**README.v4.1 for Human Brain GeoMx® Spatial Dataset**

Table of Contents

Sample Data…………………………………………………………………………………..….. 2

GeoMx NGS Workflow………………………………………...……………………………....... 3

Dataset Structure and File Names………………………………………………...…………… 4

Experiment and Data Processing Information…………………….…………...……………… 5

Experiment………………………………………...…………………………….............. 5

Data Processing………………………………………...……………………………...... 6

File Content Description………………………………………….…………….………………... 7

Imaging Files……………………………………………………...……………………… 7

OME-TIFF………………………………………………………………….…….. 7

OME-XML………………………………………………………………………… 7

Sequencing Files……………………………………………...…………………………. 10

FASTQ Files…………………………………………………………………….. 10

Workflow Files ……………………………………………...……………………………. 11

PKC File…………………………………………………………………..……… 11

ROI Report……………………………………………………………………….. 12

NGS Readout……………………………………………………………………. 13

DCC Files………………………………………………………………………... 15

Count Files………………………………………………………...……………………... 16

Description of Exported Count Files..……………………………………….... 16

Segment Properties……………………………………………………. 17

Target Properties………………………………………………………. 21

Target Count Matrix……………………………………………………. 21

Biological Probe Properties….………………………………………… 22

Dataset Summary……………………………………………………… 24

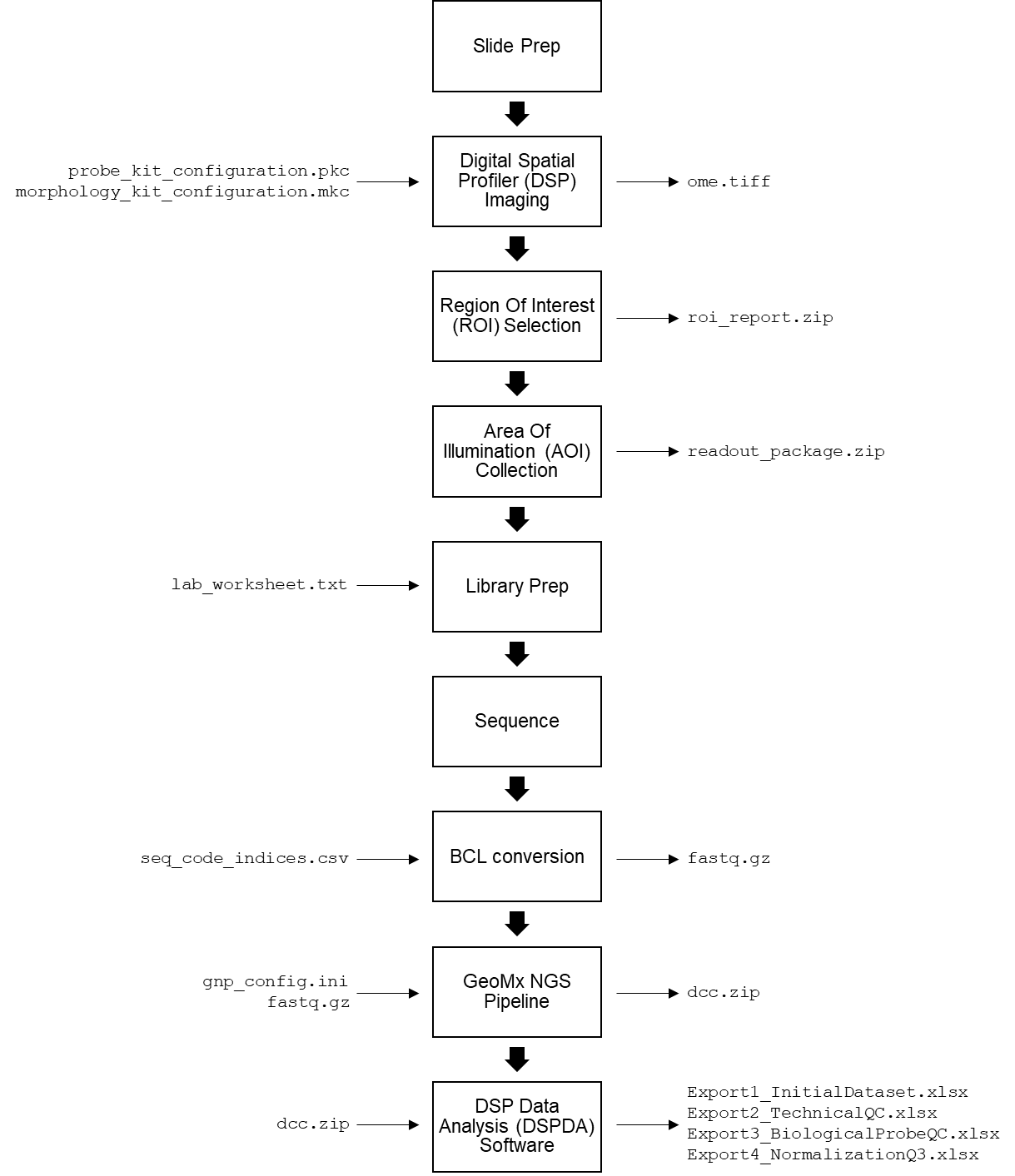
Dataset Comments…………………………………………………………...………………….. 27

GeoMx Digital Spatial Profiler is for Research Use Only. Not for use in diagnostic procedures.

Sample Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | hu\_brain\_001 | hu\_brain\_002 | hu\_brain\_003 | hu\_brain\_004a | hu\_brain\_004b |
| Sample ID | 81074A(5) | 81077A(2) | 81073A(1) | 532229A(3) | 81084A(3) |
| Sex | M | M | M | M | M |
| Age | 90 | 70 | 72 | 75 | 75 |
| Ethnicity | Caucasian | Caucasian | Caucasian | Caucasian | Caucasian |
| Sample Type | FFPE | FFPE | FFPE | FFPE | FFPE |
| Diagnosis | normal tissue | normal tissue | normal tissue | normal tissue | normal tissue |
| Date of Autopsy | 06/19/2017 | 07/05/2017 | 04/21/2017 | 09/04/2020 | 09/04/2020 |
| PMI Hours | 3 | 3 | 4.5 | 4 | 4 |
| BMI | 24.8 | 24.9 | 28.6 | 22.9 | 22.9 |
| Cause of Death | heart failure | cerebral adema | stroke, PE | acute heart failure | acute heart failure |
| Comorbidities / Pathological Findings | hypertension, CAD, atherosclerotic nephrosclerosis, diabetes mellitus type 2, chronic bronchitis, pneumosclerosis, emphysema, pyelonephritis | gastric ulcer perforation, peritonitis, acute pancreatitis | atherosclerosis, cardiosclerosis, chronic prostatitis, chronic tonsilitis, chronic pyelonephritis, pneumosclerosis, chronic hepatitis | hypertension, myocardial hypertrophy, atherosclerosis of the aorta, CAD, cerebral atherosclerosis, atherosclerotic nephrosclerosis, benign prostatic hyperplasia | hypertension, myocardial hypertrophy, atherosclerosis of the aorta, CAD, cerebral atherosclerosis, atherosclerotic nephrosclerosis, benign prostatic hyperplasia |

GeoMx NGS Workflow



Dataset Structure and File Names

> README\_hu\_brain.docx

> image\_files

* hu\_brain\_image\_files.tar.gz

> sequencing\_files

* hu\_brain\_sequencing\_files.tar.gz

> workflow\_and\_count\_files

> workflow

> pkc

* Hs\_R\_NGS\_WTA\_v1.0.pkc\_.zip

> roi\_report

* hu\_brain\_001.zip
* hu\_brain\_002.zip
* hu\_brain\_003.zip
* hu\_brain\_004a.zip
* hu\_brain\_004b.zip

> readout\_package

* 01SEPT2021\_HWTA\_20210910T1632
* 21June2021\_huWTA\_20210622T1805
* 30Apr2021\_HWTA\_20210506T1429

> dcc

* run1\_DCC.zip
* run2\_DCC.zip
* run4\_DCC\_combined.zip

> count

* Export1\_InitialDataset.xlsx
* Export2\_TechnicalQC.xlsx
* Export3\_BiologicalProbeQC.xlsx
* Export4\_NormalizationQ3.xlsx

> count\_results

* hu\_brain\_count\_results.tar.gz
  + Export3\_BiologicalProbeQC.xlsx
  + Export4\_NormalizationQ3.xlsx
  + README\_abridged.docx

> hu\_brain\_md5sums

* hu\_brain\_md5sums.txt

Experiment and Data Processing Information

> Experiment

* Number of individuals = 4
* Sample type = 5 µm FFPE tissue section
* Number of biospecimens = 5
* Name of biospecimens
* hu\_brain\_001
* hu\_brain\_002
* hu\_brain\_003
* hu\_brain\_004a
* hu\_brain\_004b
* Type of GeoMx RNA Assay = Human Whole Transcriptome Atlas (WTA)
* Number of Morphology Markers = 4

|  |  |  |  |
| --- | --- | --- | --- |
| Channel Name | Fluorescent Dye | Marker | Cell / Structure Marked |
| FITC/525 nm | Alexa 488 | GFAP | Astrocytes |
| Cy3/568 nm | SYTO 83 | DNA | Nuclei |
| Texas Red/615 nm | Alexa 594 | Iba-1 | Microglia |
| Cy5/666 nm | Alexa 647 | NeuN | Neurons |

* Number of scans = 5
* Number of Regions of interest (ROIs) = 144
* Number of Areas of illumination (AOIs) = 252
* Number of OME-TIFF images = 5
  + Version of DSP Control Software for exporting OME-TIFF = v2.4
  + Software used to extract OME-XML file = Python 3.8.8. with ome-types 0.2.9 library. For compatibility with the OME-TIFF, the xmlschema library was downgraded to v1.4.1. <https://pypi.org/project/ome-types/>
* Number of ROI Reports = 5
  + Version of DSP Control Software for exporting reports = v2.4
* Number of NGS Readout Packages = 3
  + Version of DSP Control Software for exporting packages = v2.2, v2.3
* Number of indexed libraries (i.e. NTCs + AOIs) = 257
  + Library Prep Kit = Nanostring GeoMx Seq Code
* Sequencing platform = Illumina NovaSeq 6000
  + Sequencing kit
    - S4 flow cell v1.5, 35 cycles
  + Read type = paired-end
  + Read lengths
    - Read 1 = 27 cycles
    - Index 1 = 8 cycles
    - Index 2 = 8 cycles
    - Read 2 = 27 cycles
  + Total raw reads (i.e. read pairs) = 7.27 B

> Data Processing

* Number of FASTQ files (i.e. R1 and R2 for all lanes) = 2072
  + BCL conversion software = bcl2fastq v2.20
* Number of Digital Count Conversion (DCC) files = 257
  + Number of GeoMx NGS Pipeline runs = 3
  + Version of pipeline = v2.0.0 (run1), v2.3.4 (run2 and run4)
* Version of DSPDA Software used for QC processing of count files = v2.4
  + Version for exporting count files = v2.4

File Content Description

> Imaging Files

**OME-TIFF** – Open Microscopy Environment file type containing both pyramidal TIFF image(s) as well as an extensive metadata header.

<https://docs.openmicroscopy.org/ome-model/5.6.3/ome-tiff/>

**OME-XML** – The XML header of the OME-TIFF as specified by the Open Microscopy Environment, specifying extensive metadata of imaging data.  
<https://docs.openmicroscopy.org/ome-model/5.6.3/ome-xml/>

Below is a list of XML tags for given metadata element as reference:

* <OME> tag: Creator information, including Company name, Software name, and Software version.

e.g.

<OME xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openmicroscopy.org/Schemas/OME/2016-06 http://www.openmicroscopy.org/Schemas/OME/2016-06/ome.xsd" Creator="Nanostring GeoMx 2.3.0.179" xmlns="http://www.openmicroscopy.org/Schemas/OME/2016-06">

* <Image> tag: Scan Name

e.g.

<Image ID="Image:0" Name="AC\_BVT\_032921\_01">

* Reagent within the <Screen> tag: Probe Kit(s)

e.g.

<Reagent ID="Reagent:0" ReagentIdentifier="(v1.0) Human Immune Cell Profiling Protein Core" />

* Custom annotation attached to the <Image> tag: Morphology Kit

e.g.

<CommentAnnotation ID="Annotation:5">

<Description>MorphologyKit</Description>

<Value>Human Melanoma TME, P</Value>

</CommentAnnotation>

* <Channel> tag: Channel Name, Fluorophore, Associated pseudocolor

e.g.

<Channel ID="Channel:0" Name="FITC/525nm" Color="65279" Fluor="SYTO 13" SamplesPerPixel="1">

* <Plane> tag: Exposure Time

e.g.

<Plane TheC="0" TheT="0" TheZ="0" ExposureTime="50" ExposureTimeUnit="µs" />

* Custom annotation attached to the <Channel> tag: Biological Class and Biological Target

e.g.

<XMLAnnotation ID="Annotation:1">

<Value>

<ChannelInfo>

<Name>Blue</Name>

<Dye>SYTO 13</Dye>

<DyeDisplayName>FITC</DyeDisplayName>

<DyeWavelength>525nm</DyeWavelength>

<BiologicalClass>Unknown</BiologicalClass>

<BiologicalTarget>DNA</BiologicalTarget>

<IsFocus>true</IsFocus>

</ChannelInfo>

</Value>

</XMLAnnotation>

* <ROI> tag: ROI Name, ROI Type, ROI shape and location description

e.g.

<Label ID="Shape:0" Text="001" FontStyle="Normal" FontFamily="sans-serif" FontSize="64" FontSizeUnit="pt" StrokeColor="-1" X="19087" Y="37791" />

<Ellipse ID="Shape:1" X="19145.3477" Y="38382.7422" RadiusX="495.705257338519" RadiusY="495.705257338519" />

* Mask attached to the <ROI> tag: Segment ID (identifier), Segments Name, Segment Location, Segment Color, Channel Thresholds

e.g.

<Mask ID="Shape:2" FillColor="-858465408" X="18649" Y="37887" Width="991" Height="991" Text="Segment 2">

* Custom annotation attached to the <Image > tag: Segment Definitions

e.g.

<XMLAnnotation ID="Annotation:8">

<Value>

<SegmentDefinitions>

<Segment>

<Name>Segment 1</Name>

<CollectionOrder>1</CollectionOrder>

<DisplayColor>#ccd4db</DisplayColor>

<BlueSelection>1</BlueSelection>

<GreenSelection>3</GreenSelection>

<YellowSelection>3</YellowSelection>

<RedSelection>3</RedSelection>

<Erode>1</Erode>

<Dilate>2</Dilate>

<HoleSize>160</HoleSize>

<ParticleSize>50</ParticleSize>

</Segment>

</SegmentDefinitions>

</Value>

</XMLAnnotation>

* <Instrument> tag: Microscope (Manufacturer and Model)

e.g.

<Microscope Manufacturer="Nanostring" Model="GeoMx" />

* Annotation within the <StructureAnnotation> tag: Instrument ID (Name)

e.g.

<CommentAnnotation ID="Annotation:0">

<Description>InstrumentName</Description>

<Value>GEOMX-B0008</Value>

</CommentAnnotation>

* Custom annotation within the <Image> tag: Slide Name

e.g.

<CommentAnnotation ID="Annotation:4">

<Description>Slide</Description>

<Value>AC\_052620\_Slide1A</Value>

</CommentAnnotation>

* <Pixel> tag: Pixel Sizes

e.g.

<Pixels ID="Pixels:0" BigEndian="false" Type="uint16" SignificantBits="16" Interleaved="false" DimensionOrder="XYCZT" PhysicalSizeX="0.398422241" PhysicalSizeY="0.398994535" SizeX="32768" SizeY="49152" SizeC="3" SizeZ="1" SizeT="1" PhysicalSizeXUnit="µm" PhysicalSizeYUnit="µm">

> Sequencing Files

**FASTQ** **Files** – Standard sequencing file type generated by Illumina sequencers, containing base calls, quality scores, and run metadata.

<https://help.basespace.illumina.com/articles/descriptive/fastq-files/>

e.g.

FGGGGGGGGGGGGGGGGGGGGGCCCCC

@M04116:184:000000000-JNFCN:1:1101:20182:5009 1:N:0:AATCCGGT+TTAAGGCA

CCGATCTCGTATGCCGTCTTCTGCTTG

+

GGFGGGGGGGGGGGGGGGGGGGCCCCC

@M04116:184:000000000-JNFCN:1:1101:4883:5179 1:N:0:AATCCGGT+TTAAGGCA

CGAGATACCGGATTGTGACTGGAGTTC

+

GGGFFGGGGGGGGGGGGGGGGGCCCCC

@M04116:184:000000000-JNFCN:1:1101:17502:5452 1:N:0:AATCCGGT+TTAAGGCA

CGAGATACCGGATTGTGACTGGAGTTC

+

GGGGGGGGGGGGGGGGGGGGGGCCCCC

@M04116:184:000000000-JNFCN:1:1101:26775:5467 1:N:0:AATCCGGT+TTAAGGCA

TCCGATCTAATGATACGGCGACCACCG

+

GGGGGGGGGGGGFCGFBGFEFCCCCCC

@M04116:184:000000000-JNFCN:1:1101:16378:5757 1:N:0:AATCCGGT+TTAAGGCA

CGAGATACCGGATTGTGACTGGAGTTC

+

GGGGFGGGGGGGGGGGGGGGGGCCCCC

@M04116:184:000000000-JNFCN:1:1101:10790:6375 1:N:0:AATCCGGT+TTAAGGCA

CGATCTCGTATGCCGTCTTTCTGCTTG

+

> Workflow Files

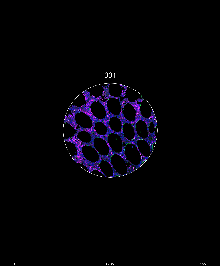
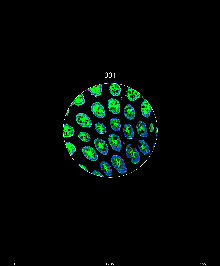
**Probe Kit Configuration (PKC) File** – File input during the GeoMx DSP run set up to specify the RNA or protein panel metadata associated with the panel used during slide preparation. PKC files are coded in JSON format. <https://www.json.org/json-en.html>

|  |  |  |
| --- | --- | --- |
| Name | Description | Example |
| AnalyteType | Type of biological target, either RNA or Protein | RNA |
| DisplayName | Name given at both the target and probe level; target level name may not be unique whereas probe level name is unique; in the case of multiple gene targets only one gene name is given | CAMK1 (target) and CAMK1\_01 (probe) |
| CodeClass | Class of target used for analysis; value is either Control, Negative, or Endogenous | Endogenous |
| RTS\_Seq | 12 bp barcode sequence that is sequenced to identify target; RTS\_Seq = Readout Tag Sequence | GTTCGCCTCAGC |
| GenomeBuild | Genome build corresponding to the given genome coordinates | GRh38.p13 |
| ProbeID | Nanostring unique identification number for each probe | 42048 |
| SystematicName | Name of gene(s) with >=95% identity over the length of probe; human gene names follow HUGO gene nomenclature | CAMK1 |
| Accession | List of NCBI RefSeq accession numbers with >=95% identity over the length of probe | XM\_017007354.1, NM\_003656.5, XM\_005265516.2 |
| GenomeCoordinates | Genome coordinates corresponding to the target sequence | chr3:9761715-9762920 |
| TargetSequence | First 35 bp sequence of the target sequence that hybridizes with the probe; target sequence may be truncated as each probe length is between 35-50 bp | TCTCAAGCCAGAGAATCTGCTGTACTACAGCCTGG |
| RTS\_ID | Nanostring unique name associated with the 12 bp target barcode sequence; RTS\_ID = Readout Tag Sequence Identification | RTS0028011 |
| GeneID | NCBI gene ID for each gene listed in the SystematicName field | 8536 |

**ROI Report** –Collection of ROI selections from a scanned slide. The ROI Report is exported from the DSP Control Center.

* **PNG Files** -Collection ofPNG images from a scanned slide. Image shows the FOV for an ROI and include segmented images if applicable.

e.g. three FOV images for one ROI (PanCK-, PanCK+, All Segments)

****  

* **HTML** **File** – ROI selection summary from one scan displayed in a webpage format, includes metadata (slide name, scan name, channels, and morphology markers), full scanned image, and each FOV image associated with an ROI.

**Readout Package Folder** – After finalizing the readout group(s) on the DSP Control Center, this zipped folder is downloaded. The NGS Readout Package v2.3 contains three file types needed for the NGS readout workflow, see below.

1. **Lab Worksheet File** – Library prep protocol that includes AOI indexing information and AOI annotations.

Experiment Summary – Information on the readout group

e.g.

|  |  |
| --- | --- |
| Readout Group Name | 07June2021\_WTA |
| Date | 6/8/2021 14:36 |
| Readout Mode | NGS |
| Number of Collection Plates | 1 |
| Number Of AOIs | 89 |
| Library Prep Protocol Version |  |

Library Prep Summary – Protocol for which Seq Code primer plate(s) and row(s) to use with each DSP collection plate. Total area of all AOIs in the readout group is in µm2 for calculating sequencing depth.

e.g.

|  |  |  |  |
| --- | --- | --- | --- |
| Library Prep Plate | Collection Plate | Primer Plate | Rows |
| 07June2021\_WTA-H | 1012550000101 | GeoMx Seq Code H | A - H |
| Total Area | 5957573 |  |  |

Annotations – List of each AOI from the DSP collection plate with associated metadata, including area in µm2. A unique *Sample\_ID* is given to each AOI and used to track through Lab Worksheet, Seq Code Indices, and Configuration File. The *Sample\_ID* is in the format of the Platform + DSP collection plate number + Seq Code Primer Plate Letter + Seq Code Primer Plate well (e.g. DSP-1012550000101-H-A01).

e.g.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample\_ID | Slide Name | Scan name | Panel | ROI | Segment | AOI | Area | Tags |
| DSP-1012550000101-H-A01 | No Template Control |  |  |  |  |  |  |  |
| DSP-1012550000101-H-A02 | 025247T1 | 025247T1 | (v1.0) Human NGS Whole Transcriptome Atlas RNA | 001 | PanCK+ | PanCK+-aoi-001 | 88329.44 |  |
| DSP-1012550000101-H-A03 | 025247T1 | 025247T1 | (v1.0) Human NGS Whole Transcriptome Atlas RNA | 001 | PanCK- | PanCK+-aoi-001 | 72605.12 |  |

1. **Seq Code Indices** **File** –AOIindexinginformationfor transferring into the [Data] section of the Illumina sample sheet for FASTQ generation (i.e. BCL conversion). The Illumina sample sheet is the input file for demultiplexing an Illumina sequencing run based on Index 1 (i7) and Index 2 (i5) sequences. Entry of the *Sample\_ID* name into the Illumina sample sheet outputs FASTQ file names containing the *Sample\_ID* corresponding to an AOI. The FASTQ file names are then transferred to the DCC file names during GeoMx NGS Pipeline data processing resulting in the *Sample\_ID* name in the DCC file that then can be tracked back to each AOI within the DSP Control Software. Illumina sample sheet format here: <https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/sequencing-sheet-format-specifications-technical-note-970-2017-004.pdf>

e.g.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample\_ID | I7\_Index\_ID | index | I5\_Index\_ID | index2\_Reverse\_Complement |
| DSP-1012550000101-H-A01 | i7DSP0673 | AGCCATTC | i5DSP0673 | GATATACA |
| DSP-1012550000101-H-A02 | i7DSP0674 | AGAGCGCG | i5DSP0674 | TACGTCAG |

1. **GeoMx NGS Pipeline (GNP) Configuration File** – Input file needed for processing FASTQ files through the GeoMx NGS Pipeline. File contains readout group information a user entered into DSP Control Software when finalizing a readout group, as well as information on pipeline parameters, AOIs and targets. Configuration file is coded in *.ini* format. <https://en.wikipedia.org/wiki/INI_file>

e.g.

[Sequencing]

platform = "NovaSeq 6000"

readPattern = "27x27"

libraryPrep = "Direct\_PCR"

[Processing\_v2]

dcc-metadata = false

save-interim-files = false

quality-trim-score = 20

2color-trimming = True

adapter1 = AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

adapter2 = AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

adapter-trim-match-length = 10

adapter-trim-max-mismatch = 3

barcode-max-mismatch = 1

stitching-max-mismatch = 2

dedup-hd = 1

threads = 4

[AOI\_List]

DSP-1012550000101-H-A01 = 1012550000101,A01,0.00

DSP-1012550000101-H-A02 = 1012550000101,A02,88329.44

DSP-1012550000101-H-A03 = 1012550000101,A03,72605.12

…

[Targets]

RTS0020877 = CGCTTACATGCA

RTS0020878 = CTGCAGTGAACC

RTS0020879 = CTAGCAATATAT

…

**Digital Count Conversion (DCC)** **File** – Primary count file outputted from the GeoMx NGS Pipeline during processing of the FASTQ files. Each DCC file corresponds to the counts from one AOI specified in the configuration file. The pipeline outputs a collection of DCC files in a dcc.zip folder that is uploaded back into the DSP Control Software for further data processing. Each DCC file name contains the *Sample\_ID* name corresponding to the AOI.

e.g.

<Header>

FileVersion,0.02

SoftwareVersion,"GeoMx\_NGS\_Pipeline\_2.3.4"

Date,2021-5-20

</Header>

<Scan\_Attributes>

ID,DSP-1012300100510-H-A01

Plate\_ID,1012300100510

Well,A01

</Scan\_Attributes>

<NGS\_Processing\_Attributes>

SeqSetId,A01356:36:H2FCLDMXY

Raw,49477

Trimmed,44795

Stitched,246

Aligned,1

umiQ30,0.9945

rtsQ30,0.9956

</NGS\_Processing\_Attributes>

<Code\_Summary>

RTS0028413,1

</Code\_Summary>

> Count Files

To generate count files, the DCC files are uploaded back into GeoMx DSP Data Analysis (DSPDA) software. The count data within DCC files are matched with the metadata and any probes that have a count of 0 is automatically given 1. Once the data is imported, the DSPDA initiates the QC workflow. There are two steps in the QC process, Segment QC and Biological Probe QC. After QC, it is common to normalize WTA data using the Q3 method to compare counts between segments. For Q3 method, segments are normalized such that the 75th percentile value (i.e. third quartile) is the same across segments. The normalization factor for each segment is Q3/geomean(Q3 of all segments) and the normalized count is count/normalization factor. More information on the DSPDA workflow can be found in the Nanostring GeoMx NGS Data Analysis Manual. <https://blog.nanostring.com/geomx-online-user-manual/Content/DataAnalysis/GeoMx-NGS_Data_Analysis/NGS.DA.HTML.htm>

Four exported count files are included in this dataset to represent each major data processing step in DSPDA software workflow. Note that the worksheet names within an exported count file may be shared between the four export files, however, the same worksheet may slightly differ depending on what step the dataset was exported in during the analysis workflow.

Below we describe where in the analysis workflow each count file was exported. We then describe the columns for each of the five possible worksheets included in the exported count file. Some columns in a worksheet may be absent or present depending on the data processing step.

* **Export1\_InitialDataset** – DSPDA export of initial count file prior to QC processing in DSPDA Software. This file is an excel workbook that contains two worksheets (SegmentProperties, BioProbeCountmatrix).
* **Export2\_TechnicalQC** - DSPDA export of an intermediate count file after QC of the segments in DSPDA Software as specified in the Data Summary worksheet. This file is an excel workbook that contains three worksheets (SegmentProperties, BioProbeCountMatrix, Data Summary).
* **Export3\_BiologicalProbeQC** - DSPDA export of resulting count file after QC of probes in DSPDA Software as specified in the Data Summary worksheet. This file is an excel workbook that contains five worksheets (SegmentProperties, TargetProperties, TargetCountMatrix, BioProbeProperties, Dataset Summary). TargetCountMatrix worksheet in Export3 represents the QC processed but unnormalized count data matrix typically used as the starting point for downstream tertiary analysis.
* **Export4\_NormalizationQ3** - DSPDA export of Q3 (third quartile) normalized count file after QC processing in DSPDA Software as specified in the Data Summary worksheet. This file is an excel workbook that contains five worksheets (SegmentProperties, TargetProperties, TargetCountMatrix, BioProbeProperties, Dataset Summary). TargetCountMatrix worksheet in Export4 represents the normalized count data matrix typically used in tertiary analysis.

Segment Properties

|  |  |  |
| --- | --- | --- |
| Column Header | Description | Example |
| SlideName | Name of slide as tracked in DSP instrument, user input | hu\_brain\_001 |
| ScanLabel | Name of scan as tracked in DSP instrument, user input; multiple scans may be generated from one slide | 74 |
| ROILabel | Numeric identifier for ROI; all segments within one ROI will have the same ROI number (e.g. PanCK+ and PanCK- segments from same ROI are both labeled ROI "001") | 001 |
| SegmentLabel | Label from segmenting on morphology marker or else will be "Geometric Segment" or “Full ROI” which denotes that the ROI was not segmented, Tag of “Geometric Segment” occurs in GeoMx control software v2.3 and earlier and replaced with the term “Full ROI” in v2.4 | Geometric Segment |
| SegmentDisplayName | Unique name per segment (i.e. AOI) as combination of "ScanLabel | ROILabel | SegmentLabel" | 74 | 001 | Geometric Segment |
| [Morphology Marker Name] | Additional tag columns with each column containing a custom tag added to segments or any autogenerated segment tags. "True" or "False" value is dependent on whether the user selected this tag during the segmentation selection step | FALSE |
| QC Flags | List of QC Flag(s) from Segment QC in DSPDA using the thresholds recorded in the associated Data Summary worksheet, value is blank for initial dataset prior to QC; if after QC, value is blank if segment is not flagged | Low Nuclei Count |
| AOISurfaceArea | Segment surface area, in pixels after conversion of scan in mm | 100501.661 |
| AOINucleiCount | Segment nuclei count derived using the following algorithm in DSPDA v2.4 or earlier: To implement nuclear counting, the nuclear signal is first detected and masked from the nuclear stained images using the minimum cross entropy thresholding method [1]. Background subtraction is used to remove auto-fluorescence signals. Within the initial nuclear mask, a fine level segmentation is performed to extract precise cell boundaries using Bernsen dynamic thresholding of gray-level images [2]. This method sets dynamic threshold values that depend on the local image contrast. This ensures that cells with varying level of intensities can be counted. If there are cell membranes present in the scan / ROIs, they will be detected and used to enhance borders between nuclei regions using the Laplacian of Gaussian (LoG) filter [3] prior to thresholding. In tissues, there are areas of contiguous and overlapping cells. To count individual cells, the overlapping mask areas are separated using watershed [4]. Marker-based watershed algorithm from EmguCV [5] is used. Markers (seeds from where the cells will grow) are obtained by finding local maxima in a topology map where heights represent distances of internal mask pixels to the closest boundaries. The topology map is computed using distance transform [6]. From each marker, watershed will fill/flood the local area / basin until they form separate cells. To reduce noise and spurious splitting, the input mask shapes is smoothed, and all holes formed from the local thresholding method are filled. Finally, small tiny particles from debris are removed and all individual nuclear components extracted from watershed are counted using connected component algorithm [7].  References [1] Li C.H. and Tam P.K.S. "An Iterative Algorithm for Minimum Cross Entropy Thresholding", Pattern Recognition Letters, 18(8): 771-776, 1998. [2] Bernsen J. "Dynamic Thresholding of Grey-Level Images", Proc. of the 8th International Conference on Pattern Recognition, pp. 1251-1255, 1986. [3] The Laplacian of Gaussian: https://en.wikipedia.org/wiki/Blob\_detection#The\_Laplacian\_of\_Gaussian [4] Fernand Meyer, “Color image segmentation.” International Conference on Image Processing and its Applications, 1992, pp. 303–306. 1992. (EmguCV). [5] .NET OpenCV wrapper: http://www.emgu.com/wiki/index.php/Main\_Page. [6] Gunilla Borgefors. “Distance transformations in digital images.” Computer vision, graphics, and image processing, 34(3):344–371, 1986. (EmguCV). [7] Kesheng Wu, Ekow Otoo, and Kenji Suzuki. “Optimizing two-pass connected-component labeling algorithms.” Pattern Analysis and Applications, 12(2):117–135, Jun 2009. (EmguCV). | 157 |
| ROICoordinateX | x-coordinate of the ROI center on the Raw Channel Image exported from DSPDA, x starts on the left with 0 and extends right, in pixels | 14218 |
| ROICoordinateY | y-coordinate of the ROI center on the Raw Channel Image, y starts at the top with 0 and extends down, in pixels | 30294 |
| RawReads | Number of read pairs initially inputted into the GeoMx NGS Pipeline | 10617132 |
| AlignedReads | Number of reads remaining after alignment to the barcode whitelist (i.e. 12 bp RTS\_ID sequence); step after StitchedReads in GeoMx NGS Pipeline | 10142882 |
| DeduplicatedReads | Number of reads remaining after removal of PCR duplicates (i.e. unique UMIs); step after AlignedReads in GeoMx NGS Pipeline | 259343 |
| TrimmedReads | Number of read pairs remaining after quality and adapter trimming; step after initial input into GeoMx NGS Pipeline | 10604787 |
| StitchedReads | Number of reads remaining after stitching overlapping sequences of Read 1 and Read 2; step after TrimmedReads in GeoMx NGS Pipeline | 10537456 |
| SequencingSaturation | Sequencing saturation is related to sequencing depth and dependent on the given library complexity and reads sequenced; Proportion of UMIs in library sequenced more than once, sequencing saturation = (1 - deduplicated reads / aligned reads) | 97.44310345 |
| SequencingSetID | Sequencing run information passed to DSPDA from DCC files as combination of "SequencingInstrumentName:RunID:FlowCellID" | A01356:83:HKKJKDSX2 |
| UMIQ30 | Fraction of total base calls in UMI (14 bases) with Illumina quality score of 30 or above | 0.9986 |
| RTSQ30 | Fraction of total base calls in RTS\_ID (12 bases) with Illumina quality score of 30 or above | 0.9985 |
| GeoMxNgsPipelineVersion | Version of the GeoMx NGS Pipeline used to produce DCC from FASTQ | "GeoMx\_NGS\_  Pipeline\_ 2.0.0" |
| LOT\_Human\_NGS\_  Whole\_Transcriptome\_  Atlas\_RNA | Lot number for each probe kit for each scan, entered by user during scan setup | HWTA12001 |
| [Annotation Tags] | User input annotation tag given during analysis in DSPDA software | Cortex |
| ROIID | Alphanumeric identification per ROI, segments from the same ROI will have the same ROIID | 51297460-d49f-40ad-9452-b8a4d716dbf0 |
| SegmentID | Alphanumeric identification per segment; if ROI was not segmented, then a SegmentID is still given; all SegmentIDs are unique | 1dda4aff-e8b0-458b-b916-46779f62c681 |
| ScanWidth | Width of the largest rendered scan image, in pixels after conversion of scan width from mm; used to calculate the ROI center of an exported image using the following equation (result in pixels with origin (0,0) at the bottom left): Export X coordinate = (ROICoordinateX -ScanOffsetX) \* (Export width/ScanWidth) | 33520.27734 |
| ScanHeight | Height of the largest rendered scan image, in pixels after conversion of scan height from mm; used to calculate the ROI center of an exported image using the following equation (result in pixels with origin (0,0) at the bottom left): Export Y coordinate = Export height - ((ROICoordinateY -ScanOffsetY) \* (Export height/ScanHeight)) | 33393.21094 |
| ScanOffsetX | x offset of the start of the rendered scan image relative to the raw image, in pixels; used to calculate the ROI center of an exported image using the following equation (result in pixels with origin (0,0) at the bottom left): Export X coordinate = (ROICoordinateX -ScanOffsetX) \* (Export width/ScanWidth) | 7816 |
| ScanOffsetY | y offset of the start of the rendered scan image relative to the raw image, in pixels; used to calculate the ROI center of an exported image using the following equation (result in pixels with origin (0,0) at the bottom left): Export Y coordinate = Export height - ((ROICoordinateY -ScanOffsetY) \* (Export height/ScanHeight)) | 7879 |
| LOQ | The limit of quantitation (LOQ) is a user-defined threshold above which a target is considered detected with high confidence. LOQ is equivalent to the geomean of negative probes in a given segment plus a user-defined number of standard deviations [geomean(neg probes)\*geoSD(neg probes)^X, where X is a user-defined value (default = 2)]. The default value of 2 should be used if a slightly higher false positive rate (<5%) is acceptable in exchange for higher sensitivity. Consider changing the value to 2.5 to minimize false positive calls (~1%), although this may decrease sensitivity slightly. The LOQ threshold can be used downstream to filter out targets that are not detected in at least a certain percentage of segments or when making binary detected/not detected calls. | 22.46606149 |
| NormalizationFactor | Segment normalization factor by which each count is divided to get the normalized count | 2.003061353 |
| ExpressionFilteringThreshold (Human NGS Whole Transcriptome Atlas RNA) | Threshold for calling a gene above LOQ. ExpressionFilteringThreshold = LOQ unless a lower threshold was set (i.e. 2), and then it equals that lower threshold | 22.46606149 |
| Group | User added column for histological structure annotation | Cortical layer I |

Target Properties

|  |  |  |
| --- | --- | --- |
| Column Header | Description | Example |
| TargetName | Name given to target that may include the gene name; target level name may not be unique whereas probe level name is unique; in the case of multiple gene targets only one gene name is given; value from "DisplayName" at target level in PKC file | CHGB |
| HUGOSymbol | Comma-delimited list of name of gene(s) with >=95% identity over the length of probe; human gene names follow HUGO gene nomenclature; value from "SystematicName" in PKC file | CHGB |
| TargetGroup | Target group as defined by the DSPDA software during analysis; multiple groups of targets can be user-defined; “All Targets” is a default group containing all targets in an assay | All Targets |
| AnalyteType | Type of biological target, either RNA or Protein; value from "AnalyteType" in PKC file | RNA |
| CodeClass | Class of target used for analysis; value is either Control, Negative, or Endogenous; value from "CodeClass" in PKC file | Endogenous |
| ProbePool | Integer value that identifies a single pool of probes; in the case multiple assays are used together, there would be more than one integer value in this list | 01 |
| MinCount | Minimum count observed for the target from all segments | 1 |
| MaxCount | Maximum count observed for the target from all segments | 3721 |
| MedianCount | Median count measured for the target from all segments | 82.5 |
| GeomeanCount | Geometric mean of all counts for the target from all segments | 81.97986105 |
| NumberOfProbesIncluded | Number of probes included in the target in analysis after QC | 1 |
| NumberOfProbesTotal | Number of probes included in the target prior to QC | 1 |
| GeneID | The field identifies the one NCBI gene ID from the "SystematicName" column recommended for use in pathway analysis | 1114 |

Target Count Matrix

|  |  |  |
| --- | --- | --- |
| Column Header | Description | Example |
| TargetName | Name given to target that may include the gene name; in the case of multiple gene targets only one gene name is given; value from "DisplayName" at target level in PKC file | CHGB |
| [SegmentDisplayName] | List of columns with SegmentDisplayName as unique header (e.g. 74 | 001 | Geometric Segment), each column header is an AOI, values in the column are the counts of the indicated TargetName on each corresponding row, counts may be unnormalized or normalized depending on the step of the DSPDA workflow dataset was exported (see corresponding Dataset Summary tab to determine) | 27 |

Biological Probe Count Matrix

|  |  |  |
| --- | --- | --- |
| Column Header | Description | Example |
| ProbeName | NanoString unique identification number for each probe, value from "ProbeID" in PKC file | 42855 |
| ProbeDisplayName | Unique name given for probe that includes the target name and number; value from "DisplayName" at probe level in PKC file | JAK2\_01 |
| TargetName | Name given to target that may include the gene name; target level name may not be unique whereas probe level name is unique; in the case of multiple gene targets only one gene name is given; value from "DisplayName" at target level in PKC file | JAK2 |
| HUGOSymbol | Comma-delimited list of name of gene(s) with >=95% identity over the length of probe; human gene names follow HUGO gene nomenclature; value from "SystematicName" in PKC file | JAK2 |
| Accessions | List of NCBI RefSeq accession numbers with >=95% identity over the length of probe; value from "Accession" in PKC file | XM\_017026369.1,NM\_001007248.3 |
| GenomeBuild | Genome build corresponding to the given genome coordinates; value from "GenomeBuild" in PKC file | GRCh38.p13 |
| GenomicPosition | Genome coordinates corresponding to the target sequence; value from "GenomeCoordinates" in PKC file | chr9:5055753-5064896 |
| AnalyteType | Type of biological target, either RNA or Protein; value from "AnalyteType" in PKC file | RNA |
| CodeClass | Class of target used for analysis; value is either Control, Negative, or Endogenous; value from "CodeClass" in PKC file | Endogenous |
| ProbePool | Integer value that identifies a single pool of probes; in the case multiple assays are used together, there would be more than one integer value in this list | 01 |
| TargetGroup | Target group as defined by the DSPDA software during analysis; multiple groups of targets can be user-defined; “All Targets” is a default group containing all targets in an assay | All Targets |
| GlobalOutlier | Status if the probe was called a global outlier using one of two criteria: 1) by failing the Grubb's outlier test in a greater proportion of segments than a user-specified value (i.e. >20% of segments) or 2) by having a geometric mean value across segments less than a user-specified fraction of the geometric mean value of all probes with the same target (i.e. geomean probe in all segments / geomean probes within target <= 0.1) | TRUE |
| GlobalOutlierReason | Reason for probe being called global outlier of the two options explained above | Failed Grubbs test and excluded from all AOIs, AOI proportion is: 7.063492E-01 |
| OutlierFrequency | Proportion of segments in which a probe was called an outlier | 0.706349206 |
| GeneID | The field identifies the one NCBI gene ID from the "SystematicName" column recommended for use in pathway analysis | 3717 |
| [SegmentDisplayName] | List of columns with SegmentDisplayName as unique header (e.g. 74 | 001 | Geometric Segment), each column header is an AOI, values in the column are the deduplicated counts of the indicated ProbeDisplayName or TargetName on each corresponding row | 17 |

Dataset Summary

|  |  |  |
| --- | --- | --- |
| Row Header | Description | Example |
| Analysis name | User given analysis name | Human brain organ book 4 sample analysis |
| Analysis created by | User identification | avanschoiack-admin |
| Analysis last modified | Date and time stamp | 10-06-2021 23:08:09 |
| Software version | Version notation | 2.4.0.146 |
| Dataset name | Dataset name used for analysis | Export2\_TechnicalQC |
| Dataset created by | User identification | avanschoiack-admin |
| Dataset last modified | Date and time stamp | 10-06-2021 22:20:39 |
| Segment QC: Dataset name | Dataset name used for analysis | Export2\_TechnicalQC |
| Segment QC: Created by | User identification | avanschoiack-admin |
| Segment QC: Last modified | Date and time stamp | 10-06-2021 22:20:39 |
| Segment QC: Derived from | Dataset used as input to analysis | Export1\_InitialDataset |
| Segment QC: Filter Targets | Status whether to filter on QC flag (yes or no) | No |
| Segment QC: Used raw reads threshold | Flag status activated (yes or no) | Yes |
| Segment QC: Raw read threshold | Flag if less than threshold of raw reads, default = 1000 raw reads | 1000 |
| Segment QC: Used aligned reads threshold | Flag status activated (yes or no) | Yes |
| Segment QC: Aligned reads threshold | Flag if percent aligned reads is less than threshold; percent aligned = (aligned reads / raw reads) \* 100; default = 80% | 80 |
| Segment QC: Used stitched reads threshold | QC Flag activated (yes or no) | Yes |
| Segment QC: Stitched read threshold | Flag if percent stitched reads is less than threshold; percent stitched = (stitched reads / raw reads) \* 100, default = 80% | 80 |
| Segment QC: Used trimmed reads threshold | QC Flag activated (yes or no) | Yes |
| Segment QC: Trimmed read threshold | Flag if percent trimmed reads is less than threshold; percent trimmed = (trimmed reads / raw reads) \* 100, default = 80% | 80 |
| Segment QC: Used sequencing saturation threshold | QC Flag activated (yes or no) | Yes |
| Segment QC: Sequencing saturation threshold | Flag if percent sequencing saturation is less than threshold; percent sequencing saturation = (1 - deduplicated reads / raw reads) \* 100, default = 50% | 50 |
| Segment QC: Used negative probe count geomean threshold | QC Flag activated (yes or no) | Yes |
| Segment QC: Negative probe count geomean threshold | Flag if geomean of negative probes, deduplicated counts is less than threshold, default = 10 counts | 10 |
| Segment QC: Used NTC threshold | QC Flag activated (yes or no) | Yes |
| Segment QC: No Template Control Count | Flag if number of raw reads for NTC (i.e. A01 well) is equal or greater than threshold, default = 1000 raw reads | 1000 |
| Segment QC: Used surface area threshold | QC Flag activated (yes or no) | Yes |
| Segment QC: Surface area Segment QC: QC threshold | Flag if surface area is less than threshold, default = 16000 µm2 | 16000 |
| Segment QC: Used nuclei count threshold | QC Flag activated (yes or no) | Yes |
| Segment QC: Nuclei count QC threshold | Flag if nuclei count is less than threshold, default = 200 nuclei | 200 |
| Probe QC: Dataset name | User given analysis name | Export3\_BiologicalProbeQC |
| Probe QC: Created by | User identification | avanschoiack-admin |
| Probe QC: Last modified | Date and time stamp | 10-06-2021 21:44:13 |
| Probe QC: Derived from | Dataset name used for analysis | Export2\_TechnicalQC |
| Probe QC: Remove Failed Segments before executing Probe QC | Status on whether segments that failed Segment QC were removed prior to ProbeQC (yes or no) | No |
| Probe QC: Used exclude if below probe ratio threshold | Status on whether probe excluded from target count if (geomean probe in all segments) / (geomean probes within target) is less than or equal to threshold value (yes or no) | Yes |
| Probe QC: Ratio of probe geomean across all segments | Threshold value of which (geomean probe in all segments) / (geomean probes within target), default = 0.1 | 0.1 |
| Probe QC: Used exclude if below Grubbs fail percent threshold | Status to exclude probe from target count if fails Grubbs outlier test in greater than or equal to threshold value (yes or no) | Yes |
| Probe QC: Percent AOIs threshold for Grubbs test | Threshold value for Grubbs outlier test, default = 20% of segments | 20 |
| Probe QC: Used exclude local segment outliers | Status on whether probes that fails Grubbs outlier test in a segment were excluded from target count calculation in that segment (yes or no) | No |
| Probe QC: Standard deviation amount for the LOQ | Number of standard deviations to define confidence level of LOQ; LOQ = geomean(neg probes)\*geoSD(neg probes)^X, where X is a user-defined standard deviation value, default = 2 | 2 |
| Expression Filtering: Dataset name | User given analysis name | Probe QC filtered |
| Expression Filtering: Created by | User identification | avanschoiack-admin |
| Expression Filtering: Last modified | Date and time stamp | 10-06-2021 22:16:12 |
| Expression Filtering: Derived from | Dataset name used for analysis | Export3\_BiologicalProbeQC |
| Expression Filtering: Expression filtering mode | Expression threshold used to remove segments or targets with expression at or lower than selection threshold value at or above specified frequency; expression filter mode options are “LOQ”, “User defined value”, and “Higher of LOQ and user define value” | HIGHER\_OF\_LOQ\_AND\_  USER\_DEFINED |
| Expression Filtering: Filtering by | Expression filtering parameter, options are segment or target | TARGET |
| Expression Filtering: Frequency | Frequency to keep targets that are greater or equal to threshold, default is 10% segments above threshold | 5 |
| Expression Filtering: User-defined value | Expression threshold for the expression filtering mode option of “User defined value”, default = 2 counts | 2 |
| Normalization: Dataset name | User given analysis name | Export4\_NormalizationQ3 |
| Normalization: Created by | User identification | avanschoiack-admin |
| Normalization: Last modified | Date and time stamp | 10-06-2021 22:18:51 |
| Normalization: Derived from | Dataset name used for analysis | Export3\_BiologicalProbeQC |
| Normalization: Filter segment | Status of whether segments were filtered prior to normalization (yes or no) | No |
| Normalization: Filter Targets | Status of whether targets were filtered prior to normalization (yes or no) | Yes |
| Normalization: Reference targets | List of targets used to normalize | Cannot display name list, please select 2000 targets or less. |
| Normalization: Method | Method used to normalize | THIRD\_QUARTILE |

Dataset Comments

* To extract the OME-XML header from the OME-TIFFs, we used the Python library ome-types v0.2.9 (<https://pypi.org/project/ome-types/>). For compatibility with DSP v2.4 OME-TIFFs, xmlschema was downgraded from 1.70 to 1.41. The resulting XMLs were missing the PhysicalSizeUnitX and PhysicalSizeUnitY fields when extracted this way, although the values for these fields are always microns (µm).
* For the numbering of ROIs, OME-TIFF is 0-based indexing while DSP file exports are 1-based indexing. For example, ROI 0 in OME-XML file corresponds to ROI 1 in DSP export files.
* Tag annotation of “Neuronal environment” is equivalent to “neuropil.”
* Use the Updated\_SegmentLabel column in the Segment Properties worksheet for final annotation of segments.
* Morphology marker information:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Marker** | **Cell / Structure Marked** | **Channel Name** | **Fluorescently-labeled By** | **Host** | **Clone ID** | **Vendor** | **Catalog Number** | **Concentration** | **Notes** |
| GFAP | Astrocytes | 488 | Vendor | Mouse | GA5 | Thermo | 53-9892-82 | 5.00 ug/mL (RNA)  2.50 ug/mL (Protein) | Add after secondaries |
| Iba1 | Microglia | Unconjugated (594) | Secondary Ab | Mouse | 20A12.1 | Sigma | MABN92 | 1:100 | Secondary a-Ms 594, 1:400, Thermo Fisher, cat.# A-11005 |
| NeuN | Neurons | Unconjugated (647) | Secondary Ab | Rabbit | Polyclonal | Millipore | ABN78 | 1:40 | Secondary a-Rb 647, 1:400, Thermo Fisher, cat.# A32733 |
| DNA | Nuclei | 532 | N/A | N/A | N/A | Thermo | S11364 | 200nM | Syto 83 |

* Human brain sections are through the medial temporal lobe and contain the hippocampus, along with surrounding areas of cerebral cortex and subcortical white matter. The hippocampus is part of the limbic system with important roles in learning and memory. The cerebral cortex is associated with higher order cognitive functions including decision making, thought and emotion and the subcortical white matter enables communication between different regions of the brain. With GeoMx DSP, regions of interest (ROIs) are selected by the user and then molecularly profiled. These ROIs can be either profiled in entirety (i.e., geometric profiling) or subdivided based on the visible morphology markers into compartments and profiled separately (i.e., segmented profiling). The segmented profiling carried out in the human brain sections are NeuN+ (Neuronal nuclei and perinuclear cytoplasm), NeuN- (Neuropil), Iba1+ (Microglia), GFAP (Astrocytes), and Syto83 (Nuclei).
* README v4 updated and released on 5-23-2022. README v3 updated and released on 11-1-2021. README v2 updated and released on 10-18-2021. README v1 released on 10-10-2021.
* README v4.1 updated on 10-17-2022.