TypeSeq HPV User Guide

[This document contains the analysis sections from the full user guide]

Stage 7 – Custom TypeSeq Analysis

Three files are required as inputs to the custom TypeSeq analysis: a sample manifest defining the sample IDs and their corresponding dual barcodes, a list of control definitions and a file describing custom groupings which can be used to generate custom outputs. Templates and examples may be downloaded from the docs folder at https://github.com/cgrlab/TypeSeq_HPV. **All files must be in .csv format.**

TypeSeq analysis file list and overview:

1. **Sample manifest**: Describes barcodes 1 and 2 combinations used for each sample; contains additional metadata columns for each sample *(refer to the template for which columns are required; filename: SAMPLE_MANIFEST_TypeSeq_Template)*
2. **Control definitions:** Describes the expected positive/negative results per type for each assay control using the template provided in the Sequencing Analysis Files archive (*filename: CONTROLS_TypeSeq_Control_Definitions*)
3. **Custom Groupings:** Defines customized Type masking and/or Type groupings for reporting format – grouped results may be configured, or types masked from the output if desired; all reporting formats may be described in this table regardless of whether they will be utilized in every run *(refer to the instructions below on how to trigger this feature; filename: GROUPINGS_TypeSeq_Definitions)*
4. **Barcode sequences *[Illumina analysis only]:*** List of sequences corresponding to all possible barcodes (may contain additional barcodes to those used in each batch); see Stage 7b – Illumina Custom TypeSeq Analysis for further details

Instructions on preparing the Analysis Files:

Sample manifest:

Prepare a sample manifest using the template provided in the Ion Sequencing Analysis Files archive (*filename: SAMPLE_MANIFEST_TypeSeq_Template; archive also contains an example completed manifest*); first, delete row 2 containing information about which columns are required from the template, then add sample information; copy and paste barcode 1 and 2 combinations (columns “BC1”, “BC2”) into the manifest for each barcode plate used from the lists in the TypeSeq_Ion_Analysis_Files spreadsheet, matching with the corresponding Owner_Sample_ID. Custom metadata columns may also be added by inserting new columns according to the desired layout in the output files. Unwanted columns in the template file may be deleted (if not marked as being required).

Control definitions:

Prepare a CSV table defining the expected results for each assay control by editing the template provided in the Sequencing Analysis Files archive (*filename: CONTROLS_TypeSeq_Control_Definitions*). Add or remove rows as needed.

All assay controls may be included in this table regardless of whether they will be included in every run; the “Control_Code” strings from the Control definitions file will be searched against the “Owner_Sample_ID” entries in the sample manifest to find matching strings of characters, identifying which samples within the manifest are

controls versus specimens. The Owner_Sample_ID names may have extra characters before or after the Control_Code, if the Control_Code string is not interrupted (e.g. Control_Code = “NTC” will match Owner_Sample_IDs such as “PCR-NTC”, “A1_well_NTC”, “NTC_rep2”; Control_Code = “SiHa-HeLa” will not match an Owner_Sample_ID of “SiHa-and-HeLa”). Differences in capitalization are tolerated.

Use the “Control_type” column to define each assay control as either positive or negative for appropriate grouping in the summary PDF report (use “pos” or “neg”).

Custom Groupings:

Download the GROUPINGS_TypeSeq_Definitions.csv file from Github. Customize masking or grouping of HPV types to suit as follows:

Mask = removes a type from the output results; this type will be absent from the samples_only_matrix, but will still be present in the full_pn_matrix (to facilitate reporting of assay control results, and to ensure complete results are available at a later date if required, without the need for re-analysis)

Masking example: In the GROUPINGS_TypeSeq_Definitions.csv file, the example panel name “Carcinogenic_only” has low risk types masked, by entering “mask” in columns with types that should not be displayed in the samples_only_matrix; all types that are blank for that row will have results displayed

Group = all types noted for custom grouping will be condensed into a single group result of positive of negative; for example, if a grouped HPV16 and 18 result is defined, the grouped result will be positive if either or both HPV16 or 18 is positive, but will be negative only if both types are negative; grouped types will not be reported individually in the samples_only_matrix

Grouping example: In the GROUPINGS_TypeSeq_Definitions.csv file, the example panel “Carcinogenic_Grouped” has an entry in the column “Group_Name”, of “Onco” (though any group name may be defined by the user); this creates a new group that will be called “Onco” in the samples_only_matrix file, which will give a positive result if 1 or more types within the group are positive, and a negative result if all types within the group are negative

Masking and grouping: a combination of both masking and grouping may be implemented (an example of this in the groupings file on Github is “Carcinogenic_only_Grouped”), however a single type cannot be both masked and grouped

Use of this feature will be triggered by matching the “Panel” column entry in the sample manifest file with the “Sample_Sheet_Panel_Name” column entry *(user defined; please only include letters, numbers, - or _ symbols, with no spaces; all characters must match exactly between the two files)* in the *GROUPINGS_TypeSeq_Definitions.csv* file; default is “All”, which has no grouping or masking of types

VARIATION NOTES: *Follow either Stage 7a or 7b for Ion or Illumina analysis steps*

Stage 7a – Ion Custom TypeSeq Plugin Analysis

Before first use of the plugin, refer to https://github.com/cgrlab/TypeSeq_HPV for installation instructions. Follow the instructions in the Appendix (**Error! Reference source not found.**) to import the custom reference and barcode files to Torrent Suite.

Running the TypeSeq Plugin:

1. After the sequencing run analysis has completed, click the "Plugins -> Select the plugins to run" buttons from within the sequencing run results page
2. Click on the TypeSeq-HPV plugin; this will open a plugin setup page
3. Select "choose file" for the Sample Sheet line and navigate to the sample manifest in .csv format containing sample and barcode information; select the file and click "open"
4. Repeat for Control Definitions and Report Grouping Definitions to select the matching files
5. When all three files have been selected, click "submit"
6. Plugin should complete within ~30 – 60 minutes depending on the chip size and number of reads
7. After the plugin completes, a PDF report and zipped file hyperlink will appear; click on each to download and access results files

Stage 7b – Illumina Custom TypeSeq Analysis

Reads shorter than 32 bp are automatically removed during FASTQ generation on the MiSeq, so the first step in the workflow is to regenerate the FASTQ files with custom parameters to recover the short R2 barcode reads. Alternatively, heterogeneity spacers could be designed and added to the R2 primers as well, and 2 x 75 bp sequencing be performed if desired.

Expansion of Illumina S3 barcodes:

- New barcodes can be created for synthesis of a larger group of S3 barcoded primers (see **Error! Reference source not found.** for design instructions)
- To add new barcodes to the barcode list for analysis, copy the sequences into the TypeSeq_Illumina_Barcode_List file after converting the barcode sequence to the reverse complement of the 5' to 3' sequence used for primer synthesis, for both the forward and reverse barcode sequence; follow the barcode naming conventions used for the existing barcodes

Running the TypeSeq Illumina analysis workflow:

1. Re-generate the R1 and R2 fastq files by running bcl2fastq with custom parameters as follows:

```
bcl2fastq --runfolder-dir [location of run directory (local or networked drive) ] \
--output-dir [ any directory] \
--with-failed-reads \
--minimum-trimmed-read-length 11 \
```

```
        --mask-short-adapter-reads 11
```

   a. *This should generate two new fastq files, one for R1 and one for R2*

2. Refer to https://github.com/cgrlab/TypeSeqHPV for the most up to date instructions on how to run the Illumina workflow

TypeSeq Analysis Output Files

The custom TypeSeq analysis workflow generates several types of outputs. Filenames are customized for each analysis, incorporating the "Assay_Batch_Code" and/or "Project" names specified in the sample manifest. If a run contains more than one Project code, one file per unique project code will be generated for the samples_only_matrix output.

The possible results for the "human_control" column are either "pass" or "failed_to_amplify". "Failed_to_amplify" is used rather than "fail", for clearer differentiation for negative controls results, where "failed_to_amplify" is the desired result and is not considered a fail. A "failed_to_amplify" result occurs when the number of human control and/or HPV reads are below an optimized minimum threshold.

Output file list:

1. **"[Assay_Batch_Code]_full_pn_matrix.csv"** - positive/negative HPV calls for all samples and assay controls in a run in matrix/table format; also includes a "pass" or "failed_to_amplify" result for the human control gene
2. **"[Assay_Batch_Code]_control_results.csv"** – pass/fail results for all samples identified as assay controls (matching the control definitions file); for rapid QC assessment of assay control performance; a pass result for each requires a 100% match to the content in the control definitions file for all HPV types and the human control result (i.e. all the expected negatives must be negative in addition to matching positives)
3. **"[Project]_[Assay_Batch_Code]_samples_only_matrix.csv"** – contains positive/negative calls for all samples not identified as assay controls
4. **"[Assay_Batch_Code]_failed_samples_matrix.csv"** – contains positive/negative calls for non-control samples failing human control detection; for easy identification of samples needing repeat testing
5. **"[Assay_Batch_Code]_qc_report.pdf"** – A report containing run metadata from the sequencer, assay control performance summary and various plots for QC purposes

A CSV file containing the raw read numbers before positive/negative determination (but after dual de-multiplexing and quality filtering) will be generated within the plugin/workflow analysis folder. If this file is viewed it is important to be aware that the read numbers do not correspond in any way to viral load.