## Description of Xpress-seq data deliveries

This document describes the content and format of Xpress-seq preprocessing data output and delivery.

### Overview

In short, the data the customer will receive in return after completed library service consist of the following:

- Raw Data
- Preprocessing Pipeline Output
- Count Tables
- Report File
- Metadata and other files

### Raw Data

Raw sequencing data is provided in gzipped fastq format. Depending on the ordered read depth and number of Xpress-seq plates, the raw data may consist of partial, full or multiple flow cells.

Xpress-seq raw data consist of two read-types: (1) UMI-containing reads derived from the 5′ end of transcripts. The UMI position is bases 1-10 of read 1 in UMI-containing reads (Xpress-seq v2 chemistry). (2) internal reads spanning the full-length of transcripts (no UMI).

We use dual 10 bp barcodes to denote cell identity.

### Preprocessing Pipeline Output

We utilize a closed-source pipeline adapted from zUMIs (Parekh et al., 2018). The pipeline performs following steps:

1. Read filtering (based on quality)
2. Barcode error correction
3. Determination of UMI-containing reads and extraction of UMI sequences.

The resulting output is filtered, parsed, as unmapped reads into a .bam format (Xpress.filtered.tagged.unmapped.bam).

Next, the pipeline proceeds with:

4. Mapping to the reference genome (using STAR; and including any transgenes if applicable and provided to us)
5. Assignment of mapped reads to gene features (introns & exons)
6. Error correction of UMI sequences.

The resulting output is a coordinate sorted .bam file with all reads including gene assignment, UMI and barcode tags:
(Xpress.filtered.Aligned.GeneTagged.UBcorrected.sorted.bam).

The utilized tags within the file are specified as follows:

- BC --- Barcode sequence
- BX --- Raw barcode before correction
- QB --- Barcode sequence quality scores
- UB --- UMI sequence
- UX --- Raw UMI sequence before correction
- QU --- UMI sequence quality scores
- ES --- featureCounts status for exon assignment
- EN --- Number of overlapping exons
- GE --- GeneID for exon assignment
- IS --- featureCounts status for intron assignment
- IN --- Number of overlapping introns
- GI --- GeneID for intron assignment

Lastly, the pipeline calculates:

7. Count tables (dgecounts.rds object in zUMIs_output/expression/)
8. Summary statistics (zUMIs_output/stats/ folder)

Certain preprocssing summary statistics that may be used for QC during downstream analysis:

1. Mapping statistics per cell.
   (zUMIs_output/stats/Xpress.readspercell.txt)
2. Overview of amount of UMIreads and internal reads per cell
   (zUMIs_output/Xpress_barcodes_binned.txtBCUmistats.txt)

## Count Tables

Certain gene expression count tables are also provided in plain text format for convenience. We provide four quantifications:

1. Unique UMI molecule counts
    a. Exon-mapped reads
    b. Intron+Exon-mapped reads
2. All read counts (UMI + internal, no deduplication)
    a. Exon-mapped reads
    b. Intron+Exon-mapped reads

All count matrices have cells in columns and genes in rows of the table.

## Report File

We provide an automatically generated report file (ReportXG.html) that can be viewed in a browser. The report contains summary statistics and plots that can facilitate a first look into the generated sequencing data. Note that the results shown are automatically generated and data needs to be inspected and filtered by the customers for a full analysis.

## Metadata and other files

In addition to the previous files and outputs, data submissions contain the following:

- Metadata file (barcode_annotation.txt): This file lists the mapping of cell barcodes to their corresponding submitted 384-well plate (denoted by plate barcode) and well position (eg. A1, B1, etc).
- Barcode combination file (barcode_shares.txt): Each sequenced cell obtains two unique cell barcodes in Xpress-seq v2 chemistry. In this file we list the combination of matching cell barcodes, with each line corresponding to one single cell.
- Pipeline configuration file in YAML format with settings used to process the raw data.

For questions, please contact us at service@xpress-genomics.com.