.3.4	Typer Manager	11
2.3.5	Application Settings.....	11
2.4	File browser	11
2.4.1	Navigation in the file browser	11
2.4.2	Samples and analyses.....	12
2.5	Traffic lights.....	12
2.6	Context menu	12
2.7	Simple Genotyping	12
2.8	Advanced Genotyping.....	13
2.8.1	Introduction	13
2.8.2	Protocol	13
2.8.3	Advanced Options	13
2.9	Genotyping Reanalysis	15
2.10	Export results	16
2.10.1	General information.....	16
2.10.2	HML export	16
	General information.....	16
	HML export wizard	16
2.10.3	PDF export	16

2.10.4 JSON export	17
Export options.....	17
Export format and content	18
2.10.5 HPRIM export.....	21
Filter analysis results	21
Select analysis results.....	22
Export configuration	22
2.11 Filtering.....	22
2.11.1 Access	22
2.11.2 Usage	22
Targeted genes.....	22
Filtering options.....	23
3 Genotyping analysis result	24
3.1 Introduction	24
3.1.1 Commenting.....	24
3.2 Top panel.....	24
3.2.1 Top row.....	24
3.2.2 Middle row	24
3.2.3 Bottom row	25
3.3 Interpreting Results	25
3.3.1 Troubleshooting missing results.....	25
3.4 Filtering Alleles.....	26
3.5 Traffic Lights.....	26
3.5.1 Quality control light	27
3.6 Assigning Alleles.....	27
3.7 Export table	27
3.8 Linkage disequilibrium details.....	28
3.8.1 Overview.....	28
3.8.2 Case by case guide	28
3.8.3 Notes.....	29
3.8.4 DQA1, DQB1 allele dropout warning	29
4 Genotyping sample result	30
4.1 Top panel.....	30

4.1.1	Top row.....	30
4.1.2	Middle row.....	30
4.1.3	Bottom row	30
4.2	Available functions.....	31
4.2.1	Opening the browser	31
4.2.2	Detailed genotyping information.....	31
4.2.3	Opening the browser	31
4.2.4	Detailed genotyping information.....	31
4.2.5	Customizing displayed results	31
4.2.6	Assigning alleles.....	31
4.2.7	Commenting.....	31
4.2.8	PIRCHE® epitope matching.....	32
4.3	Genotype	32
4.3.1	Tree view	32
4.3.2	Troubleshooting missing results.....	32
4.3.3	Adding alleles manually.....	32
4.3.4	Removing additional alleles.....	33
4.4	Quality Control	33
4.5	Data statistics.....	34
4.5.1	Overview.....	34
4.5.2	Allele imbalance.....	35
4.5.3	Fragment size	35
4.5.4	Read quality	35
4.5.5	Troubleshooting missing results.....	32
4.6	Genotyping mismatch result.....	36
4.7	Genotyping result novelties	36
4.8	Allele frequency information.....	36
5	Gene Browser	37
5.1	Introduction	37
5.2	Top panel.....	38
5.2.1	Top row.....	38
5.2.2	Middle row	38
5.2.3	Bottom row	38
5.3	Settings and Functions	38

5.3.1	Filtering the Allele Candidates	38
5.3.2	Exporting Sequences	39
	Export Gene Browser Tracks	39
	Export Novel Sequence.....	39
5.3.3	Exporting Sequences	39
5.3.4	Zooming	39
5.3.5	Jumping around.....	39
5.3.6	Changing short read track settings	39
5.3.7	Track management.....	39
5.3.8	Display setup	39
5.3.9	Exporting Gene browser data	40
5.3.10	Rotating the browser	40
5.4	Context menu	40
5.5	Metadata Panel	40
5.5.1	Metadata Panel functions in the Gene Browser:	41
5.6	Shortcuts	41
5.7	Manage tracks	41
6	Settings dashboard.....	42
6.1	General information.....	42
6.2	Sidebar.....	42
6.2.1	General	42
6.2.2	Database.....	42
6.2.3	Administration	42
6.2.4	Automation.....	42
6.2.5	Export Settings.....	42
6.2.6	Screen Settings	42
6.3	Sample and Result Folders.....	43
6.4	Analysis Protocols.....	43
6.4.1	Configure protocol	43
6.4.2	Create protocol	43
	Introduction	43
	General	43
	Sample data type	44
	Analysis options	44

Sample data type	44
Analysis options	44
6.5 Install New Database	45
6.5.1 Download problems	45
6.6 Select Active Database.....	45
6.7 Configure Database Extensions	45
6.8 User Management.....	46
6.9 Upload Licence.....	46
6.10 Configure Automatic Analysis	47
6.11 Gene Browser	47
6.11.1 Display configuration.....	47
6.11.2 Color setup	47
6.11.3 Data options	47
6.11.4 Metadata options.....	47
6.11.5 Advanced settings.....	47
7 Detailed automation guide.....	48
7.1 Overview.....	48
7.2 Configuration	48
7.3 Management	49
7.4 Logging	49
8 Notifications	50
8.1 IMGT Database	50
8.2 License Expiry.....	50
8.3 Extended allele notification	50
8.4 Deviating from the Factory default protocol	50
9 HTTPS support	51
10 Application performance tuning	52
11 Omixon database relocation	53
12 Fastq and BAM filter tool	54
12.1 Installation	54
12.2 Filter Tool execution	55

12.3	NextSeq Filter Tool execution	55
12.4	Custom configuration	56
12.5	Logging	56
13	PacBio standalone tools	57
13.1	HDF5 file converter	57
13.1.1	Installation	57
13.1.2	Converter execution.....	57
13.2	Demultiplexer.....	57
13.2.1	Installation	57
13.2.2	Demultiplexer execution	57
13.3	Barcode file format	58
13.3.1	Custom configuration	58
14	Glossary	60
15	List of shortcuts.....	62
15.1	Generic shortcuts	62
15.2	Genotyping dashboard	62
15.3	Genotyping analysis result	62
15.4	Genotyping sample result	62
15.5	Gene browser	62
16	Acknowledgements	64
16.1	Collaborators.....	64
16.2	Third party tools and databases	64
16.3	Citations	64



1 Company information

This product is manufactured by Omixon Biocomputing Ltd.

Address:

H-1117 Budapest
Fehérvári út 50-52.
Hungary, EU

Website: <http://www.omixon.com>

Technical contact: support@omixon.com¹

Sales contact: sales@omixon.com²

¹ <mailto:support@omixon.com>

² <mailto:sales@omixon.com>

2 Genotyping dashboard

2.1 General overview

This is the home dashboard of the software. All Genotyping functions are available from here.

The dashboard consists of the following subscreens:

- Top panel: this contains all the main functions and some high level information about the current folder and sample. It also shows information about the current user and active database and provides some navigation functions.
- File browser: this part of the screen can be used for navigation between all accessible folders.
- Bottom panel: analysis configuration details are displayed in the bottom of the screen when hovering over or selecting an analysis result

2.2 Top panel

On the top of the screen you can see three rows of functions.

2.2.1 Top row

On the left, you can find the *Toggle fullscreen* button which hides the less important parts of the current screen, while on the right you can see the following things:

- the ID of the current user,
- the memory usage widget panel,
- the status panel of the Typer Scheduler,
- the status panel of the Event Log,
- the welcome tutorial button,
- the logout button,
- and the exit button.

Memory Usage Widget Panel

In case of the desktop version, the actual memory usage can be observed at the top right corner of the screen. This displays the amount of currently used memory compared to the allowed maximum. Note, that the software might be allocating more memory from the operating system compared to the reported actual memory usage to make further computations faster by keeping some previously allocated memory pieces for later use, thus making the allocation of new memory unnecessary during further computations. By clicking on the memory report widget the software tries to free up as much memory as possible on a best effort basis: this is only an attempt, there are many technical factors which might block returning all memory to the operating system. It is also important to mention that by explicitly asking to free up memory the subsequent computations might become slightly slower, since the option of reusing certain structures allocated before becomes impossible. Only use this option when all planned analyses have been finished and you would like to use other programs more intensively in parallel while investigating the analysis results.

Typer Scheduler

In this window you can see all genotyping tasks listed by processing order. The tasks on top of the list will be processed first. You can change the processing order by moving single tasks up or down in the queue using the *Move* buttons on top of the screen. It is also possible to stop or remove items from the list using the inline *Stop* and *Remove* buttons.

Event Log

The event log contains data on all past actions of the following types:

- Genotyping
- Result approval (workflow state changes)
- Result file export
- Result table export

- Reference database import

You have two tabs to choose from: *Tasks* and *Events*. The information displayed on these two tab panels is basically the same, only the structure of the information is different.

On the *Events* tab one entry belongs to a single event while on the *Tasks* tab events are grouped in tasks. For example: you can start analyses in a batch. A batch analysis would have a single task which starts when you trigger the analysis and ends when all analyses in the batch are finished. The same batch analysis will have many state changes during its lifetime: each and every analysis in the batch will have its own status changes from CREATED to RUNNING to SUCCEEDED, CANCELLED, FAILED or ABORTED.

Note: Approval events do not have tasks, they can only be viewed under the *Events* tab.

Use the task type filters on top of the screen to display or hide certain types of tasks. It is also possible to show tasks and events from within a specific time period, using the *Select dates* function.

2.2.2 Middle row

There are four navigation buttons that are available on every screen of the tool. These general navigation buttons work very similar to navigation functions available in most widely used internet browsers:

- The *Back* button will take you to the previous screen.
- The *Forward* button will take you to the next screen.
- The *Up* button will move you one level up in the application.
- The *Home* button will take you to the starting screen (i.e. the Genotyping dashboard).

Right from the navigation button, in the information panel, you can find relevant high-level information about the current screen. On the *Genotyping dashboard*, you can see the release date and version of the active IMGT database.

On the far right hand side of the middle row, you can find the bookmark and context specific help buttons. The help tool shows relevant usage information and tips about functions and data available on the current screen of the application and can also be opened by pressing F1.

You can search the Help for any expression by using the Magnifier glass icon displayed on the popup window. Relevant hits are displayed in the left panel where you can select them and open them in the right hand reading pane. You can close the search by clicking the X displayed beside the search field.

2.2.3 Bottom row

The bottom row of the top panel contains a series of buttons which contain the main functions available on the screen.

2.3 Genotyping and other functions

2.3.1 Typing and data analysis options

- **Simple Genotyping** - This function starts the standard analysis with the Factory default protocol which is fully configured for the Omixon Holotype assay and is the only typing function available for all users.
- **Advanced Genotyping** - This function starts the Advanced Genotyping wizard which is only available for users with superuser privileges. This advanced wizard allows the setting of several advanced parameters for the Genotyping algorithms.
- **View Results** - Leads to the more detailed Genotyping analysis result screen which is suitable for working with multiple genotyping results.
- **View Details** - Opens the Genotyping sample result screen where locus level results, detailed QC metrics and data statistics can be reviewed for the selected genotyping result.
- **Export Results** - By pressing this button, the Export results wizard can be opened. Using this wizard, genotyping results can be exported in the standard HML, PDF or JSON format.
- **Re-Analyse** - This function allows the re-analysis of previously typed samples.

Note

When you create a new protocol and set is as default, after pressing the Simple Genotyping or the Analysis button, you will be warned about the deviation from the factory default protocol. The reason for the warning is to avoid clinical usage of results which are for research use only.



2.3.2 Typing and data analysis options

- **Simple Genotyping** - This function starts the standard analysis with the Factory default protocol which is fully configured for the Omixon Holotype assay and is the only typing function available for all users.
- **Advanced Genotyping** - This function starts the Advanced Genotyping wizard which is only available for users with superuser privileges. This advanced wizard allows the setting of several advanced parameters for the Genotyping algorithms.
- **View Results** - Leads to the more detailed Genotyping analysis result screen which is suitable for working with multiple genotyping results.
- **View Details** - Opens the Genotyping sample result screen where locus level results, detailed QC metrics and data statistics can be reviewed for the selected genotyping result.
- **Export Results** - By pressing this button, the Export results wizard can be opened. Using this wizard, genotyping results can be exported in the standard HML, PDF or JSON format.
- **Re-Analyse** - This function allows the re-analysis of previously typed samples.
- **Filter data** - This function is only enabled for large files (whole exome and whole genome sequencing data). By pressing this button, a filtering process is started which aligns all the reads in the selected sample against the active IMGT database and creates a new, filtered sample in FASTQ format which is suitable for genotyping. *Note, that a default analysis is not suitable for genotyping from the filtered data!* The following advanced settings are suggested: all genes should be set as targeted genes, all reads should be processed and quality based subsampling should be turned off.

2.3.3 File browser functions

- **Sample Filter** - Toggles the file browser between showing and hiding samples.
- **Analysis Filter** - Toggles the file browser between showing all analyses, showing the most recent analysis per sample and hiding all analyses.
- **Analysis State Filter** - Toggles the file browser between showing all analyses or analyses that are in progress, waiting for approval or approved.
- **Filter Files** - Opens the *Filter analyses and samples* wizard which can be used for filtering the samples/analyses in the current folder using simple string matching, string matching with wildcards or regular expressions.
- **Sort Files** - Sorts samples and analyses by name.

2.3.4 Typer Manager

This function is only available in distributed server configurations.

In a distributed server configuration, if you would like to genotype samples, you have to deploy one or more typer nodes through this dialogue. The typer node list contains the typer node configurations previously set in the *typer.conf* file, which is located in the server installation directory. For each typer node you have the option to deploy or undeploy it. The number of deployed typer nodes determines the number of genotyping jobs that can be run simultaneously.

2.3.5 Application Settings

This function opens the Settings dashboard, where tool-wide settings and administrative functions can be found.

2.4 File browser

2.4.1 Navigation in the file browser

The file browser works very similarly to a high number of well known file browsers. Navigation is single click based: folders can be entered using a single left mouse click. Moving up to the parent folder can be achieved by left clicking on the top ".." folder or by using the folder breadcrumbs at the top of the file browser.

Folders without read permission are shown in red color.

2.4.2 Samples and analyses

Only short read files (in the supported file formats) and genotyping results (with a htr extension) are shown in the file browser. Paired-end or mate-pair fastq files are automatically paired and the sample is represented by the R1 file within the file browser to avoid redundancies and make navigation easier. For each sample, the pairing status and total size of the short read files are shown. On the right hand side of each sample row you can find the *Results* and *Details* buttons which provide shortcuts for the *Genotyping analysis result* and *Genotyping sample result* views, respectively.

Those analyses which were produced using Custom analysis - typing parameters have been altered from the default protocol - are displayed with red font.

The same markup is used for those results where audit details are not available. It is possible to lose this data if the database had been deleted or damaged due to an error, or if the result had been produced with a previous version of the software which allowed removing these details.

2.5 Traffic lights

Sample level results are annotated using a "traffic light" system with the following output:

- (green): all QC tests gave passed or info results and the genotyping results of the two algorithms were fully concordant.
- (red): at least one of the QC measures failed and/or the concordance between the two algorithms was under two fields.
- (yellow): in all other cases, a yellow light is shown.

Sample level results are annotated using a "traffic light" system with the following output:

- (green): all QC tests gave passed or info results.
- (red): at least one of the QC measures failed.
- (yellow): in all other cases, a yellow light is shown.

2.6 Context menu

Like in most parts of the tool, a context specific menu, containing additional functions can be opened using the right button of the mouse (or touchpad). On the *Genotyping dashboard* the following functions are available in this right-click menu:

- *Toggle select all*: selects or deselects all files in the current browser.
- *Filter files*: opens the file filter wizard.
- *Reduce/Expand file name*: shortens sample and analysis names using regular expressions.
- *View selected results*: takes you to the *HLA Typing analysis* result dashboard of the selected analyses.
- *Show protocol*: shows the parameters used for the selected analysis.
- *Save as protocol*: saves the parameters used for the selected analysis
- *Delete selected items*: deletes the selected items.
- *Move selected items*: moves selected samples and analyses to a new location on the file system.
- *Rename selected result*: allows the user to rename the selected result file.
- *Copy selected result original name*: copies the full name of the analysis to the clipboard.
- *Create new folder*: creates new folder on the file system.
- *Cut/Copy/Paste*: cuts, copies or pastes selected samples and analyses.

2.7 Simple Genotyping

This function allows you to analyse data produced with the Omixon Holotype assay. All parameters are pre-set and fine-tuned for this assay. Datatype is autodetected at the beginning of the typing and the amount of data processed is adjusted accordingly. For running genotyping with other settings, please use the *Advanced Genotyping* function. You need to select files for the input. This function will allow you to *process multiple samples at the same time*, i.e. multiple pairs of fastq (or fastq.gz) files. The two input files in a file pair are assumed to have

the exact matching reads, in the exact same order (there is currently no built-in check for this). A background task will be started, and once it's finished a new item will appear in the main list on the *Genotyping dashboard* where you will be able to see the result.

Note

When you create a new protocol and set is as default, after pressing the Simple Genotyping or the Analysis button, you will be warned about the deviation from the factory default protocol. The reason for the warning is to avoid clinical usage of results which are for research use only.

This function allows you to analyse data with the Factory default typing protocol. All parameters are pre-set and fine-tuned. Datatype is autodetected at the beginning of the typing and the amount of data processed is adjusted, but note, that this typing option is not suitable for analyzing whole exome or whole genome data. For running genotyping with other settings, please use the *Advanced Genotyping* function. You need to select files for the input. This function will allow you to *process multiple samples at the same time*, i.e. multiple pairs of fastq (or fastq.gz) files. The two input files in a file pair are assumed to have the exact matching reads, in the exact same order (there is currently no built-in check for this). A background task will be started, and once it's finished a new item will appear in the main list on the *Genotyping dashboard* where you will be able to see the result.

2.8 Advanced Genotyping

2.8.1 Introduction

This function is available from the Genotyping dashboard.

When an analysis is started with the Advanced Genotyping option a warning will be displayed to alarm the user about potential deviation from the Factory Default Protocol. The reason for the warning is to avoid clinical usage of results which are for research use only.

First you need to select one or multiple samples to analyse. This function allows you to process multiple samples at the same time, that is, multiple pairs of fastq (or fastq.gz) files. The two input files in a file pair are assumed to have the exact matching reads in the exact same order (currently there is no in-built check for this). Before the analysis you also need to select which sequencer created the short read data. After starting the analysis, a background task will be started. When the task is finished, the result will appear in the main list on the *Genotyping dashboard*.

2.8.2 Protocol

Protocols are stored sets of Genotyping analysis parameters. Advanced Genotyping allows you to select stored protocols. If a particular set of parameters work well for your samples, you can save these as a protocol and apply it to other samples later. Selecting an existing protocol can speed up the usage of the Advanced Genotyping wizard, as most of the parameters are already preset in the protocol. Protocols can be copied, edited, removed and created at any time.

2.8.3 Advanced Options

Analysis options

- *Typing method*: You can set the method used for the genotyping.
- *Maximum pairs processed per locus*: Twin uses locus based subsampling to optimize results. This value determines how many reads will be used for each locus. The datatype is assessed at the beginning of each typing and the amount of reads processed is adjusted accordingly. Note, that only reads longer than 75 bases are selected during subsampling (for paired data, both reads in the pair have to be over this minimum length threshold). For samples produced with the Holotype assay and other targeted data sets the default 4000 read pairs per locus works well. In case you run into problematic samples increasing the read count can help resolve some of these issues.
- *Disable quality based subsampling*: If this checkbox is checked, subsampling will be based on the per locus read count and the number and type of targeted loci. This option is used for generating mappability statistics.
- *Ignore rare alleles*: Using this option, very rare alleles will be ignored in the analysis and will not be included or reported. If this checkbox is unchecked, rare alleles will be marked in the results. If this option is greyed out, you need to install a new database in order to import the rare alleles.

- **SG analysis mode:** When using the Statistical genotyping method, you can determine which parts of the genetic information will be used for the analysis. The **Exons only** option uses exonic information only. The **Whole gene** option allows the algorithm to use information from the intronic and UTR parts of the locus as well. The regions will be displayed in the Gene Browser according to this selection.
- **Treat regions separately:** Using this option, reads are aligned to each region separately. This is reflected in the alignment in the Gene Browser.
- **Alternative method execution:** You can set the alternative method execution based on the QC report status or it can be switched off. Using this option the analysis performance can be improved with high quality samples. **Always** option runs the alternative typing method always. **QC inspect/investigate/failed** option runs alternative typing method for those loci where QC reported inspect/investigate/failed status and for loci with novel alleles regardless of the QC status (this is the default configuration). **QC failed** option runs alternative typing method for those loci where QC failed and for loci with novel alleles regardless of the QC status. **Never** option turns off the alternative typing method completely.

Note: When using the Twin typing method, statistical genotyping is never run for the locus HLA-DRB3.

- **Novel allele detection:** This feature detects SNPs and indels in the short read allele alignments. Based on the detected variants and the original reference sequence of the allele, novel allele sequences are generated. The name of a novel allele candidate consists of the name of the original allele plus a hashtag and a number. For example, HLA-A*01:01:01:01#1 is a novel allele generated based on the sequence of HLA-A*01:01:01:01 and one or more SNPs found in the short read alignment of this allele. Novel alleles have a double reference track in the Gene Browser, which means that both the reference of the novel allele candidate (Novel ref) and the reference sequence of the related allele (Rel ref) - i.e. the sequence of the allele the novel allele candidate originated from - are shown. Novel positions are accepted only when at least 90% of the reads and at least 10-fold coverage depth support the novel reference value. Note that in many cases multiple similar novel allele candidates are generated from different alleles. There are three versions of the novel allele detection settings:
 - not running novel allele detection
 - running novel allele detection only when exon mismatches are present (note: this function will try to resolve both exon and intron novelties)
 - running novel allele detection if any mismatches are present (whole gene novel allele detection).
- **Pipeline execution mode:** This feature allows the user to change the behavior of the analysis pipeline. The main difference between the two is the analysis time performance. **Legacy mode** will behave similarly to older versions of HLA Twin. In comparison, **Fast mode** achieves better analysis times with the help of optimization, and also the result files will be significantly smaller. Thus the long-term storage requirements became lower.

Gene targeting

You can select the targeted gene family and loci set. The predefined sets are based on the different Holotype HLA Kit configurations (5, 7, 11-loci) and there is an option to set custom loci for the genotyping, which can be useful for reanalysis purposes. When the 'Auto' option is selected, gene family and targeted loci selection is disabled since these parameters will be determined from the sample during the analysis.

Analysis options

- **Maximum pairs processed per locus:** Twin uses locus based subsampling to optimize results. This value determines how many reads will be used for each locus. The datatype is assessed at the beginning of each typing and the amount of reads processed is adjusted accordingly. Note, that only reads longer than 75 bases are selected during subsampling (for paired data, both reads in the pair have to be over this minimum length threshold). You can experiment with this value to find the best match for your data. For samples produced with the Holotype assay and other targeted data sets the default 4000 read pairs per locus works well. In case you run into problematic samples increasing the read count can help resolve some of these issues.
- **Disable quality based subsampling:** If this checkbox is checked, subsampling will be based on the per locus read count and the number and type of targeted loci. This option is used for generating mappability statistics.
- **Ignore rare alleles:** Using this option, very rare alleles will be ignored in the analysis and will not be included or reported. If this checkbox is unchecked, rare alleles will be marked in the results. If this option is greyed out, you need to install a new database in order to import the rare alleles.
- **Alternative method execution:** You can set the alternative method execution based on the QC report status or it can be switched off. Using this option the analysis performance can be improved with high quality samples. **Always** option runs the alternative typing method always. **QC inspect/investigate/failed** option runs alternative typing method for those loci where QC reported inspect/investigate/failed status and for loci with novel alleles regardless of the QC status (this is the default configuration). **QC failed** option runs alternative typing method for those loci where QC failed and for loci with novel alleles regardless of the QC status. **Never** option turns off the alternative typing method completely.

Note: When using the Twin typing method, statistical genotyping is never run for the locus HLA-DRB3.

- **Novel allele detection:** This feature detects SNPs and indels in the short read allele alignments. Based on the detected variants and the original reference sequence of the allele, novel allele sequences are generated. The name of a novel allele candidate consists of the name of the original allele plus a hashtag and a number. For example, HLA-A*01:01:01:01#1 is a novel allele generated based on the sequence of HLA-A*01:01:01:01 and one or more SNPs found in the short read alignment of this allele. Novel alleles have a double

reference track in the Gene Browser, which means that both the reference of the novel allele candidate (Novel ref) and the reference sequence of the related allele (Rel ref) - i.e. the sequence of the allele the novel allele candidate originated from - are shown. Novel positions are accepted only when at least 90% of the reads and at least 10-fold coverage depth support the novel reference value. Note that in many cases multiple similar novel allele candidates are generated from different alleles. There are three versions of the novel allele detection settings:

- not running novel allele detection
 - running novel allele detection only when exon mismatches are present (note: this function will try to resolve both exon and intron novelties)
 - running novel allele detection if any mismatches are present (whole gene novel allele detection).
- *Pipeline execution mode*: This feature allows the user to change the behavior of the analysis pipeline. The main difference between the two is the analysis time performance. **Legacy mode** will behave similarly to older versions of HLA Twin. In comparison, **Fast mode** achieves better analysis times with the help of optimization, and also the result files will be significantly smaller. Thus the long-term storage requirements became lower.

Gene targeting

You can select the targeted gene family and loci set. The predefined sets are based on the different Holotype HLA Kit configurations (5, 7, 11-loci) and there is an option to set custom loci for the genotyping, which can be useful for reanalysis purposes. When the 'Auto' option is selected, gene family and targeted loci selection is disabled since these parameters will be determined from the sample during the analysis.

Sample data type

You can select the used sequencing technology on the Sample data type tab. There is an option to analyse data from the PacBio technology, but this module is in Alpha stage, so it may not contain all of the features planned for the final version. It is likely to contain a number of known or unknown bugs. The QC metrics for PacBio are not yet optimized and they should not be fully trusted.

Analysis options

- *Big data analysis*: When using this option, QC metrics will be adjusted to deal with big data.
- *Maximum pairs processed per locus*: Explore uses locus based subsampling to optimize results. This value determines how many reads will be used for each locus. The datatype is assessed at the beginning of each typing and the amount of reads processed is adjusted accordingly. Note that only reads longer than 75 bases are selected during subsampling (for paired data both reads have to be over this minimum length threshold). You can experiment with this value to find the best match for your data. For targeted data sets the default 4000 read pairs per locus works well. If you run into problematic samples, you can increase the read count to resolve these issues.
- *Disable quality based subsampling*: If this checkbox is checked, subsampling will be based on the per locus read count and the number and type of the targeted locus. This option is used for generating mappability statistics.
- *SG analysis mode*: when using the Statistical genotyping method, it's possible to determine which parts of genetic information will be used for the analysis. The **Exons only** option utilizes exonic information only while the **Whole gene** option allows the algorithm to use information from the intronic and UTR parts of the locus as well. The regions will be displayed in the Gene Browser according to this selection.
- *Ignore rare alleles*: Using this option, very rare alleles will be ignored in the analysis and will not be included or reported. If this checkbox is unchecked, rare alleles will be marked in the results. If this option is greyed out, you need to install a new database in order to import the rare alleles.
- *Treat regions separately*: reads are aligned to each region separately. This is reflected in the alignment in the Gene Browser.

Note that for analysing filtered whole exome or whole genome data, the following advanced parameters are suggested: all reads should be processed, quality based subsampling should be turned off and all loci should be set as targeted.

Gene targeting

You can select the targeted gene family and loci set. The predefined sets are based on the different Holotype HLA Kit configurations (5, 7, 11-loci) and there is an option to set custom loci for the genotyping, which can be useful for reanalysis purposes. When the 'Auto' option is selected, gene family and targeted loci selection is disabled since these parameters will be determined from the sample during the analysis.

2.9 Genotyping Reanalysis

Any analysis can be rerun for the same set of input files - samples - which it had been setup with. Apart from the input files, any parameters and settings can be changed. This allows you to correct any mistakes - like choosing an inappropriate parameter - and also enables examination of results with different settings or IMGT database versions. The number of free re-analyses is unlimited.

2.10 Export results

2.10.1 General information

Genotyping results can be exported in the standard HML or in PDF format using the Export Results wizard available from the Genotyping dashboard.

On the first screen of the wizard, export format and the preferred location for the exported files can be selected.

If you check the *Use LD* checkbox, results will be exported as shown on the *Genotyping analysis result* screen when LD functionality is switched on. If the option is unchecked LD information will not be included in the exports.

2.10.2 HML export

General information

The currently supported HML Report format is version 1.0.1. For reference and variant reporting the first best match allele pair of the result is used and the reference is extracted from the IMGT database. If no allele pair is available then we fall back to the GRCh38 reference: in this case the coordinates are not as accurate therefore those results/exports shouldn't be used for further processing except some special research cases. A HML report can be generated from imported results as well.

HML export wizard

Export options

All alleles or only assigned alleles can be exported by choosing the All or Assigned options, respectively.

The *Export precision* options lets you refine the precision level of the reported genotypes/alleles:

- P group: export P groups instead of alleles
- 2 fields: export 2 field allele names
- G group: export G groups instead of alleles
- 3+2 fields: export 3 field allele names for class I and 2 field allele names for class II
- 3 fields: export 3 field allele names
- 4 fields: export 4 field allele names

The *Precision reduction for novelties* option allows further automatic refinement of the precision: in case of novelties/mismatches it reduces the precision automatically to 3 field. Test Id and Test Id Source can also be specified on this screen.

Advanced export options

On this screen, values for several HML fields can be specified.

By default, one HML file is created per sample and file names are generated using analyses names. Single file export can be selected to export all selected results with the specified file name.

NMDP export options

This optional step of the wizard can be used for specifying packing list files. A default center code can also be specified in this step of the wizard.

2.10.3 PDF export

A PDF report can be generated based on the results of an analysis. The contents are highly customizable, you can select which metrics, statistics and other details should appear in the report. The alleles can also be selected by restricting the reported alleles to the Assigned alleles only. If you check the *Use LD* checkbox, results will be exported as shown on the *Genotyping analysis result* screen when LD functionality is switched on. If the option is unchecked LD information will not be included in the exports.

The organization of the report is the following:

- Header: mandatory section on top of each page - contents cannot be customized.
- Sample name: as displayed in the software.
- Analysis date: date and time when the analysis has run.
- Analyzed by: user name used to produce the analysis.
- Approval date: if the analysis results have been approved the date of the approval is displayed. In case the analysis have been approved more than once the last date is displayed.
- Approved by: if the analysis results have been approved the user name of the approved is displayed.
- Sent for Approval by: if the analysis results have been sent for approval the user's name who sent the results is displayed. If the results have not been sent for approval, the field contains the 'Not Sent' message.
- Genotyping result: contains allele candidates for each locus and the related P&G groups or the phenotype where applicable. When a result is too ambiguous to report - more than 10 allele pairs would be included in the report - the report will contain a warning about this as it is difficult to display so many alleles. You can reduce the amount of displayed alleles for example by assigning only the relevant ones and re-generating the report.
- Analysis configuration: contains the details of the protocol and a warning in case the result had not been produced with the Holotype protocol.
- Quality control summary: concordance and overall QC result for each locus
 - Mismatches
 - Novelities
- Alignment statistics: coverage and average coverage depth for each allele candidate on each locus
- Data Statistics: ratios of processed and skipped reads, noise and other details also displayed on the Data Statistics tab
- Comments: user comments related to samples or loci
- Genotyping result: contains allele candidates for each locus and the related P&G groups or the phenotype where applicable. When a result is too ambiguous to report - more than 10 allele pairs would be included in the report - the report will contain a warning about this as it is difficult to display so many alleles. You can reduce the amount of displayed alleles for example by assigning only the relevant ones and re-generating the report. The user added alleles are marked with a "(+)" sign in the report.
- Analysis configuration: contains the details of the protocol used to produce the result.
- Quality control summary: concordance and overall QC result for each locus
 - Mismatches
 - Novelities
- Alignment statistics: coverage and average coverage depth for each allele candidate on each locus
- Data Statistics: ratios of processed and skipped reads, noise and other details also displayed on the Data Statistics tab
- Comments: user comments related to samples or loci
- Genotyping result: contains allele candidates for each locus and the related P&G groups or the phenotype where applicable. When a result is too ambiguous to report - more than 10 allele pairs would be included in the report - the report will contain a warning about this as it is difficult to display so many alleles. You can reduce the amount of displayed alleles for example by assigning only the relevant ones and re-generating the report. The user added alleles are marked with a "(+)" sign in the report.
- Analysis configuration: contains the details of the protocol used to produce the result.
- Quality control summary: overall QC result for each locus
- Alignment statistics: coverage and average coverage depth for each allele candidate on each locus
- Data Statistics: ratios of processed and skipped reads and other details also displayed on the Data Statistics tab
- Comments: user comments related to samples or loci
- Locus specific content: details relevant for each locus - contents are customizable.
- Quality control details: detailed quality control metric values for each allele candidate in each locus

Selected report customization options are pertained for each user separately so you won't have to select the required contents again if you are satisfied with the setup as it was the last time you used it.

2.10.4 JSON export

Export options

You can choose to export all alleles from a typing result or assigned alleles only.



Export format and content

Results will be exported in standard JSON format.

i Sample JSON file

```
{
  "sampleName": "Sample",
  "analysisRunName": "Sample_result",
  "analysisDate": "Sep 8, 2017 12:55:14 PM",
  "executedBy": "Unknown",
  "approvedBy": "Not approved",
  "sampleComment": "Sample level comment",
  "genotypes": [
    {
      "gene": "HLA-A",
      "genotype": "01:03#1+02:40:01",
      "genotypeGGroup": "No G group+No G group",
      "genotypePGroup": "02:40P+No P group"
    },
    {
      "gene": "HLA-B",
      "genotype": "08:74#1+53:21#1",
      "genotypeGGroup": "No G group+No G group",
      "genotypePGroup": "No P group+No P group"
    }
  ],
  "geneResults": [
    {
      "gene": "HLA-A",
      "frequentQcMetrics": [
        {
          "name": "Read count",
          "value": "3464",
          "code": "READ_COUNT"
        }
      ],
      "rareQcMetrics": [
        {
          "name": "Read quality",
          "value": "36.22",
          "code": "READ_QUALITY"
        }
      ]
    }
  ]
}
```

```
],
"alleles": [
  {
    "allele": "01:03",
    "averageCoverageDepth": 20
  },
  {
    "allele": "02:40:01",
    "averageCoverageDepth": 26
  }
],
"alleleFrequencies": [

],
"comment": "Locus level comment"
},
{
  "gene": "HLA-B",
  "frequentQcMetrics": [
    {
      "name": "Read count",
      "value": "3498",
      "code": "READ_COUNT"
    }
  ],
  "rareQcMetrics": [
    {
      "name": "Read quality",
      "value": "35.97",
      "code": "READ_QUALITY"
    }
  ],
  "alleles": [
    {
      "allele": "08:74",
      "averageCoverageDepth": 6
    },
    {
      "allele": "53:21",
```

```
"averageCoverageDepth": 8
}
],
"alleleFrequencies": [

],
"comment": "No comment available"
}
],
"configuration": {
  "sequencer": "ILLUMINA",
  "dataType": "PAIRED",
  "typingMethod": "TWIN",
  "maxReadsPerLocus": 4000,
  "disableQualityBasedSubsampling": false,
  "ignoreRareAlleles": false,
  "novelAlleleDetectionMode": "EXON_NOVELTIES_ONLY",
  "applicationVersion": "2.2.0",
  "applicationBuildId": "0ab106d522677cac281333ab6aa7e45981e0b7e9",
  "databaseVersion": "3.25.0_3(+E)",
  "defaultConfig": true
}
}
```

2.10.5 HPRIM export

If Omixon HLA is configured to have a LIMS connection you can use this wizard to launch the process of exporting analysis results into the configured LIMS system.

Filter analysis results

By default all analysis results in the result table are selected for export. You have the option to narrow down the list based on following predefined criteria:

- *Approved only*: if this is turned on then only the approved results are shown (default is *on*)
- *Exported only*: if this is turned on then only the previously exported results are shown - in other words this filter can be used for re-exporting results after modification (default is *off*)
- *Show approved results since*: if the *Approved only* checkbox is on then you can set an initial date as an additional filter based on the analysis execution time (default is one week before the actual date)
- *Show exported results since*: if the *Exported only* checkbox is on then you can set an initial date as an additional filter based on the analysis result export time (default is one week before the actual date).

Select analysis results

On the next tab only those analysis results are listed which are matching all the filter criteria set in the previous wizard step. This screen allows you to select one or multiple results to be exported.

For each analysis result the following attributes are shown to help identifying them:

- *Analysis name*: name of the result as shown in other screens in the application
- *Execution date*: date and time of the sample analysis execution
- *Approved by*: login name of the user who has approved the result
- *Approval date*: date and time of the approval

Export configuration

In the last step of the export wizard the following export options can be set:

- *Export method*: *Export by sample* or *Export by batch* - the former one defines that each typing result should be placed into a separate file (this is the default setup) and in the latter case all typing results will be exported into one file.
- *Export location*: the folder where export files should be placed.

Once you have finished the wizard the export generation is executed automatically in the background. The process can be monitored just like any other background task in the *Event log*

2.11 Filtering

Filtering is a built in wizard that helps with increasing the precision and efficiency of big data analysis.

2.11.1 Access

When adding a folder on the Genotyping Dashboard, Explore monitors all input files: whenever the reading process finds a file that surpasses a size threshold and certain amount of (short) reads (approx. 5 million), it automatically falls into big data category, which grants the option of filtering.

Whenever there's a file that fulfills the conditions, the Filter button appears on the right side of the sample, next to the Analyse and Custom Analysis buttons.

2.11.2 Usage

The wizard can be initiated by clicking the Filter button. This tool has two tabs:

- Targeted genes
- Filtering options

Targeted genes

This tab offers the option to narrow the analysis to selected genes. The following options are available:

- **Holotype 5-Locus Kit**: alleles amplified by Holotype 5-Locus kit (HLA-A, HLA-B, HLA-C, HLA-DQB1, HLA-DRB1)
- **Holotype 7-Locus Kit**: alleles amplified by Holotype 7-Locus kit (HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1)
- **Holotype 11-Locus Kit**: alleles amplified by Holotype 11-Locus kit (HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5)
- **MIC**: MICA alleles
- **ABO**: ABO alleles
- **Custom loci**: this option enables processing whole genome sequences, and any combination of supported alleles, which includes most Class I and Class II HLA genes and pseudogenes.




Filtering options

Two options are available on this screen:

- **Process all reads checkbox:** When this option is selected, the software will attempt to process every read that is contained in the file. It is the recommended setting if analyzing whole genome sequences.
- **Maximum collected per locus:** This option is available if the one above is unchecked. You can determine the maximum reads that will be processed for each locus, which will help if the amplification is imbalanced (ie. some alleles are represented by several million reads, while others by only a couple thousand), can also help with speeding up the analysis process.

When the configuration is done, the wizard is closed by either cancelling the process by clicking on Close, or running the configured filtering by clicking Finish. By clicking Finish, the task will be registered in the process manager, and after the filtering is done, two files will be created from the filtered reads: a single and a paired read one, which are now available for further analysis.

 The initial file will not be manipulated, all changes are saved in new files.

3 Genotyping analysis result

3.1 Introduction

High-level overview of the Genotyping results.

A high-level overview table of the genotyping results. This can include the summary of results for many samples. The result table can be collapsed using the *Collapse/Expand* button to make it even more concise. In the collapsed view only the first best matching allele is shown. If there are further best matching alleles, the level of ambiguity is denoted by coloring the affected fields red in the allele name. Each Sample can be selected and unselected (Ctrl+click). The following options are available:

- *Sample Details* - leads to the *Genotyping sample result* dashboard which contains detailed data and statistics.
- *Browse Alignment* - opens the first pair of consensus sequences with the best matching allele candidates in the Gene browser.

The results can be exported in TXT (tab delimited text), CSV (comma separated text), XLS (Excel) or XLSX (Excel XML) format.

You can export the results with or without the flags that are displayed on the screen by selecting between the Overview and the Overview with allele flags options.

Tip - If you would like to export only the assigned allele candidates first click the **Assigned Only** button, then export the results.

3.1.1 Commenting

Comments can be added to a selected sample using the Sample Comment button. Commented samples are marked with a small red triangle in the top right corner of the sample name field on this screen. These comments are by default included in the PDF report but they can be excluded if necessary.

Comments can be removed by clicking the Sample Comment button again and removing the text.

3.2 Top panel

On the top of the screen you can see three rows of functions.

3.2.1 Top row

On the left, you can find the *Toggle fullscreen* button which hides the less important parts of the current screen, while on the right you can see the following things:

- the ID of the current user,
- the memory usage widget panel,
- the status panel of the Typer Scheduler,
- the status panel of the Event Log,
- the welcome tutorial button,
- the logout button,
- and the exit button.

In case of the desktop version, the actual memory usage can be observed at the top right corner of the screen. This displays the amount of currently used memory compared to the allowed maximum. Note, that the software might be allocating more memory from the operating system compared to the reported actual memory usage to make further computations faster by keeping some previously allocated memory pieces for later use, thus making the allocation of new memory unnecessary during further computations. By clicking on the memory report widget the software tries to free up as much memory as possible on a best effort basis: this is only an attempt, there are many technical factors which might block returning all memory to the operating system. It is also important to mention that by explicitly asking to free up memory the subsequent computations might become slightly slower, since the option of reusing certain structures allocated before becomes impossible. Only use this option when all planned analyses have been finished and you would like to use other programs more intensively in parallel while investigating the analysis results.

3.2.2 Middle row

There are four navigation buttons that are available on every screen of the tool. These general navigation buttons work very similar to navigation functions available in most widely used internet browsers:

- The *Back* button will take you to the previous screen.

- The *Forward* button will take you to the next screen.
- The *Up* button will move you one level up in the application.
- The *Home* button will take you to the starting screen (i.e. the Genotyping dashboard).

Right from the navigation button, in the information panel, you can find relevant high-level information about the current screen. On the *Genotyping dashboard*, you can see the release date and version of the active IMGT database.

On the far right hand side of the middle row, you can find the bookmark and context specific help buttons. The help tool shows relevant usage information and tips about functions and data available on the current screen of the application and can also be opened by pressing F1.

You can search the Help for any expression by using the Magnifier glass icon displayed on the popup window. Relevant hits are displayed in the left panel where you can select them and open them in the right hand reading pane. You can close the search by clicking the X displayed beside the search field.

3.2.3 Bottom row

The bottom row of the top panel contains a series of buttons which contain the main functions available on the screen.

3.3 Interpreting Results

By default, only the *best matching* candidates are displayed in the results table. These are the pairs of candidates with *the lowest exon and non-exon mismatch counts* compared to the pair of consensus sequences generated from the short reads. If an ambiguous best match is returned, then there will be more than two best matching candidates for the locus. Hover your mouse over any result to display a tooltip that contains further details about the results.

Novel alleles are named relative to an existing allele in the database. The base allele is selected in a way to take into account the biochemical similarity of the novel allele and the base allele as much as possible. This means that sometimes alleles with more mismatches/novelties on the non key exons are used as a base allele if this leads to a higher similarity/lower mismatch count - ideally even equivalence - on the key exons (exon 2 and exon 3 for class I genes and exon 2 for class II genes).

In case only the Statistical genotyping method was used Best matching alleles are allele pairs whose alignments contain reads specific to them among other reads which are present in the alignment of other allele pairs as well. Based on this unique read count based ordering method the software selects those allele pairs as the best result which has the highest statistical rank based on the alignments. If an ambiguous best match is returned, then there will be more than two best matching candidates for the locus.

The result will not contain any novel allele candidates since there is no novel allele detection in the Statistical genotyping method.

Best matching alleles are allele pairs whose alignments contain reads specific to them among other reads which are present in the alignment of other allele pairs as well. Based on this unique read count based ordering method the software selects those allele pairs as the best result which has the highest statistical rank based on the alignments. If an ambiguous best match is returned, then there will be more than two best matching candidates for the locus. Hover your mouse over any result to display a tooltip that contains further details about the results.

By pushing the *Best Matches Only* button, it is possible to see other close allele candidates.

By selecting a sample (clicking on it in the table) you will be presented with more options for drilling down into the details of the results. You can see details for a sample, you can browse allele candidates and consensus alignments in the browser, and you can see the pair(s) of candidates for each locus. More information is available within the help page of each of these functions.

A gene is homozygous if the same allele is present on both of the chromosomes and heterozygous if two different alleles of the same gene are present. In case only one copy of a gene is present, the locus is hemizygous. Details about the allele display can be found under the Traffic Lights, Quality control light section. Alleles from different genes are not in random recombination if the genes are located close to each other in the chromosome. This phenomenon is called 'linkage disequilibrium and this information is used for determining if a particular locus is hemizygous.

3.3.1 Troubleshooting missing results

When no alleles could be reported for a targeted gene, a markup describing the possible reason for the missing allele call is shown. For additional information, hover over the info icon next to the markup and read the tooltip. The following cases can be reported:

For non DRB3/4/5 loci:

- *Not targeted* - Meaning that the locus was not targeted in the specific analysis. This markup is a placeholder for loci which were targeted in other analyses displayed in the table.
- *No data present* - No data present means that the locus has dropped out during sequencing and should be re-sequenced.
- *Insufficient or low quality data* - There is insufficient data or the data is of low quality in the sample. Quality control results should be checked for more detail.

For DRB3/4/5:

- *Not targeted* - Meaning that the locus was not targeted in the specific analysis. This markup is a placeholder for loci which were targeted in other analyses displayed in the table.
- *Allele not expected* - There is no allele expected at this locus based on known linkage disequilibrium with HLA-DRB1 and HLA-DQB1.
- *Expected allele not found* - This markup means that based on known linkage disequilibrium information, data was expected for the locus/allele but was not found.
- *Unexpected allele found* - Data was found for a locus/allele, which was not expected based on known linkage disequilibrium information.
- *Insufficient or low quality data* - There is insufficient data or the data is of low quality in the sample. Quality control results should be checked for more detail.

When no alleles are reported for a targeted gene it is suggested to rerun the sample in question using a higher number of reads. (The number of processed reads can be set in the *Advanced Genotyping wizard*.) The reasons behind the missing allele level results can be that the coverage does not reach the minimum threshold on the allele or on the exons, or the coverage depth is too small. Processing more reads can help making the signals that support the correct alleles stronger.

3.4 Filtering Alleles






A summary of the currently active filters is displayed to the top right of the Genotyping analysis results screen. The following filter options are available:

- **Setup Loci** - A set of visible loci can be selected.
- **Best Matches Only** - Toggles remainder alleles.
- **Assignment State** - In combination with the Best Matches Only filter, the following combinations can be shown:
 - Assigned alleles, best matches and remainders
 - Only assigned alleles
 - Best matches and remainders
- **Workflow State** - This is a sample level filtering option that toggles between showing samples in progress, waiting for approval and only approved samples.





3.5 Traffic Lights



Results are annotated using a "Traffic light" system. Similarly to an actual traffic light, three different colours are used, all with different meanings. Unlike in a real traffic light, "mixed colours" are also available.

In the topmost row for each locus, the locus specific markups are displayed. These include:

- Quality control traffic lights:
 - These lights are based on the locus level quality control measures and can be one of the following:
 -  (green) - PASSED: the locus passed all QC tests,
 -  (yellow/green) - INFO: one or more QC tests produced lower than average results,
 -  (yellow) - INSPECT: one or more QC tests produced concerning results, manual inspection of the results needed,
 -  (red/yellow) - INVESTIGATE: one or more QC tests showed low result quality, manual inspection and possibly reanalysis is needed,
 -  (red) - FAILED: one or more QC tests showed very low result quality, manual inspection is needed to determine the cause and the locus or sample likely needs re-sequencing or re-typed by alternative methods.






Other locus level markups are also presented beside the Quality lights:

Zigosity markups: Heterozygous loci have the  markup while homozygous loci are marked by . Hemizygous loci are marked with . In case a locus is hemizygous only one allele is displayed and the other cell is left empty. In case the zygosity of a locus can't be determined based on the data available, it is marked with .










- Novelty markups: loci with alleles containing exonic (or exonic and intronic) novelties are marked with , while loci with novel alleles containing only intronic novelties are marked with .

3.5.1 Quality control light

The quality control traffic light is based on locus level quality control (QC) measures. It is displayed in the topmost row for each targeted locus in an analysis result. These lights are based on the locus level quality control measures. They are displayed in the topmost row for each targeted locus in an analysis result.

-  (green) - PASSED: the locus passed all QC tests,
-  (yellow/green) - INFO: one or more QC tests produced lower than average results,
-  (yellow) - INSPECT: one or more QC tests produced concerning results, manual inspection of the results needed,
-  (red/yellow) - INVESTIGATE: one or more QC tests showed low result quality, manual inspection and possibly reanalysis is needed,
-  (red) - FAILED: one or more QC tests showed very low result quality, manual inspection is needed to determine the cause and the locus or sample likely needs re-sequencing or re-typed by alternative methods.

Some other markings can also be presented for alleles:

- Alleles displayed with the blue font are homozygous.
- Serological equivalent antigens: If information regarding the associated serological equivalent antigens is available for the locus, the tooltip of the *antigen icon*  will contain that information.
- Rare alleles are marked with an exclamation point icon .
- Novel alleles containing exonic (or exonic and intronic) novelties are marked with , while novel alleles containing only intronic novelties are marked with .
- The presence of additional imbalanced remainder alleles that are close to the best match in coverage, but have significantly lower coverage depth is marked with a scale icon . These imbalanced alleles are displayed with italic font.
- Alleles with extended allele sequence are marked with a *plus sign* .
- If a minor allele with well-known low amplification is present in the imbalanced minor allele list, the allele is marked with . In this case validation of the homozygous result using an alternate genotyping method (e.g. SSO) is strongly suggested.
- Hemizygous loci are marked with . In case a locus is hemizygous only one allele is displayed and the other cell is left empty. In case the zygosity of a locus can't be determined based on the data available, it is marked with .

3.6 Assigning Alleles

For each locus, one or more allele pair candidates can be assigned. Candidates can be assigned manually by clicking on the check mark before the allele's name. The following assignment options are available via the Assignment button:

- All alleles with a certain level of overall QC result or above can be assigned.
- Assign best matches
- Assign only heterozygous results
- Assign only unambiguous results
- Assign novel alleles

All assignments can be deleted by the *Clear assignments* button in the wizard. Note, that the 'Assign ...' options affect only displayed samples and alleles - they don't affect alleles that are not being displayed due to filtering criteria. If you prefer to do the assignments separately for each sample/allele, you can find similar functions on the *Genotyping sample result* dashboard.

Allele assignment only works for samples in "In progress" state. The function is blocked in "Ready" and "Approved" states. Similarly assignment is blocked when genotype precision is set to lower than the maximum precision in the analysis result.

To reach the assignment history for an analysis result, press the clock icon in the top-right corner of the analysis result name column in the result table.

3.7 Export table

Every Genotyping result table can be exported in either CSV (comma separated text), TXT (tab delimited text), XLS or XLSX format. Note, that only visible results are exported (i.e. if an allele is filtered out based on low coverage, it won't be shown in the exported file either).

There are several similar export table functions:

On the *Genotyping analysis results* dashboard, you can export all the results from an analysis. If you select the *Overview* export, only the allele names and quality control measures are exported. The *Detailed* table also contains flags and statistics for each allele candidate.

Flag abbreviations:

- R - rare,
- I - imbalanced,
- E - extended allele.

Quality control test result abbreviations:

- P - passed,
- I1 - info,
- I2 - inspect,
- I3 - investigate,
- F - failed.

Concordance test result abbreviations:

- C - fully concordant,
- C2 - concordant on a two digit level (1st field),
- C4 - concordant on a four digit level (2nd field),
- C6 - concordant on a six digit level (3rd field),
- D - discordant.

3.8 Linkage disequilibrium details

3.8.1 Overview

Based on the known genetic rules related to linkage disequilibrium (LD) and the database that is integrated into the software in case of DRB3, DRB4 and DRB5 loci additional information can be reported next to the actual genotyping results. These can be either markups describing the relation between the alleles in the genotyping result and the alleles expected based on LD or can be actual allele display changes in the result - e.g. showing up an allele only once for hemizygous results.

Note

If more than 50 best matching results can be found on a locus, LD will not be calculated. If you press the Show LD details button, no information will be displayed.

The output of matching the genotyping result with the LD based expected result is reported in the software as follows:

Matching genotyping result with linkage disequilibrium	Allele present in genotyping result	Allele <i>not</i> present in genotyping result
Allele expected based on LD	Allele displayed, no markup	Allele not displayed, bold markup says: 'Expected allele not found'
Allele <i>not</i> expected based on LD	Allele not displayed, bold markup says: 'Unexpected allele found'	Allele not displayed, markup says: 'Allele not expected'
There is no information about the allele in the LD database	Allele displayed, bold markup says: 'No linkage information available'	Allele not displayed, markup says: 'Allele not expected'

3.8.2 Case by case guide

Specific cases can be detailed as follows.

Expected allele not found

- Allele present in the LD based expected result but not present in the genotype

- DRB4*01:01 has an extra annotation: “This allele is known to drop out, please consider checking this sample with an alternative typing technology (e.g. SSO/SSP).”
- The only allele is null in the genotype of DRB4 meanwhile the LD based expected result contains DRB4*01:01 too

Unexpected allele found

- Genotyping result present for all of DRB3/4/5 but the LD based expected result contains the alleles of the other two loci only
- Genotyping result present for two of DRB3/4/5 but the LD based expected result contains the alleles of one other locus only
- Heterozygous genotype but the LD based expected result contains only one of the alleles and another locus
- Heterozygous genotype but the LD based expected result contains only one of the alleles and denotes the other allele to be not present
- Second allele of a heterozygous genotype where both alleles match the LD based expected result and the other pair matches another locus
- Second allele of a heterozygous genotype where both alleles match the LD based expected result and the other pair refers to an allele known to be not present

Allele not expected

- Allele is the only result for a locus and it matches one of the LD based expected result pairs meanwhile the other pair denotes the other allele to be not present
- Genotype not present and no LD based expected result is available

No linkage info available

- Allele not present in the LD based expected result
- Ambiguous allele meanwhile only other allele(s) match the LD based expected result

Unknown

- Multiple contradicting rules among the above ones could be applied to the result

3.8.3 Notes

- Results are assumed to be hemizygous by default
- Alleles are considered up to 2 fields precision
- If result is not present on DRB1 then we abandon the LD based result evaluation
- Novel alleles are interpreted by taking their base allele, the novelty part is not considered

3.8.4 DQA1, DQB1 allele dropout warning

As allele dropouts are very hard to distinguish from naturally homozygous results, LD information is used to detect possible DQA1 or DQB1 allele dropouts.

In case DQA1 or DQB1 is found homozygous by the algorithm, then it is checked if one of the alleles of the other locus is linked with an allele which is not the homozygous one. In case there is such allele(s) then a warning is displayed next to the homozygous result. (E.g. if DQB1 is homozygous for 06:03 and DQA1 is heterozygous (01:03+05:05): DQA1*01:03 is linked with DQB1*06:03 and DQA1*05:05 is linked with DQB1*03:01, so we display a warning about the possible DQB1*03:01 dropout.)

If both DQA1 and DQB1 are homozygous then they are checked against DRB1 the same way as described above.

4 Genotyping sample result

4.1 Top panel

On the top of the screen you can see three rows of functions.

4.1.1 Top row

On the left, you can find the *Toggle fullscreen* button which hides the less important parts of the current screen, while on the right you can see the following things:

- the ID of the current user,
- the memory usage widget panel,
- the status panel of the Typer Scheduler,
- the status panel of the Event Log,
- the welcome tutorial button,
- the logout button,
- and the exit button.

In case of the desktop version, the actual memory usage can be observed at the top right corner of the screen. This displays the amount of currently used memory compared to the allowed maximum. Note, that the software might be allocating more memory from the operating system compared to the reported actual memory usage to make further computations faster by keeping some previously allocated memory pieces for later use, thus making the allocation of new memory unnecessary during further computations. By clicking on the memory report widget the software tries to free up as much memory as possible on a best effort basis: this is only an attempt, there are many technical factors which might block returning all memory to the operating system. It is also important to mention that by explicitly asking to free up memory the subsequent computations might become slightly slower, since the option of reusing certain structures allocated before becomes impossible. Only use this option when all planned analyses have been finished and you would like to use other programs more intensively in parallel while investigating the analysis results.

4.1.2 Middle row

There are four navigation buttons that are available on every screen of the tool. These general navigation buttons work very similar to navigation functions available in most widely used internet browsers:

- The *Back* button will take you to the previous screen.
- The *Forward* button will take you to the next screen.
- The *Up* button will move you one level up in the application.
- The *Home* button will take you to the starting screen (i.e. the HLA Typing dashboard).

Right from the navigation button, in the information panel, you can find relevant high-level information about the current screen. On the *Genotyping dashboard*, you can see the release date and version of the active IMGT database.

On the far right hand side of the middle row, you can find the bookmark and context specific help buttons. The help tool shows relevant usage information and tips about functions and data available on the current screen of the application and can also be opened by pressing F1.

You can search the Help for any expression by using the Magnifier glass icon displayed on the popup window. Relevant hits are displayed in the left panel where you can select them and open them in the right hand reading pane. You can close the search by clicking the X displayed beside the search field.

4.1.3 Bottom row

The bottom row of the top panel contains a series of buttons which contain the main functions available on the screen.

4.2 Available functions

4.2.1 Opening the browser

Both chromosomes or a single consensus sequence and the corresponding alleles can be viewed in the Gene browser, using the *Browse Alignment*, *Browse Allele 1* or *Browse Allele 2* functions. Note that the same settings can be chosen from within the Gene browser.

4.2.2 Detailed genotyping information

Mismatch and coverage statistics and detailed mismatch and novelty lists for a selected allele pair can be viewed using the *Genotype Details*, *Show Mismatches* and *Show Novelty* functions.

4.2.3 Opening the browser

Candidates for both chromosomes can be viewed in the Gene browser, using the *Browse Alignment*, *Browse Allele 1* or *Browse Allele 2* functions. Note that the same settings can be chosen from within the Gene browser.

4.2.4 Detailed genotyping information

Detailed coverage statistics can be viewed using the *Genotype Details* functions.

4.2.5 Customizing displayed results

Loci shown in this screen can be customized using the *Setup Loci* function. Alleles can be filtered by status, using the *Best Matches Only* and *Assigned Only* functions. Result resolution can be changed using the *Genotype Precision* button.

4.2.6 Assigning alleles

For each locus, one or more allele pair candidates can be assigned. Candidates can be assigned manually by clicking on the check mark before the allele's name. The following assignment options and assignment history are available via the Assignment option:

- All alleles with a certain level of overall QC result or above can be assigned.
- Assign best matches
- Assign only heterozygous results
- Assign only unambiguous results
- Assign novel alleles

All assignments can be deleted by the *Clear assignments* button in the wizard.

Note, that the 'Assign ...' options affect only displayed samples and alleles - they don't affect alleles that are not being displayed due to filtering criteria. You also have the possibility to assign allele pairs one by one by clicking the assignment status icon before the allele pairs in the *Genotype Tree*.

Allele assignment only works for samples in "In progress" state. The function is blocked in "Ready" and "Approved" states. Similarly assignment is blocked when genotype precision is set to lower than the maximum precision in the analysis result.

To reach the assignment history for an analysis result, press the clock icon in the top-right corner of the analysis result name column in the result table.

4.2.7 Commenting

Comments can be added to a selected sample or each loci within a selected sample using the Sample Comment or the Locus comment buttons. Commented samples and loci are marked with a small red triangle in the top right corner of the sample name field or the allele field on the Genotyping Analysis result screen. These comments are by default included in the PDF report but they can be excluded if necessary. Comments can be removed by clicking the Sample Comment or Locus comment button again and removing the text.

4.2.8 PIRCHE® epitope matching

With this function you can send selected allele results to the PIRCHE website for epitope matching. Once the website was displayed, use your PIRCHE account to sign in and choose the option that suits your needs best. The allele names for the loci you have selected in the *Send to PIRCHE* dialog will be filled in automatically. On the website you can manually edit the allele names before matching.

For more information on the PIRCHE® technology and access request, please visit the *Help* section of the PIRCHE website or contact the PIRCHE team at info@pirche.com³.

4.3 Genotype

4.3.1 Tree view

The Genotype tab displays the results in a tree format. The different levels of the tree refer to different properties in different gene families. In case of HLA the first levels of the tree are the P group and the G group - where available. Using the *Genotype Precision* button reduces the result display level to these groups.

In case of ABO results the first level of the tree is the Phenotype information derived from the results.

4.3.2 Troubleshooting missing results

When no alleles could be reported for a targeted gene, a markup describing the possible reason for the missing allele call is shown. For additional information, hover over the info icon next to the markup and read the tooltip. The following cases can be reported:

For non DRB3/4/5 loci:

- *Not targeted* - Meaning that the locus was not targeted in the specific analysis. This markup is a placeholder for loci which were targeted in other analyses displayed in the table.
- *No data present* - No data present means that the locus has dropped out during sequencing and should be re-sequenced.
- *Insufficient or low quality data* - There is insufficient data or the data is of low quality in the sample. Quality control results should be checked for more detail.

For DRB3/4/5:

- *Not targeted* - Meaning that the locus was not targeted in the specific analysis. This markup is a placeholder for loci which were targeted in other analyses displayed in the table.
- *Allele not expected* - There is no allele expected at this locus based on known linkage disequilibrium with HLA-DRB1 and HLA-DQB1.
- *Expected allele not found* - This markup means that based on known linkage disequilibrium information, data was expected for the locus/allele but was not found.
- *Unexpected allele found* - Data was found for a locus/allele, which was not expected based on known linkage disequilibrium information.
- *Insufficient or low quality data* - There is insufficient data or the data is of low quality in the sample. Quality control results should be checked for more detail.

When no alleles are reported for a targeted gene it is suggested to rerun the sample in question using a higher number of reads. (The number of processed reads can be set in the *Advanced Genotyping wizard*.) The reasons behind the missing allele level results can be that the coverage does not reach the minimum threshold on the allele or on the exons, or the coverage depth is too small. Processing more reads can help making the signals that support the correct alleles stronger.

4.3.3 Adding alleles manually

Right click on the Genotype panel to find the context menu which provides the "Add genotype" option to add alleles to the displayed result. A popup window will be displayed where the original result is highlighted. Browse the displayed P and G groups to locate the allele pair candidate that you wish to add. Please note that an already present result cannot be added again. In case of heterozygous loci it is only possible to add allele pairs but not single alleles. If the locus is marked with a hemizygous or a hemi- or homozygous markup, you can only

³ <mailto:info@pirche.com>

choose a single allele to add. You can add allele pair candidates from any P and G group and the allele pair will be displayed on the Genotype panel accordingly with a different blue markup in front of the pair.

Adding alleles manually is only possible for samples in "In progress" state. The function is not available in "Ready" and "Approved" states.

4.3.4 Removing additional alleles

The manually added allele pair candidates can be removed by selecting them and using the right mouse button to summon the context menu. Select the "Delete user added genotype" option to remove the allele pair candidate.

Removing additional alleles is only possible for samples in "In progress" state. The function is not available in "Ready" and "Approved" states.

4.4 Quality Control

Several quality control measures are calculated for every loci. Each measure for each loci is marked with a traffic light system. Thresholds for light colour assignment are presented as a tooltip in the right hand side of every cell. The QC measures are the following:

Overall

Overall QC status of the loci, calculated as the value of the worst QC measure or measures.

Primary QCs for Interpretation

- *Read count* - total number of reads mapped to a locus.
- *Noise ratio* - number of reads considered as noise relative to the total number of reads mapped to a locus.
- *Key-exon spot noise ratio* - maximum number of reads assumed to be systematic noise relative to the total number of reads at a certain position in the key exon regions (i.e. exons 2 and 3 for Class I loci and exon 2 for Class II genes in the HLA family).
- *Consensus coverage key exon minimum depth* - lowest number of reads supporting the coverage across all key exon positions.
- *Key exon allele imbalance* - average ratio of the alleles for all key exon regions. Values around 50%-50% indicate a balanced heterozygous result, while values close to 100%-0% indicate imbalance in the proportion of reads derived from the two chromosomes. Homozygous results get 50%-50% assigned. Note, that it is not always possible to distinguish between homozygous regions with a small amount of contamination or artifact reads and highly imbalanced heterozygous regions.
- *Genotype available* - a Yes value indicates that a genotype could be identified for a locus while No indicates that no best matching alleles were found.

Secondary QCs for Interpretation

- *Fragment size* - average length of the sequenced fragments identified by read pairs mapped to a locus. Fragment size is calculated the following way: alignment length for read 1, plus alignment length for read 2, plus average distance between read 1 and read 2 based on all alignments for the read pair.
- *Read quality* - average quality of all reads mapped to a locus.
- *Other exon spot noise ratio* - maximum number of reads assumed to be systematic noise relative to the total number of reads at a certain location in the non key exon regions.
- *PCR crossover artifact ratio* - number of reads assumed to be PCR crossover artifacts relative to the total number of reads mapped to the locus.
- *Key exon mismatch count* - number of variants identified between the alleles in the result pair and the consensus sequences in key exonic regions.

Warnings for Troubleshooting

- *Read length* - average length of all reads mapped to a locus.
- *Crossmapping (intergenic ambiguity)* - number of reads mapping to multiple genes relative to the total number of reads mapped to a locus.
- *Ambiguous layout (intragenic ambiguity)* - number of reads with an ambiguous relation to other reads relative to the total number of reads mapped to a locus.
- *Non-exon spot noise ratio* - maximum number of reads assumed to be noise relative to the total number of reads at a certain location in the non exonic regions.
- *Continuous consensus* - continuous consensus sequences get a value of 1, while non-continuous consensus sequences get a value of 0.
- *Fully phased consensus* - fully phased consensus sequences get a value of 1, while not fully phased sequences get a value of 0.
- *Consensus coverage other exon minimum depth* - lowest number of reads supporting the coverage across non key exon positions.
- *Consensus coverage non-exon minimum depth* - lowest number of reads supporting the coverage across all non exon positions.
- *Other exon allele imbalance* - average ratio of the alleles for non key exon regions with consensus sequences. Values around 50%-50% indicate a balanced heterozygous result, while values close to 100%-0% indicate imbalance in the proportion of reads

derived from the two chromosomes. Homozygous results get 50%-50% assigned. Note, that it is not always possible to distinguish between homozygous regions with a small amount of contamination or artifact reads and highly imbalanced heterozygous regions.

- *Non-exon allele imbalance* - average ratio of the alleles for the non-exon region with consensus sequences. Values around 50%-50% indicate a balanced heterozygous result, while values close to 100%-0% indicate imbalance in the proportion of reads derived from the two chromosomes. Homozygous results get 50%-50% assigned. Note, that it is not always possible to distinguish between homozygous regions with a small amount of contamination or artifact reads and highly imbalanced heterozygous regions.
- *Other exon mismatch count* - number of variants identified between the alleles in the result pair and the consensus sequences in non key exons.
- *Non-exon mismatch count* - number of variants identified between the alleles in the result pair and the consensus sequences in non-coding regions (introns and UTRs).
- *Novel position count* - the number of novel positions identified throughout the gene. A high number of novel positions can indicate incorrect phasing or low consensus quality.

When only the Statistical genotyping method is used, a limited number of Quality metrics is available.

Overall

Overall QC status of the loci, calculated as the value of the worst QC measure or measures.

Primary QCs for Interpretation

- *Read count* - total number of reads mapped to a locus.
- *Key exon allele imbalance* - average ratio of the alleles for key exon regions (i.e. exons 2 and 3 for Class I genes and exon 2 for Class II genes). Values around 50%-50% indicate a balanced heterozygous result, while values close to 100%-0% indicate imbalance in the proportion of reads derived from the two chromosomes. Homozygous results get 50%-50% assigned. Note, that it is not always possible to distinguish between homozygous regions with a small amount of contamination or artifact reads and highly imbalanced heterozygous regions.
- *Genotype available* - a Yes value indicates that a genotype could be identified for a locus while No indicates that no best matching alleles were found.

Secondary QCs for Interpretation

- *Fragment size* - average length of the sequenced fragments identified by read pairs mapped to a locus.
- *Read quality* - average quality of all reads mapped to a locus.

Warnings for Troubleshooting

- *Read length* - average length of all reads mapped to a locus.
- *Crossmapping (intergenic ambiguity)* - number of reads mapping to multiple genes relative to the total number of reads mapped to a locus.
- *Other exon allele imbalance* - average ratio of the alleles for the non key exon regions. Values around 50%-50% indicate a balanced heterozygous result, while values close to 100%-0% indicate imbalance in the proportion of reads derived from the two chromosomes. Homozygous results get 50%-50% assigned. Note, that it is not always possible to distinguish between homozygous regions with a small amount of contamination or artifact reads and highly imbalanced heterozygous regions.

QC results can be exported using the *Export results* button. Note, that you can find some troubleshooting tips in the Advanced user guide section of the Omixon HLA user manual.

4.5 Data statistics

4.5.1 Overview

Read counts and proportions are available for several different steps of the analysis. In the two columns on the left "remaining reads" (i.e. reads that will be used in the next step of the analysis) are shown, while on the right "filtered reads" (i.e. reads that have been removed from the analysis for some reason) are presented. The following counts are listed:

- *Processed* - total number of processed reads from the input data files.
- *Skipped* - reads that don't map to any of the targeted loci or were skipped to balance the amount of reads between targeted loci.
- *Invalid orientation* - currently, only FR orientation is accepted, read pairs with FF or RR orientation are ignored.
- *Crossmapping* - reads mapping to multiple loci, by default these are ignored.
- *Dropped during filtering* - the total number of reads that were not aligned (this includes crossmapped reads, off-target reads, reads with invalid orientation and reads that were dropped because their pair could not be aligned to the HLA region).
- *Kept after filtering* - reads that were mapped to targeted loci and were not filtered due to any of the reasons above.
- *Lost during consensus contig generation* - reads lost due to their ambiguous relation to other reads.

- *Assembled* - total number of reads in the consensus contigs.
- *Removed as noise* - reads removed as contamination (based on supporting read proportions for variants, basically consensus fragments introducing very lowly supported variants are removed).
- *PCR crossover artifact* - reads removed as chromosomal crossover artifact. Note, that currently only artifacts within a single region are removed.
- *Used for final consensus generation* - the total number of reads used for final consensus generation.

When only the Statistical genotyping method is used, a limited number of statistics is available.

- *Processed* - total number of processed reads from the input data files.
- *Skipped* - reads that don't map to any of the targeted loci or were skipped to balance the amount of reads between targeted loci.
- *Invalid orientation* - currently, only FR orientation is accepted, read pairs with FF or RR orientation are ignored.
- *Crossmapping* - reads mapping to multiple HLA loci, by default these are ignored.
- *Dropped during alignment* - the total number of reads that were not aligned (this includes crossmapped reads, off-target reads, reads with invalid orientation and reads that were dropped because their pair could not be aligned to the HLA region).
- *Kept after alignment* - reads that were mapped to targeted loci and were not filtered due to any of the reasons above.

4.5.2 Allele imbalance

This figure shows the per-region allelic imbalance for all the genes. On the y-axis you can see the allelic imbalance measure (%) while on the x-axis, you can see all loci. The allele imbalance measure reflects the imbalance between reads originating from different chromosomes. The measure is calculated using heterozygous positions and presented as a pair of percentages. Percentage pairs, where both values are around 50% represent balanced heterozygous positions, while low-high percentage pairs mark imbalance. Homozygous regions are presented as 50%-50%. Note, that it is not always possible to distinguish between homozygous regions with a small amount of contamination and highly imbalanced heterozygous regions.

Overall allele imbalance: average allelic imbalance, only considering heterozygous regions.

4.5.3 Fragment size

This histogram shows the fragment size distribution of paired reads. (Fragment size is calculated as the length of both reads in a pair, plus the inner mate distance between R1 and R2, which can be negative if the reads overlap.) On the y-axis, you can see the number of reads and on the x-axis you can see the fragment length. Note, that the x-axis shows "binned" values: values with non-zero read counts are grouped by 10, therefore columns often (and especially with higher values) represent a relatively big "stretch" of fragment length. Again, by hovering over a column with the mouse, you can see the exact values on both the x and y-axis for that specific column.

4.5.4 Read quality

On this graph, the base quality per 5 bases is shown for the processed reads. Read positions are on the x-axis while on the y-axis quality values are shown. Besides the 5-base average, the 90th percentile (High quality) and 10th percentile (Low quality) quality values are shown. Note, that by hovering the mouse over a group of columns, the exact position quality values can be viewed.

The average quality and length for processed and aligned reads are shown. Note, that reads shorter than 75 bp are filtered out, so all processed reads are longer than this threshold (when paired data is used, both reads in every processed pair have to pass the length filter). After read processing, but before the alignment, quality based trimming is performed. Because of this step, aligned reads are usually somewhat shorter and have a little higher average quality than processed reads.

4.5.5 Troubleshooting missing results

When no alleles could be reported for a targeted gene, a markup describing the possible reason for the missing allele call is shown. For additional information, hover over the info icon next to the markup and read the tooltip. The following cases can be reported:

For non DRB3/4/5 loci:

- *Not targeted* - Meaning that the locus was not targeted in the specific analysis. This markup is a placeholder for loci which were targeted in other analyses displayed in the table.
- *No data present* - No data present means that the locus has dropped out during sequencing and should be re-sequenced.
- *Insufficient or low quality data* - There is insufficient data or the data is of low quality in the sample. Quality control results should be checked for more detail.

For DRB3/4/5:

- *Not targeted* - Meaning that the locus was not targeted in the specific analysis. This markup is a placeholder for loci which were targeted in other analyses displayed in the table.
- *Allele not expected* - There is no allele expected at this locus based on known linkage disequilibrium with HLA-DRB1 and HLA-DQB1.
- *Expected allele not found* - This markup means that based on known linkage disequilibrium information, data was expected for the locus/allele but was not found.
- *Unexpected allele found* - Data was found for a locus/allele, which was not expected based on known linkage disequilibrium information.
- *Insufficient or low quality data* - There is insufficient data or the data is of low quality in the sample. Quality control results should be checked for more detail.

When no alleles are reported for a targeted gene it is suggested to rerun the sample in question using a higher number of reads. (The number of processed reads can be set in the *Advanced Genotyping wizard*.) The reasons behind the missing allele level results can be that the coverage does not reach the minimum threshold on the allele or on the exons, or the coverage depth is too small. Processing more reads can help making the signals that support the correct alleles stronger.

4.6 Genotyping mismatch result

On this screen, a detailed list of mismatches between the selected alleles and the corresponding consensus pair are available. After clicking on a specific mismatch, you can jump to the selected mismatch in the Gene browser. You can filter mismatches based on allele status (*best match* or *assigned*) or by *chromosome*. You can export this detailed list using the *Export Results* option.

4.7 Genotyping result novelties

On this screen, a detailed list of novelties (i.e. base changes that have been applied to the allele from which the novel allele was created) is available. After clicking on a specific novelty, you can jump to the selected novelty in the Gene browser. You can filter novelties based on allele status (*best match* or *assigned*) or by *chromosome*. You can export this detailed list using the *Export Results* option.

4.8 Allele frequency information

Information from the Allele Frequencies database (<http://www.allelefrequencies.net>) is incorporated into our downloadable IMGT databases from version 3.24. It is available for the alleles that are listed on the Allele Frequencies website and can be viewed on the *Allele Frequency* tab of the *Genotype Details* window which is accessible from the *Genotyping Sample Result* screen.

5 Gene Browser

5.1 Introduction

The Gene Browser allows visual inspection of genomics data. Multiple allele candidates can be browsed together.

With default settings, the following tracks are available in the browser:

- *Position track* - Shows the coordinates for all visible tracks. Numbering starts from one.
- *Phasing track group*:

Phasing track - This track contains annotations for continuously phased regions (aka Phasing regions).

Variants track - Shows the number of overlapping read pairs between two consecutive heterozygous positions (i.e. two position where the two consensus sequences differ from each other). The 'Straight' label shows the number of reads for each consensus that supports the phasing shown in the browser, while the 'Cross' label shows the number of supporting reads for the other possible phasing of the two positions.

- *Consensus sequence 1* - The generated consensus sequence for one of the chromosomes.
- *Coverage depth for consensus 1* - Shows the depth of coverage for every position of the consensus sequence 1 *assembly*.
- *Consensus sequence 2* - The generated consensus sequence for the other chromosome.
- *Coverage depth for consensus 2* - Shows the depth of coverage for every position of the consensus sequence 2 *assembly*.
- *Allele 1 sequence* - Nucleotide sequence of the allele that matches the first consensus the best.
- *Region annotation for allele 1* - Annotations for exons, introns and UTRs are shown for allele 1.
- *Coverage depth track for allele 1* - Shows the depth of coverage for every position of the allele 1 *alignment*.
- *Allele 2 sequence* - Nucleotide sequence of the allele that matches the second consensus the best.
- *Region annotation for allele 2* - Annotations for exons, introns and UTRs are shown for allele 2.
- *Coverage depth track for allele 2* - Shows the depth of coverage for every position of the allele 2 *alignment*.

For novel alleles, two reference tracks are shown: the reference sequence of the novel allele (*Novel ref*) and the reference sequence of the closely related allele (*Rel ref*) from which the novel allele was derived.

Note that consensus sequences and the corresponding short reads can be viewed in the browser, even when no allele match pairs are found.

Additional tracks:

- *Noise track* - Shows systematic noise filtered out during consensus assembly. The noise consensus contains the major nucleotide for every position.
- *Amino acid track* - Shows the amino acid sequence for all allele and consensus sequences, including novel alleles, colored based on amino acid hydrophobicity.

The Gene Browser allows visual inspection of genomics data. Multiple allele candidates can be browsed together.

With default settings, the following tracks are available in the browser:

- *Allele 1 sequence* - Nucleotide sequence for all the defined exons of the allele selected for one of the chromosomes.
- *Region annotation for allele 1* - Exon annotations for allele 1.
- *Coverage depth track for allele 1* - Shows the depth of coverage for every position of the allele 1 *alignment*.
- *Allele 2 sequence* - Nucleotide sequence for all the defined exons of the allele selected for the other chromosome.
- *Region annotation for allele 2* - Exon annotations for allele 2.
- *Coverage depth track for allele 2* - Shows the depth of coverage for every position of the allele 2 *alignment*.

Additional tracks:

- *Amino acid track* - Shows the amino acid sequence for all allele and consensus sequences, including novel alleles, colored based on amino acid hydrophobicity.

By default, detailed coverage tracks are displayed for the allele alignments, alongside region annotations. The coverage track has a built-in base statistics visualization support: for bases in reads different from the actual consensus/reference base the corresponding coverage depth is shown with the associated nucleotide base color proportionally.

Additional modes for short read tracks

Other than the default *coverage depth mode*, the following alternative short read visualization modes are available for the short read track:

- *Short read mode* - Shows short reads displayed in a stranded fashion, so that forward strand reads (pink) and reverse strand reads (yellow) can be easily distinguished within the display.
- *Fragment mode* - Paired visualization mode that shows the corresponding forward and reverse reads in pairs in the same line. Overlapping sections between read pairs are marked with blue, while non-overlapping reads are connected with a thin line.

In both of the above modes, the short reads track can be *collapsed* which gives a summary view of the short reads (and does not allow each read to be inspected in detail).

5.2 Top panel

On the top of the screen you can see three rows of functions.

5.2.1 Top row

On the left, you can find the *Toggle fullscreen* button which hides the less important parts of the current screen, while on the right you can see the following things:

- the ID of the current user,
- the memory usage widget panel,
- the status panel of the process manager,
- the welcome tutorial button,
- the logout button,
- and the exit button.

In case of the desktop version, the actual memory usage can be observed at the top right corner of the screen. This displays the amount of currently used memory compared to the allowed maximum. Note, that the software might be allocating more memory from the operating system compared to the reported actual memory usage to make further computations faster by keeping some previously allocated memory pieces for later use, thus making the allocation of new memory unnecessary during further computations. By clicking on the memory report widget the software tries to free up as much memory as possible on a best effort basis: this is only an attempt, there are many technical factors which might block returning all memory to the operating system. It is also important to mention that by explicitly asking to free up memory the subsequent computations might become slightly slower, since the option of reusing certain structures allocated before becomes impossible. Only use this option when all planned analyses have been finished and you would like to use other programs more intensively in parallel while investigating the analysis results.

5.2.2 Middle row

There are four navigation buttons that are available on every screen of the tool. These general navigation buttons work very similar to navigation functions available in most widely used internet browsers:

- The *Back* button will take you to the previous screen.
- The *Forward* button will take you to the next screen.
- The *Up* button will move you one level up in the application.
- The *Home* button will take you to the starting screen (i.e. the Genotyping dashboard).

Right from the navigation button, in the information panel, you can find relevant high-level information about the current screen. On the *Genotyping dashboard*, you can see the release date and version of the active IMGT database.

On the far right hand side of the middle row, you can find the bookmark and context specific help buttons. The help tool shows relevant usage information and tips about functions and data available on the current screen of the application and can also be opened by pressing F1.

You can search the Help for any expression by using the Magnifier glass icon displayed on the popup window. Relevant hits are displayed in the left panel where you can select them and open them in the right hand reading pane. You can close the search by clicking the X displayed beside the search field.

5.2.3 Bottom row

The bottom row of the top panel contains a series of buttons which contain the main functions available on the screen.

5.3 Settings and Functions

5.3.1 Filtering the Allele Candidates

By clicking the *Assignment State* button, alleles can be filtered based on their assignment state. Consensuses and allele candidates can be filtered by chromosome, using the *Displayed Allele(s)* function.

5.3.2 Exporting Sequences

Export Gene Browser Tracks

DNA sequences shown in the current browser session can be exported in FASTA format using the *Export Sequences* function. By default all displayed sequences will be exported with aligned consensus sequences. Choose the *Only export consensus(es)* option to export only noise and consensus sequences.

Export Novel Sequence

This function helps to export novel sequences for use with the *GenBank BankIt* allele submission tool.

The function is only available if exactly one novel allele is displayed in the *Gene Browser*. The novel sequence will be exported from first to last fully defined region. The sequence in the generated FASTA file will always be continuous, all gaps will be omitted. In order to export a novel sequence the *Sequence ID*, *Note* and *Output fasta file* fields must all be filled in. Please be aware that in the *Sequence ID* field only letters, digits and the following characters are allowed: *, -, ., :, #.

On the *Feature table file* tab you can choose to export a feature table file alongside the FASTA. Use the *Product* field to briefly describe the protein product of the exported allele. Optionally, notes can be added in the *Note* field. *Product* and *Output feature table file* are mandatory to fill in.

5.3.3 Exporting Sequences

DNA sequences shown in the current browser session can be exported in FASTA format using the *Export Sequences* function. By default all displayed sequences will be exported.

5.3.4 Zooming

By default the browser starts in *Drag* mode, where the mouse wheel (or '+' and '-' keys) can be used to zoom in and out, and the display can be moved by clicking and dragging with the mouse. *Drill* mode can also be used, where the region highlighted by a mouse click and drag will be 'drilled into' once the mouse button is released.

5.3.5 Jumping around

You can jump to a position in the reference using *Jump To Position*. If you have previously selected a feature within the display (such as an annotation or short read), then you can jump back to that feature at any time by using the small arrow button at the bottom of the display.

5.3.6 Changing short read track settings

You can display more short reads by using the *collapsed* pile-up view. You can also *page* around the short reads using the small arrows at the top of the short read track. You can set the page size using the up and down arrows, and the current page by using the left and right arrows. This is useful when you have deep sequencing and want to scroll through a few hundred short reads at a time.

5.3.7 Track management

You can configure which tracks you would like to see in the display.

5.3.8 Display setup

You can hide or display strand, indels and soft-clips. The scaling of the coverage track can be changed between an absolute and a relative (global) scale. You can also set whether nucleotide and quality sequences for short reads should be showed on the bottom metadata panel or not.

5.3.9 Exporting Gene browser data

You can *Copy to Clipboard* the details of individual, selected items in the display. You can also copy the reference sequence belonging to a selected item (read or annotation), by clicking on the *Copy reference to clipboard* button.

You can also *Capture a Screenshot* of the currently visibly tracks within the Gene Browser. This will create a PNG image file at the chosen location.

5.3.10 Rotating the browser

The vertical Gene Browser view is more useful for comparing multiple allele candidates. The view can be rotated to use a horizontal display, which is a bit more traditional for Genome Browsers, and is more useful when browsing a single allele candidate.

5.4 Context menu

By clicking the right mouse button, you can open the Context menu. In the Gene Browser, this menu contains the following functions:

- *Previous locus*: Navigate to the previous locus in the result.
- *Next locus*: Navigate to the next locus in the result.
- *Show base statistics at cursor*: Shows the number of supporting reads for each base at the position of the cursor for all visible short read tracks. Read counts for indels are also shown.
- *Toggle cursor lock*: Locks/releases the cursor.
- *Find sequence*: Searches for a specific nucleotide sequence in the reference of all visible consensus and alleles.
- *Toggle fullscreen*: Toggles between full screen (i.e. only the browser is shown) and the normal view.
- *Toggle reference masked*: Hides all the bases shared by the different allele reference sequences on the screen. This way, only the differences between the allele candidates are shown.
- *Copy genotype*: Copies the allele names in the result allele pair.
- *Add custom allele(s) to 1st chromosome*: Adds one or more alleles of the currently selected gene from the IMGT database to the first consensus. Reads that can be aligned to this custom allele will be shown in the browser. Note, that allele candidates that are already in the candidate list for that gene cannot be added again.
- *Add allele(s) of other result pair(s) to 1st chromosome*: Adds one or more alleles of the currently selected gene from the not-displayed result candidates to the 1st chromosome. Note, that allele candidates that are already in the candidate list for that gene cannot be added again.
- *Remove custom allele(s) from 1st chromosome*: Removes one or more of the previously added custom alleles from the first consensus.
- *Add custom allele(s) to 2nd chromosome*: Adds one or more alleles of the currently selected gene from the IMGT database to the second consensus. Reads that can be aligned to this custom allele will be shown in the browser. Note, that allele candidates that are already in the candidate list for that gene cannot be added again.
- *Add allele(s) of other result pair(s) to 2nd chromosome*: Adds one or more alleles of the currently selected gene from the not-displayed result candidates to the 2nd chromosome. Note, that allele candidates that are already in the candidate list for that gene cannot be added again.
- *Remove custom allele(s) from 2nd chromosome*: Removes one or more of the previously added custom alleles from the second consensus.
- *Remove all custom alleles*: Removes all previously added custom candidates.
- *Convert custom allele pair to result genotype*: the manually added custom alleles can be converted to genotype results which allows them to be displayed and handled as results on other screens including assignment, approval, and result export.

5.5 Metadata Panel

At the bottom of the browser screen you can see the *Metadata Panel* (a.k.a. Bottom Panel). This section of the screen shows information about the currently viewed region and features. The Metadata Panel has three subpanels:

- On the left you can see the start and end positions of the currently shown region, the current cursor position and the per-base coverage at the cursor position.
- On the right, you can see information about the read/annotation at the current cursor location.
- In the middle, you can see metadata about the currently selected read or annotation. This middle subpanel also contains a set of options for exporting information about a single read or annotation. For a detailed explanation about these functions, see the list below!

5.5.1 Metadata Panel functions in the Gene Browser:

- *Copy to clipboard*: Copies the metadata of the selected read/annotation to the clipboard.
- *Copy reference to clipboard*: Copies a region of the reference sequence (or sequences) that overlaps with the selected read or annotation.
- *Jump to selection*: Centers the browser screen on the selected read or annotation.

5.6 Shortcuts

- F2 shows base statistics at cursor position.
- UP/DOWN arrow keys can be used for scrolling alongside the reference contig.
- LEFT/RIGHT arrow keys can be used for scrolling through the reads at the current position/area.
- PAGE UP/PAGE DOWN can be used for scrolling alongside the reference contig in bigger steps.
- HOME/END jumps to the beginning/end of the current reference contig.
- +/- zooms in and out.
- SPACE locks/unlocks the cursor (i.e. green line when unlocked, red line when locked).
- BACKSPACE: Show the whole reference contig (i.e. zooms out totally) or when the cursor is locked, jumps to the locked cursor.
- Mousewheel zooms in/out (in 'Drag' mode) or movea alongside the reference (in 'Drill' mode).
- SHIFT+mousewheel scrolls through the reads at the current position/area.
- Left mouse button can be used for dragging the display (in 'Drag' mode) or for highlighting and zooming in on a region (in 'Drill' mode by clicking and releasing the left button).
- CTRL+F can be used to find a nucleotide sequence in any of the shown consensus or all sequences.
- CTRL+M toggles fullscreen.
- CTRL+D turns the reference mask on and off. (With the mask on, only differences between allele and consensus sequences are shown).
- CTRL+A can be used for adding custom alleles to the first chromosome.
- CTRL+P can be used for adding alleles from other result pairs to the first chromosome.
- CTRL+R removes custom candidates from the first chromosome.
- CTRL+B can be used for adding custom alleles to the second chromosome.
- CTRL+Q can be used for adding alleles from other result pairs to the second chromosome.
- CTRL+S removes custom candidates from the second chromosome.
- X switches the browser between horizontal and vertical modes.
- C switches the browser between collapsed and uncollapsed modes.
- T switches the browser between drag and drill navigation modes.

5.7 Manage tracks

In this wizard, you can hide and show any tracks.

The visibility of any track can be set with the *Show Track(s)* and *Hide Track(s)* functions. Hidden tracks are marked with a darker grey color in the *Manage Tracks* wizard track list.



6 Settings dashboard

Reachable from the *Genotyping dashboard* using the *Application settings* button, the *Settings dashboard* displays an overview of the settings in the tool, allows access to administration features and display configurations. Some general information about the current version of the software and the current user is also available on this dashboard.

6.1 General information

There are three blocks of information on the Settings Dashboard:

- Omixon HLA edition: this part contains the name and version of the software, the build identifier with a dedicated copy to clipboard button and some contact and copyright information.
- Omixon HLA edition: this part contains the name, version and reference number of the software, the build identifier with a dedicated copy to clipboard button and some contact and copyright information.
- User info: this part contains the login name, first and last name of the current user.
- License info: this part shows the number of available credits and the expiration date of the license.

6.2 Sidebar

The left sidebar contains the following function sets:

6.2.1 General

In this function group you can set where analysis data and result files are stored, create and manage protocols, set targeted genes for analysis and select the assay version to be used for analysis. For details about protocols see the *Analysis Protocols* help page.

6.2.2 Database

With the *Install New Database* function, you can set up one or multiple versions of the IMGT database used for genotyping. With the *Select Active Database* function, you can specify the active version of the database. Genotyping will always be initiated using the active version. You can set whether or not to use database extensions in the *Configure Database Extensions* menu.

6.2.3 Administration

With the *User management* option, you can create, edit and remove users. With the *Display Hardware Key* option, you can display an alphanumeric identifier for your computer which can be used for generating a license for that specific machine. The *Upload Licence* option can be used for manually importing a license file into the software.

6.2.4 Automation

This function group allows you to configure automatic analysis on server-client configurations.

6.2.5 Export Settings

Here you can configure LIMS export.

6.2.6 Screen Settings

In this function group you can change the display configurations for the Gene Browser and result screens. Note, that these changes will modify the default behavior and appearance of the software. If you only want to temporarily change browser settings the *Display settings* option on the browser screen should be used. You can also modify the default filters of the Genotyping result screens. Be aware, that if you unselect a locus in this wizard, results for that locus won't be shown, regardless of the typing results. For both configuration sets,



you can set all the parameters back to the default values using the *Restore defaults* function. For details about these settings, please see the following help pages: *Analysis Result Screens* and *Gene Browser*.

6.3 Sample and Result Folders

Users can optionally set where their sample files are in the file system and where the analysis result files should be stored. If the user select a sample for analysis that can be found in the subfolder of the configured base directory then the result file will be created in the subfolder of the base result directory but the same folder hierarchy.

For example, if the configured values are

- base sample folder: /data/samples and
- base result folder: /data/results

and user select the /data/samples/folder1/SampleA_R1.fastq.gz for analysis. In this case the result will be created in the /data/results/folder1 directory.

Optionally user can set that the result files must be generated in the timestamped subfolders (i.e. /data/results/folder1/20151101_091102).

If the selected sample is outside of the base directory then the result will be created in the same folder as the sample is.

Note: the timestamped subfolder name format is yyyyMMdd_HHmm.

6.4 Analysis Protocols

Selecting the *Show protocol* option in the right-click context menu of a sample displays the *Protocol configuration* popup. The protocol configuration is the set of parameters which was applied in the selected genotyping run. The protocol of a specific analysis can be saved using the *Save as protocol* option in the *Genotyping dashboard* context menu.

6.4.1 Configure protocol

Selecting the *Show protocol* option in the right-click context menu of an analysis result on the Genotyping dashboard displays the *Protocol configuration* popup. The protocol configuration is the set of parameters which was applied in the selected genotyping analysis run.

6.4.2 Create protocol

Introduction

This function is available from the Analysis Protocols dashboard. You can view, edit and use protocols from here.

Protocols are stored sets of genotyping analysis parameters. Using an existing protocol can speed up the genotyping wizard, as most of the parameters will be set to those stored in the protocol. If you settle on a protocol that works well for your samples, you can save this protocol from any successful run and reuse it again later. You can also edit existing protocols, and create new ones at any time.

The Factory Default protocol is readily available after installation and is recommended to be used for genotyping any samples that have been produced with the Holotype HLA kits. Modification of this protocol is disabled for this reason - to preserve the recommended settings, so you cannot accidentally lose them.

The 'Create New Protocol' option is available anytime to design your own protocols. It's also possible to set any protocol as default.

General

You can name your new protocol here and you can mark it as a 'Default Protocol'.

IMPORTANT

The protocol marked as 'Default protocol' on the Protocol configuration popup will be used for genotyping 'by default' meaning that it will be utilized for every occasion when the 'Analyse' button or the 'Simple Genotyping' button is used. These options don't allow further configuration therefore please be considerate when setting a Protocol - other than the Factory Default - as a default.



Sample data type

You can select the data type (Single data, Paired data, Single data split to multiple files, Paired data split to multiple files. When using paired reads, the two input files are assumed to have the exact matching reads, in the exact same order (there is currently no built-in check for this).

Analysis options

These options dictate where the data has come from and effect how the underlying algorithms deal with the data. Some data sources (e.g. whole genome and whole exome) are more 'noisy', and the algorithms can help to filter out this extra noise.

- *Process all reads*: For whole exome sets, it is recommended to process all the reads (or pairs) because of the lower coverage that is usually found in these kind of data sets.
- *Maximum pairs processed per locus*: Twin uses locus based subsampling to optimize results. This value determines how many reads will be used for each locus. Note, that only reads longer than 75 bases are selected during subsampling (for paired data, both reads in the pair have to be over this minimum length threshold). You can experiment with this value to find the best match for your data. For samples produced with the Holotype assay and other targeted data sets the default 4000 read pairs per locus work well. In case you run into problematic samples increasing the read count can help resolve the issues.
- *Ignore rare alleles*: This option allows you to ignore very rare alleles in the analysis - they will not be included or reported in this case. If this option is not chosen, then these rare alleles will be marked in the results. If this option is greyed out, you will need to run *Install New database* again, in order to import the rare alleles database.
- *Novel allele detection*: This feature detects SNPs and indels in the short read allele alignments. Based on the detected variants and the original reference sequence of the specific allele, novel allele sequences are generated. The names of novel allele candidates contain the allele name of the original allele plus a hashtag and a number. For example, HLA-A*01:01:01:01#1 is a novel allele that has been generated based on the allele sequence of HLA-A*01:01:01:01 and one or more SNPs found in the short read alignment of this allele. Novel alleles have a double reference track in the HLA Genome Browser, both the reference of the novel allele candidate (Novel ref) and the reference sequence of the related allele (Rel ref) (i.e. the sequence of the allele the novel allele candidate originated from) are shown. Novel positions are accepted only when at least 90% of the reads and at least 10-fold coverage depth support the novel reference value. Note, that in many cases multiple similar novel allele candidates are generated from different alleles. There are three versions of the novel allele detection settings: you can choose not to run novel allele detection, you can choose to run novel allele detection only when exon mismatches are present (note, this will try to resolve both exon and intron novelties) or you can choose to run novel allele detection if any mismatches are present (i.e. whole gene novel allele detection).
- *Save read mappings*: Use this setting to save the mapped reads if you would like to browse them in the HLA genome browser.

Sample data type

You can select the sequencer (Illumina, Roche 454 or Ion Torrent) and the datatype (Single data, Paired data, Singla data split to multiple files, Paired data split to multiple files. When using paired reads, the two input files are assumed to have the exact matching reads, in the exact same order (there is currently no built-in check for this).

Analysis options

These options dictate where the data has come from and effect how the underlying algorithms deal with the data. Some data sources (e.g. whole genome and whole exome) are more 'noisy', and the algorithms can help to filter out this extra noise.

- *Process all reads*: For whole exome sets, it is recommended to process all the reads (or pairs) because of the lower coverage that is usually found in these kind of data sets.
- *Maximum pairs processed per locus*: Explore uses locus based subsampling to optimize results. This value determines how many reads will be used for each locus. Note, that only reads longer than 75 bases are selected during subsampling (for paired data, both reads in the pair have to be over this minimum length threshold). You can experiment with this value to find the best match for your data. For samples produced with the Holotype assay and other targeted data sets the default 4000 read pairs per locus work well. In case you run into problematic samples increasing the read count can help resolve the issues.
- *Ignore rare alleles*: This option allows you to ignore very rare alleles in the analysis - they will not be included or reported in this case. If this option is not chosen, then these rare alleles will be marked in the results. If this option is greyed out, you will need to run *Install New database* again, in order to import the rare alleles database.
- *Save read mappings*: Use this setting to save the mapped reads if you would like to browse them in the HLA genome browser.

6.5 Install New Database

This only needs to be run once for every IMGT database version. On the first startup an initial import process automatically imports all previously released IMGT databases. Additional database files can be downloaded from the Internet and installed by selecting them from the provided list. This should only take a few seconds.

For most of the databases a database extension (i.e. additional sequence information for some of the alleles) is available. Database extensions can be configured using the *Configure Database Extensions* option on the Settings dashboard.

6.5.1 Download problems

If for some reason you cannot download the database archive from within the tool (e.g. due to network security settings) you can download the archives manually and use the *Select local file* option instead of the *Download file* option in the wizard.

The following database versions can be downloaded:

- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.10.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.14.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.15.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.16.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.17.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.19.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.20.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.21.0.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.24.0_1.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.24.0_2.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.25.0_3.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.26.0_3.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.27.0_3.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.28.0_4.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.29.0_1_5.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.30.0_5.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.31.0_5.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.32.0_5.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.32.0_7.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.33.0_7.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.34.0_8.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.35.0_8.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.36.0_8.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.37.0_8.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.38.0_8.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.38.0_9.oxdb.gz
- http://omixon-download.s3.amazonaws.com/dbs/omixon_3.39.0_9.oxdb.gz


6.6 Select Active Database

With this function, you can select the active IMGT database version. This active version will be used for new genotyping analyses. For existing analyses, the database version used during the previous analysis must be installed but doesn't have to be selected as active, therefore previous results can be viewed, regardless of the active IMGT database version.

If a database extension (i.e. additional sequence information for some of the alleles) is available for the selected database version and extensions are not turned on, a wizard is shown after activating the database, so usage of allele extensions can be turned on or off. Database extensions can also be configured at any time using the *Configure Database Extensions* option on the *Settings dashboard*.

6.7 Configure Database Extensions

The database extension contains additional partial sequences for some of the alleles in the IMGT database. These additional sequences are not yet part of the official IMGT database most likely due to technical reasons (e.g. a full allele sequence is not available), but the information

they contain have been published and is well known within the HLA community. If the database extension is turned on (i.e. "Use genotype reference extension database?" is checked), the extended sequences are added to the official IMGT database and are used during the genotyping workflow. On the graphical user interface, extended alleles are clearly marked with a *plus sign* .

List of extended alleles in the different databases:

3.24.0_1 and 3.24.0_2:

- DRB4*01:03:01:02N (intron 1 sequence)

3.25.0_3

- C*06:116N (intron 3 sequence)
- DPB1*105:01 (intron 2 sequence)
- DRB1*09:01:02 (intron 2 and intron 3 sequences)
- DRB4*01:03:01:02N (intron 1 sequence)

3.28.0_4

- C*06:116N (intron 3 sequence)
- DRB1*09:01:02 (intron 2 and intron 3 sequences)
- DRB4*01:03:01:02N (intron 1 sequence)

3.29.0.1_5, 3.30.0_5, 3.31.0_5 and 3.32.0_5

- C*06:116N (intron 3 sequence)
- DRB4*01:03:01:02N (intron 1 sequence)

3.32.0_7 and 3.33.0_7

- C*06:116N (intron 3 sequence)
- DRB4*01:03:01:02N (intron 1 sequence)
- HLA-DQB1*03:276N (pseudo 5' UTR sequence)

3.34.0_8, 3.35.0_8, 3.36.0_8, 3.37.0_8, 3.38.0_8, 3.38.0_9 and 3.39.0_9

- C*06:116N (intron 3 sequence)
- HLA-DQB1*03:276N (pseudo 5' UTR sequence)

6.8 User Management

User Management allows User accounts to be added, edited and deactivated. There must be at least one 'Super User' at all times. The first registered user automatically becomes the Super User, this user cannot be deactivated.

The following properties must be set for all users:

- User (login name)
- Password
- User role

The following properties are optional:

- First name
- Last name
- Department

There are currently two user roles within the application:

- Analyst: this role is set up for standard activities related to sample analysis including Holotype HLA typing, View/Use Protocol, Export functions.
- SuperUser: authorised to use all functions in the application. On top of the Analyst role's permissions it also includes Administrative functions, Advanced HLA typing, Import function and Create/Save Protocol.

6.9 Upload Licence

Use this function to upload a new license which extends the number of credits or time available in the application. Please make sure that you do not modify the license file during download. Simply save it to your machine then browse and select the file for uploading then click Finish to complete the operation. A pop up message informs you if the upload was successful. Afterwards the new credits and available time will be immediately displayed on the Settings screen. In case the upload is not successful an error message informs you about the possible causes. If the license has accidentally been modified you can simply download it again from the provided link and repeat the above steps. If



the hardware key does not seem to match please check that the key under "Settings / Display Hardware Key" matches the one previously sent to the Omixon support team.

If the correct key has been provided or you get a different error message or have any questions regarding the described process please contact support@omixon.com⁴. For detailed information about licensing and pricing, please contact sales@omixon.com⁵.

6.10 Configure Automatic Analysis

You can configure Automation from the Application Settings *Configure Automatic Analysis* menu.

Automation recursively parses all subdirectories of the supplied parent directory for new *CompletedJobInfo.xml* files, which are generated after sequencing runs are completed. Directory parsing is triggered every 30 seconds by default. To override the default behavior you have the option to supply up to three hour values separated by commas. When doing so, automation only checks the folders once at each given time.

The *Parent directory* is the root folder where the analysis results are generated to. Please make sure you enter your own parent directory as it is specific to your computer.

The specific location of the sequencing results inside the supplied parent directory is not of significance, the automation service will find them in any subfolder. Set the *Autostart service* flag if you want Automation to start automatically when restarting the Omixon Server.

On the *Select automation protocol* screen you can select the protocol to be used by the Automation. If no protocol was selected the default protocol is used.

6.11 Gene Browser

You can define the starting (i.e. used when the software is freshly started) graphical and navigation settings for the browser here. These settings can be saved and will still be in effect after restart. The browser settings are user specific.

6.11.1 Display configuration

On this screen the orientation of the browser screen (horizontal or vertical), the navigation mode (*Drag & Zoom* or *Drill & Scroll*) and the short read pileup visualisation settings can be changed.

6.11.2 Color setup

On this screen, the color codes for nucleotides used in the browser can be changed.

6.11.3 Data options

On this screen you can select whether inserts or deletions should get an extra mark in zoomed out views of the browser. Soft clips and strand information can be shown or hidden as well. The default scale for the coverage tracks can also be changed on this screen.

6.11.4 Metadata options

With these settings, you can show/hide read sequences and quality information on the metadata panels in the browser (i.e. the middle and right subpanels in the bottom browser panel).

6.11.5 Advanced settings

The zooming resolution of the Gene browser can also be set on this screen by changing the *Overview factor*. This is the maximum number of nucleotides displayed per pixel. The default setting is 20. The minimum number of pixels for collapsed read width can be set as well.


⁴ <mailto:support@omixon.com>

⁵ <mailto:sales@omixon.com>

7 Detailed automation guide

7.1 Overview

Omixon HLA Twin provides an integrated automated genotyping solution. Automation can be configured and managed from the user interface. Omixon HLA Explore provides an integrated automated genotyping solution. Automation can be configured and managed from the user interface.

 Automation is only available in client-server mode. You need to have admin permission if you would like to configure and manage it.


7.2 Configuration

You can configure Automation from the Application Settings *Configure Automatic Analysis* menu.

Automation recursively parses all subdirectories of the supplied parent directory for new *CompletedJobInfo.xml* files, which are generated after sequencing runs are completed. Directory parsing is triggered every 30 seconds by default. To override the default behavior you have the option to supply up to three hour values separated by commas. When doing so, automation only checks the folders once at each given time.

The *Parent directory* is the root folder where the analysis results are generated to. Please make sure you enter your own parent directory as it is specific to your computer.

The specific location of the sequencing results inside the supplied parent directory is not of significance, the automation service will find them in any subfolder.

 **Configure Automatic Analysis**

Folders settings | Select automation protocol

You can configure folders are watched by the automation service.


Parent directory for all run folders


+


Hours of daily analysis schedule (0-23) ?


Autostart service


☐

 Help

 Previous

 Next

 Cancel

 Finish

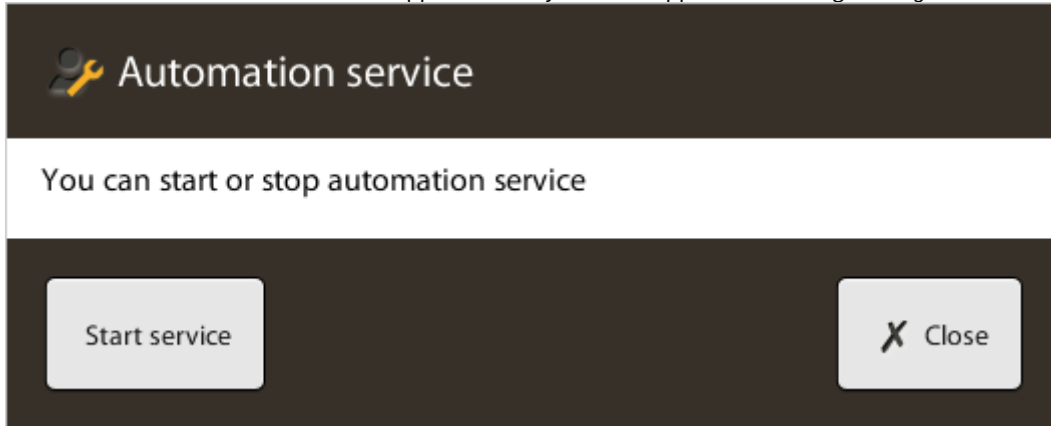
If you want Automation to start automatically when you restart the Omixon Server then set the *Autostart service* flag.

On the *Select automation protocol* screen you can select the protocol that should be used by the Automation. If no selection then the default protocol is used.

You can optionally delay the automation if your analysis files are written out in undetermined order - for example the *CompletedJobInfo.xml* marker file is written out before the last sample file is. In this case maybe not all of your sample batch is being analyzed. Setting the - *Domixon.automation.delay=120* parameter in the *omixon-hla-twin-server.vmoptions* file is used for delaying the automation by 120 sec. Setting the - *Domixon.automation.delay=120* parameter in the *omixon-hla-exp-server.vmoptions* file is used for delaying the automation by 120 sec.

7.3 Management

Automation service can be started or stopped manually from the Application Settings *Manage Automation Service* menu.



7.4 Logging

Automation generates log entries into the OMIXON_TWIN_INSTALL/logs/automation.log file. Automation generates log entries into the OMIXON_EXPLORE_INSTALL/logs/automation.log file.



8 Notifications

8.1 IMGT Database

In accordance with ASHI guidelines Omixon HLA displays a warning message if the IMGT Database used for genotyping is older than 1 year. In order to switch to a more up-to-date database you can use the database management functions available on the Application Settings Dashboard.

8.2 License Expiry

In order to help you keep your software running warnings are provided as the license approaches its expiry date. A notification will be displayed one month to license expiry and after every login within this time period.

When the notification is shown, you can choose to hide this message and it will not be shown again or you can choose to proceed without turning the notification off.

8.3 Extended allele notification

When a new database is installed in which extended allele sequences are available a notification is shown for allowing the usage of extended allele sequences. Note, that the usage of extended alleles is suggested as they can help resolving common ambiguities. You can find the list of extended alleles in the Configure Database Extensions section.

8.4 Deviating from the Factory default protocol

A notification is shown when not the Factory default is used. Please be aware, that the use of the Factory Default protocol is strongly suggested for analyzing Holotype data as all the parameters are fine tuned for this data type.



9 HTTPS support

Omixon HLA Twin supports secure connection via SSL between its nodes. The configuration details are in the Software Installation Guide document that you have received in the email with the installation files. For further information, please contact us at support@omixon.com.

10 Application performance tuning

The Omixon HLA application is implemented in the Java language. The installed Omixon HLA application contains the optimal configuration for all use cases but there are some tweaks that can be useful in case the application does not perform as well as expected. Sometimes application tuning requires only changing JVM options but there are other options. This document provides some tips and tricks that can help you to improve application performance.

Run fewer programs at the same time

Genotyping is a CPU intensive task and requires more memory than a "normal" application.

Close unused programs while Omixon HLA application runs to free up CPU and memory.

Use SSD instead of HDD

Sample and analysis result files are larger than a "normal" file, so it is crucial to decrease the I/O operations on a physical disk.

Use Solid-state drives (SSD) that are inherently faster than a hard disk drive (HDD).

- File opening speed: SSD up to ~30% faster than HDD
- File copy / write speed: SSD generally above 200 MB/s and up to 550 MB/s for cutting edge drives, and in case of HDD the range can be anywhere from 50 – 120MB/s.

Local disk vs network storage

The Omixon HLA application massively reads and writes mounted physical disks where your samples and result files are stored. Storages have inherent *latency* that refers to the increasing amount of time it takes to access a given file on the hard disk drives but in case of Network-Attached Storages (NAS) we have to calculate the distances between the files and the applications. These distances can cause delays over a wide-area-network (WAN) because the data is stored far away from where it is being used.

If possible, use local disks instead of NAS for minimizing this latency.

Too many files in one folder

When you open a folder, the Omixon HLA application parses the files - sample and result ones - before generating the file list for the user. If the folder contains too many files, the execution time is longer.

If possible, store a maximum of around 100 samples and 100 result files in one folder.

You can use file filters to get better response time, because less information is generated and sent by the application to the clients.

Distribute load on multiple CPUs

Genotyping is a CPU intensive task and by default all available CPUs are calculated and roughly 25% of available threads are kept free to provide sufficient resources for the Omixon HLA reporting functions while the genotyping process is running. Optionally, you can set the `hla.general.threadLimit` parameter in the `vmoption` file of the application to change this default behavior. If this value was set to 0, all CPUs are bound to the genotyper. If the value is set below 0 thread limit it is dynamically calculated by the application i.e. at least 2 CPU cores should not be used by the genotyper.

Set up targeted loci

You can decrease the execution time of the genotyping process if you set the targeted loci manually to match the set of loci that was sequenced. In this case filtering and analysis will only be executed for the configured loci. The advanced typing wizard can be used for specifying the list of targeted loci for a single analysis task.

Set up per locus read count

Using the default read count for Holotype samples is strongly recommended as these read counts were determined based on the characteristics of the assay and targeted loci. Running with a higher read count is only suggested for problematic loci. Note, that higher read counts can increase the memory requirements significantly.



11 Omixon database relocation

Omixon HLA application metadata is stored in a MySQL database that was configured upon the first installation of HLA Twin.

If you wish to transfer this data to another MySQL instance (or create backups), please follow usual MySQL migration procedures, or contact us at support@omixon.com



12 Fastq and BAM filter tool

In case of WXS/WGS data belonging to relevant HLA loci has to be filtered out before the execution of the genotyping process. Filtering can be executed from *Omixon HLA Explore* or can be done using the *Omixon Filter Tool* standalone application.

Another use case is the filtering of the samples generated by Illumina NextSeq sequences. In both cases we try to select reads that can be aligned to the IMGT/HLA reference database.

12.1 Installation

You need to install Oracle JDK/JRE 8 before you can start the Filter Tool (<http://www.oracle.com/technetwork/java/javase/downloads/index.html>).

The Filter Tool is packaged as *pure java* application, this means that no installation is required; it is a bundle that contains relevant classes, dependencies and start scripts. You just extract this into a custom folder.

⚠ Installation

We assume that the install folder is `C:\omixon\filter` in case of Windows and `/home/user/omixon/filter` in case of Linux.

After the installation you should set the `JAVA_HOME` environment variable in the start script. For example:

Linux startup script

```
#!/bin/sh
# Set JavaHome if it exists
if [ -f "${JAVA_HOME}/bin/java" ]; then
    JAVA=${JAVA_HOME}/bin/java
else
    JAVA=java
fi
export JAVA
PRG="$0"
OMIXON_HOME=`dirname "$PRG"`
export OMIXON_HOME
"$JAVA" -server -cp "$OMIXON_HOME/lib/*" com.omixon.hla.launcher.HlaFilterStarter $*
```

or in case of Windows

Windows startup script

```
REM Omixon HLA Filter Tool
IF DEFINED JAVA_HOME (
    SET JAVA=%JAVA_HOME%\bin\java
) ELSE (
    SET JAVA=java
)
set OMIXON_HOME=%cd%
%JAVA% -server -cp %OMIXON_HOME%\lib\* com.omixon.hla.launcher.HlaFilterStarter %*
```

12.2 Filter Tool execution

You can start the *Filter Tool* from command line by calling the `omixon-filter.sh` (or `.bat`).

Execution

```
~/omixon/filter $ ./omixon-filter.sh /mnt/hgfs/genom /mnt/hgfs/omixon/hla_reference_3.25.0_3.zip
```

where the

- 1st parameter defines the folder where the BAM/FASTQ sample files are and
- 2nd one defines the HLA reference bundle that should be used by the filtering (bundles can be downloaded from <http://omixon-download.s3.amazonaws.com>⁷).

Optionally a custom filtering configuration can be set as the 3rd parameter.

JVM settings

Optionally JVM parameters can be set; i.e. max heap size: `-Xmx7000m` (It needs to be set in the starter script file!)

All sample files in the target folder will be processed by the *Filter Tool*, regardless of file size.

The execution time depends on your hardware (CPU + memory), the size of the samples and locality of them (local or network storage).

If the generated paired sample files contain read pairs in different order - or both of them contain different reads - the common ones will be selected and read pairs will be written in the same order. This is done automatically by the built-in deinterlacer.

The filtered files are generated in FASTQ format in the same directory where the sample files are. The file name contains a timestamp value for tracing the execution time.

12.3 NextSeq Filter Tool execution

As the NextSeq sequencer can produce a very high amount of reads for the targeted genes and the distribution of data in the NextSeq files can be uneven (big blocks of data for one locus, not really mixed with data from other loci) a slightly different filtering approach is required. Additionally NextSeq reads can be shorter than Illumina Miseq/HiSeq machines can generate.

NextSeq sample

The *NextSeq Filter Tool* does not support multipart FASTQ file format. We assume the input samples are paired ones based on the Illumina naming convention.

The Illumina `bcl2fastq` tool can generate the expected FASTQ format by using the `--no-lane-splitting` flag. See more details in the *bcl2fastq* documentation: http://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html

If the files were generated without the flag, they have to be concatenated manually before the filtering step.

You can start the NextSeq Filter tool from command line by calling the `omixon-nextseq-filter.sh` (or `.bat`).

Execution

```
~/omixon/filter $ ./omixon-nextseq-filter.sh /mnt/hgfs/genom /mnt/hgfs/omixon/hla_reference_3.24.0_2.zip /omixon/filter/filter.conf
```

⁷ http://omixon-download.s3.amazonaws.com/omixon_hla_genotyping_database_3_25_0_3.zip

where the first two parameters are the same as defined in the *Omixon Filter Tool* section and the third one defines the custom filter configuration.

The filtered files are generated in FASTQ format in the same directory where the sample files are. The file name starts with a 'filtered' prefix. If you re-run the *NextSeq Filter Tool* the previous file is overwritten.

12.4 Custom configuration

In case of WXS/WGS samples no additional configuration is required.

If you would like to filter only one gene (i.e. HLA-A) you should set the following configuration:

Filter all reads

```
hla typer.targetedGenes = [HLA-A]
hla sample.maxFilteredRecordsPerLocus=20000
```

where

- the *targetedGenes* defines the genes should be filtered and
- *maxFilteredRecordsPerLocus* defines the maximum number of reads should be collected (default is 0 that means ALL aligned reads of the gene are collected).

The last parameter is used as a terminal condition of the execution: if the required number of reads are collected then filtering is finished. NextSeq samples can be prefiltered using the following configuration:

Filter NextSeq sample

```
hla typer.targetedGenes = [HLA-A, HLA-B, HLA-C, HLA-DQB1, HLA-DRB1, HLA-DQA1, HLA-DPB1]
hla sample.maxFilteredRecordsPerLocus=20000
hla sample.maxFilteredIgnoredRecords=100000
```

In this case *NextSeq Filter Tool* will select reads that can be aligned to the *targetedGenes* and if all required number of reads are collected or we dropped/ignored the configured number of reads then the filtering is finished/terminated.

In case of *maxFilteredIgnoredRecords* is 0 then ALL reads of the sample will be processed, so the execution time will be longer.

Parameters

Values of the *maxFilteredRecordsPerLocus* and *maxFilteredIgnoredRecords* depend on the characteristics of your samples. It is possible that the *Omixon Twin/Explore* will drop some the read pairs while genotyping i.e. because of one of the read pair cannot be aligned correctly or shorter than supported, etc.

12.5 Logging

While filtering is being executed the actual state of the process is logged in the logs directory of the tool. If you find any problem in the filtering please firstly check the content of the log file.

13 PacBio standalone tools

13.1 HDF5 file converter

Omixon PacBio HDF5 converter can be used for converting HDF5 sample files to FASTQ format. If you would like to use Omixon's converter tool, please contact Support (support@omixon.com).

13.1.1 Installation

The tool contains all dependencies - libraries and Java Runtime Environment (JRE) - and executables (shell/batch scripts). The tool only needs to be uncompressed into a folder.

13.1.2 Converter execution

The tool can be started from command line by calling the *omixon-pacbio-hdf5.sh* (in case of Linux environment) or *omixon-pacbio-hdf5.bat* (in case of Windows environment).

```
~/omixon/pacbio $ ./omixon-pacbio-hdf5.sh /data/genom/pb /data/output/sample.fastq
```

where the

- first parameter defines the folder where the .bax.h5 sample file is (folder is not parsed recursively, so only one sample can be converted per execution)
- second parameter is the output file (FASTQ).

If the extension of the output file is '.gz' then the file is automatically compressed into GZIP format.

⚠ Compression is useful if you want to use less disk space, but the compression time in case of large file is much higher than the conversion itself!

Logging messages of the converter tool is redirected into the *log* subfolder of the tool.

13.2 Demultiplexer

The ability to barcode samples reduces the cost for sample preparation and sequencing. Currently, we support only demultiplexing dataset that are barcoded using symmetric mode that means barcode sequences are the same on both sides of the insert. If you would like to use Omixon's demultiplexer tool, please contact Support (support@omixon.com⁸).

13.2.1 Installation

The tool contains all dependencies - libraries and Java Runtime Environment (JRE) - and executables (shell/batch scripts). The tool only needs to be uncompressed into a folder.

13.2.2 Demultiplexer execution

The tool can be started from command line by calling the *omixon-pacbio-demux.sh* (in case of Linux environment) or *omixon-pacbio-demux.bat* (in case of Windows environment).

⁸ <http://omixon.com>



```
~/omixon/pacbio $ ./omixon-pacbio-demux.sh /data/input/sample.fastq /data/input/  
barcodes.txt /data/output ./demux.conf
```

where the

- 1st parameter defines the FASTQ sample that should be demultiplexed
- 2nd parameter defines the barcode file - see the format below
- 3rd parameter is the output folder where the demultiplexed sample should be generated
- 4th parameter is optional; you can set it if a custom configuration should be used by the demultiplexer - see the sample config file later

The execution time of the demultiplexing process depends on the number of samples that needs to be demultiplexed and the available number of CPUs.

Logging messages of the converter tool is redirected into the *log* subfolder of the tool.

13.3 Barcode file format

Currently barcode files in the following format are supported:

```
sample1:TCAGACGATGCGTCAT  
sample2:CTATACATGACTCTGC  
sample3:TACTAGAGTAGCACTC  
sample4:TGTGTATCAGTACATG  
sample5:ACACGCATGACACACT  
sample6:GATCTCTACTATATGC  
sample7:ACAGTCTATACTGCTG
```

The name of the demultiplexed file is generated using the first part of the barcode configuration in the following format: *sample1-sample1.fastq*

13.3.1 Custom configuration

The user can customize the demultiplexer configuration by modifying the default values in the configuration file.

```
# Size of the leading and trailing subsequences of the read - default 75
hla.pacbio.demultiplexing.subsequenceLength = 75
# Minimum/maximum/default score values used during the demultiplexing process
# Values must be integer type between 0-16
hla.pacbio.demultiplexing.minThreshold = 6
hla.pacbio.demultiplexing.maxThreshold = 12
hla.pacbio.demultiplexing.threshold = 10

# The number of collected reads can be limited by setting this parameter. Zero means no
limit.
hla.pacbio.demultiplexing.readCountLimit = 0

# Defines the minimum read length of the processed reads; shorter reads are skipped - default
3000
hla.pacbio.demultiplexing.readMinLength = 3000

# Size of the subsequence of the read that is used by the pairwise alignment - default 30
hla.pacbio.demultiplexing.windowSize = 30

# Option to save unmatched reads. Name of the file is unmatched.fastq - default false
hla.pacbio.demultiplexing.writeUnmatched = false
```

14 Glossary

- **alignment score:** a quality score that is calculated for every alignment of a short read. The alignment score is basically based on the mismatches between the read and reference sequences.
- **allele:** a sequence version of a gene.
- **allele search:** the selection of allele pairs based on the generated consensus sequence and the IMGT/HLA database (Robinson et al. 2015).
- **ambiguity:** multiple alleles or allele pairs are called. For the statistical genotyping method these are supported by a nearly equal amount of reads, while in the consensus genotyping results multiple allele pairs can have the equal amount of mismatches due to e.g. off-target differences between the alleles or unresolved phase in the consensus.
- **assembly:** aligning and merging the fragments or short reads produced from a longer sequence in order to reconstruct the original sequence.
- **CG:** consensus based HLA genotyping method.
- **contig:** contiguous sequence, the result of de novo short read assembly.
- **continuous alignment:** a short read alignment that is generated for continuous allele region groups or consensus contigs.
- **de novo assembly:** short read assembly that doesn't use any information provided by a reference sequence (or sequences).
- **genotype:** one (or more) allele pair(s) for an HLA locus in a single sample.
- **interloci crossmapping:** reads map to multiple HLA loci (e.g. to HLA-DRB1 and HLA-DRB3).
- **locus (plural: loci):** the location of a gene.
- **long range PCR:** long range polymerase chain reaction is a type of PCR that was optimized for amplifying long (~2 to ~15 kb) DNA fragments.
- **mismatch:** difference between the consensus and reference allele sequences.
- **MSA:** multiple sequence alignment. An alignment of three or more biological sequences.
- **novelty:** a mismatch that haven't been previously described in the allele definitions of the IMGT/HLA database (Robinson et al. 2015).
- **PCR crossover artifact:** reads generated from a hybrid PCR product that contains information from multiple PCR amplicons often derived from different copies of the chromosome (e.g. see Holcomb et al. 2014).
- **phasing:** separation of signals coming from the two copies of the chromosomes (in this case chr6).
- **phase break:** a position between two fully phased subsequences between which phasing cannot be resolved (e.g. due to a long homozygous stretch or PCR crossover artifacts).
- **pooled (sample):** after the long range PCR step of the protocol, the amplicons of the different loci are mixed together. The result is a single pair of fastq files that contain reads from all targeted loci.
- **QC measures:** quality control measures that can indicate sequencing, data and result quality issues.
- **QC alignment:** short read alignment for the alleles that were selected based on the consensus.
- **random noise:** randomly distributed sequencing errors in the short reads.
- **region:** gene region: exon, intron or UTR.
- **SBT:** sequence based typing. Usually provides low resolution HLA types, as in most protocols, only exon 2 and exon 3 are sequenced using Sanger sequencing by capillary electrophoresis.
- **segmented alignment:** a non-continuous short read alignment that is generated independently for each region.
- **serological typing:** antibody based HLA typing method. Only allows low resolution typing.
- **SG:** statistics based HLA genotyping method.
- **splicing:** is a modification of the pre-mRNA in which introns are removed and exons are joined.
- **splicing ambiguity:** in some cases based solely on the consensus sequence and the allele definitions in the IMGT/HLA database the exact gene region coordinates for the consensus cannot be determined.
- **SSO:** sequence specific oligonucleotide based HLA typing. Uses hybridization probes. Only provides low resolution typing.
- **SSP:** sequence specific primer based HLA typing. PCR based technique that uses a list of primer pairs specific for alleles or allele groups. Only provides low resolution HLA typing.
- **strand bias:** an imbalance between the number of reads mapping to the forward and reverse strand.
- **systematic noise/artifact:** a group of reads that contain a systematic error that differs from both consensus sequences.
- **targeted locus (plural: targeted loci):** a gene or a set of genes for which sequencing was attempted.
- **unpooled (sample):** sequencing data generated using the long range PCR amplicons of a single locus. The result is a single pair of fastq files that contain reads from a single targeted locus.
- **variant support:** the number of reads supporting the reference and alternate bases at a heterozygous position.
- **allele:** a sequence version of a gene.
- **allele search:** the selection of allele pairs based on the IMGT/HLA database (Robinson et al. 2015).
- **ambiguity:** multiple alleles or allele pairs are called. With the statistical genotyping method these are supported by a nearly equal amount of reads.
- **genotype:** one (or more) allele pair(s) for an HLA locus in a single sample.

- **interloci crossmapping:** reads map to multiple HLA loci (e.g. to HLA-DRB1 and HLA-DRB3).
- **locus (plural: loci):** the location of a gene.
- **long range PCR:** long range polymerase chain reaction is a type of PCR that was optimized for amplifying long (~2 to ~15 kb) DNA fragments.
- **mismatch:** difference between the consensus and reference allele sequences.
- **MSA:** multiple sequence alignment. An alignment of three or more biological sequences.
- **phase break:** a position between two fully phased subsequences between which phasing cannot be resolved (e.g. due to a long homozygous stretch or PCR crossover artifacts).
- **pooled (sample):** after the long range PCR step of the protocol, the amplicons of the different loci are mixed together. The result is a single pair of fastq files that contain reads from all targeted loci.
- **QC measures:** quality control measures that can indicate sequencing, data and result quality issues.
- **random noise:** randomly distributed sequencing errors in the short reads.
- **region:** gene region: exon, intron or UTR.
- **SBT:** sequence based typing. Usually provides low resolution HLA types, as in most protocols, only exon 2 and exon 3 are sequenced using Sanger sequencing by capillary electrophoresis.
- **segmented alignment:** a non-continuous short read alignment that is generated independently for each region.
- **serological typing:** antibody based HLA typing method. Only allows low resolution typing.
- **SG:** statistics based HLA genotyping method.
- **splicing:** is a modification of the pre-mRNA in which introns are removed and exons are joined.
- **SSO:** sequence specific oligonucleotide based HLA typing. Uses hybridization probes. Only provides low resolution typing.
- **SSP:** sequence specific primer based HLA typing. PCR based technique that uses a list of primer pairs specific for alleles or allele groups. Only provides low resolution HLA typing.
- **strand bias:** an imbalance between the number of reads mapping to the forward and reverse strand.
- **targeted locus (plural: targeted loci):** a gene or a set of genes for which sequencing was attempted.
- **unpooled (sample):** sequencing data generated using the long range PCR amplicons of a single locus. The result is a single pair of fastq files that contain reads from a single targeted locus.
- **variant support:** the number of reads supporting the reference and alternate bases at a heterozygous position.

15 List of shortcuts

15.1 Generic shortcuts

- F1 key opens the Help for the current page or wizard.
- F8 key closes the window and exits the application
- F11 key switches to fullscreen display mode to utilize as much space as possible for visualization - note that certain platforms and window managers are not working properly together with this function, the first time you use you will get a warning message about this.
- ALT+F4 closes the application.
- CTRL + C/V/X copies/pastes/cuts texts or files.
- SHIFT/CTRL + left click can be used for multiple selection/deselection
- Left/Right/Up/Down arrow buttons can be used for scrolling on the screen if necessary.
- DEL can be used for deleting the selected item(s).
- F9 or CTRL-PLUS: scale up window contents
- F10 or CTRL-MINUS: scale down window contents
- F12 or CTRL-ZERO: reset window contents scaling

15.2 Genotyping dashboard

- CTRL+A can be used for selecting all samples and analyses.
- CTRL+F opens the file filter wizard.
- CTRL+R expands/reduces sample and analysis names.
- CTRL+H shows/hides previous analyses.
- DEL deletes the selected items.

15.3 Genotyping analysis result

- CTRL+C, after the selection of a sample, this shortcut copies the selected genotype to the clipboard.
- F2 key puts the cursor to the next sample with a failed locus.
- F3 key puts the cursor to the next sample with a discordant locus.
- F4 key puts the cursor to the next sample with a QC failed locus.
- CTRL+mousewheel can be used for zooming in and out.
- CTRL+C, after the selection of a sample, this shortcut copies the selected genotype to the clipboard.
- F2 key puts the cursor to the next sample with a failed locus.
- F4 key puts the cursor to the next sample with a QC failed locus.
- CTRL+mousewheel can be used for zooming in and out.

15.4 Genotyping sample result

- CTRL+C, after the selection of a genotype, this shortcut copies the selected genotype to the clipboard.

15.5 Gene browser

- F5 jumps to the same locus of the previous sample.
- F6 jumps to the same locus of the next sample.
- F2 shows base statistics at cursor position.
- UP/DOWN arrow keys can be used for scrolling alongside the reference contig.
- LEFT/RIGHT arrow keys can be used for scrolling through the reads at the current position/area.

- PAGE UP/PAGE DOWN can be used for scrolling alongside the reference contig in bigger steps.
- HOME/END jumps to the beginning/end of the current reference contig.
- +/- zooms in and out.
- SPACE locks/unlocks the cursor (i.e. green line when unlocked, red line when locked).
- BACKSPACE: Show the whole reference contig (i.e. zooms out totally) or when the cursor is locked, jumps to the locked cursor.
- Mousewheel zooms in/out (in 'Drag' mode) or move alongside the reference (in 'Drill' mode).
- SHIFT+mousewheel scrolls through the reads at the current position/area.
- Left mouse button can be used for dragging the display (in 'Drag' mode) or for highlighting and zooming in on a region (in 'Drill' mode by clicking and releasing the left button).
- CTRL+F can be used to find a nucleotide sequence in any of the shown consensus or all sequences.
- CTRL+M toggles fullscreen.
- CTRL+D turns the reference mask on and off. (With the mask on, only differences between allele and consensus sequences are shown).
- CTRL+A can be used for adding custom alleles to the first chromosome.
- CTRL+P can be used for adding alleles from other result pairs to the first chromosome.
- CTRL+R removes custom candidates from the first chromosome.
- CTRL+B can be used for adding custom alleles to the second chromosome.
- CTRL+Q can be used for adding alleles from other result pairs to the second chromosome.
- CTRL+S removes custom candidates from the second chromosome.
- X switches the browser between horizontal and vertical modes.
- C switches the browser between collapsed and uncollapsed modes.
- T switches the browser between drag and drill navigation modes.
- CTRL+F can be used to find a nucleotide sequence in any of the shown sequences.
- CTRL+M toggles fullscreen.
- CTRL+D turns the reference mask on and off.
- CTRL+A can be used for adding custom alleles to the first chromosome.
- CTRL+P can be used for adding alleles from other result pairs to the first chromosome.
- CTRL+R removes custom candidates from the first chromosome.
- CTRL+B can be used for adding custom alleles to the second chromosome.
- CTRL+Q can be used for adding alleles from other result pairs to the second chromosome.
- CTRL+S removes custom candidates from the second chromosome.
- X switches the browser between horizontal and vertical modes.
- C switches the browser between collapsed and uncollapsed modes.
- T switches the browser between drag and drill navigation modes.

16 Acknowledgements

16.1 Collaborators

We would particularly like to thank Dr. Derek Middleton for providing access to the Allele Frequency Net database.

16.2 Third party tools and databases

The IMGT/HLA database is used for HLA genotyping (1, 2).

Linkage disequilibrium data used in the software was derived from NMDP Registry Haplotype Frequencies database (3) and from Wikiversity (4).

Allele frequency information was derived from the Allele*Frequency Net database (5).

A number of tools from SamTools Picard are used within the software for handling and manipulating SAM and BAM files (6, 7).

16.3 Citations

1. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SGE: The IMGT/HLA Database. *Nucleic Acids Research* (2011), **39** Suppl 1:D1171-1176.
2. Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE: IMGT/HLA - a sequence database for the human major histocompatibility complex. *Tissue Antigens* (2000), **55**:280-287.
3. Gragert L, Madbouly A, Freeman J, Maier M: Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human Immunology* (2013), **74**(10): 1313-1320. <http://dx.doi.org/10.1016/j.humimm.2013.06.025>.
4. https://en.wikiversity.org/wiki/Genetics/Human_Leukocyte_Antigen/Linkage_Disequilibrium/DR-DQ_Blocks - Data exported: 2016.04.21
5. Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, Silva AL, Silva AL, Ghattaoraya GS, Alfievic A, Jones AR and Middleton D: Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acid Research* (2015), **28**:D784-788.
6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup: The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* (2009), **25**:2078-2079.
7. Li H: A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* (2011), **27**:2987-2993.