Files Manifest

The data in this archive refers to:

Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets
Jiayi Tang, Alex Henderson* and Peter Gardner

Analyst DOI: https://doi.org/10.1039/D0AN02155E

Another Zenodo archive contains processed versions of these data, including in MATLAB file format, for this paper. https://doi.org/10.5281/zenodo.4730312

Entire archive is 177 GB compressed and 249 GB uncompressed.

Care should be taken regarding the orientation of H&E and IR images

BR20832.csv (15 kB)

A comma separated variable file with information on the pathology classification of each tissue microarray core. More information is available from http://www.biomax.us/tissue-arrays/Breast/BR20832

BR20832_H-and-E.tif (33 MB)

A single TIF microscopy image of the H&E stained microarray. Further information on the tissue microarray can be found at http://www.biomax.us/tissue-arrays/Breast/BR20832. Note that the sample analysed is a serial section.

The core positions are designated in Table 1.

Raw data (176.8 GB, uncompressed size 248.3 GB)

The infrared spectroscopic analysis of the tissue microarray was performed in zones since the analysis area was so large. Each region comprised one or more cores. The layout of the regions is given in Table 2. Each region is in recorded in Agilent IR mosaic format and compressed to reduce file size. The file compression format is 7z, a format giving better performance than zip. Software to decompress 7z files is available from http://www.7-zip.org/

These data can be opened in MATLAB using ChiToolbox, https://bitbucket.org/AlexHenderson/chitoolbox/. Alternatively, MATLAB code is available at https://bitbucket.org/AlexHenderson/agilent-file-formats/, or Python code at https://bitbucket.org/AlexHenderson/agilentirformats/.

Table 1: Core numbering system

A16	B16	C16	D16	E16	F16	G16	H16	I16	J16	K16	L16	M16
A15	B15	C15	D15	E15	F15	G15	H15	I15	J15	K15	L15	M15
A14	B14	C14	D14	E14	F14	G14	H14	I14	J14	K14	L14	M14
A13	B13	C13	D13	E13	F13	G13	H13	I13	J13	K13	L13	M13
A12	B12	C12	D12	E12	F12	G12	H12	I12	J12	K12	L12	M12
A11	B11	C11	D11	E11	F11	G11	H11	I11	J11	K11	L11	M11
A10	B10	C10	D10	E10	F10	G10	H10	I10	J10	K10	L10	M10
A9	B9	C9	D9	E9	F9	G9	H9	I9	J9	K9	L9	M9
A8	B8	C8	D8	E8	F8	G8	H8	I8	J8	K8	L8	M8
A7	B7	C7	D7	E7	F7	G7	H7	I7	J7	K7	L7	M7
A6	B6	C6	D6	E6	F6	G6	H6	I6	J6	K6	L6	M6
A5	B5	C5	D5	E5	F5	G5	H5	I5	J5	K5	L5	~~M5~~
A4	B4	C4	D4	E4	F4	G4	H4	I4	J4	K4	L4	M4
A3	B3	C3	D3	E3	F3	G3	H3	I3	J3	K3	L3	M3
A2	B2	C2	D2	E2	F2	G2	H2	I2	J2	K2	L2	M2
A1	B1	C1	D1	E1	F1	G1	H1	I1	J1	K1	L1	M1

The core at position M5 is missing in the tissue microarray analysed.

Table 2: Region numbering system.

<table>
<tr><td colspan="2">1516AC</td><td colspan="2">1516DF</td><td colspan="2">1516GI</td><td colspan="2"></td><td rowspan="3">1114M</td></tr>
<tr><td colspan="2">1314AC</td><td colspan="2">1314DF</td><td colspan="2">1314GI</td><td colspan="2">1314JL</td></tr>
<tr><td colspan="2">1112AC</td><td colspan="2">1112DF</td><td colspan="2">1112GI</td><td colspan="2">1112JL</td></tr>
<tr><td colspan="2">910AC</td><td colspan="2">910DF</td><td colspan="2">910GI</td><td colspan="2">910JL</td><td rowspan="2">610M</td></tr>
<tr><td colspan="2" rowspan="2">78AC</td><td colspan="2" rowspan="2">78DF</td><td colspan="2" rowspan="2">78GI</td><td rowspan="3">68JK</td><td rowspan="4">58L</td></tr>
<tr><td rowspan="3">57M</td></tr>
<tr><td colspan="2" rowspan="2">56AC</td><td colspan="2" rowspan="2">56DF</td><td colspan="2" rowspan="2">56GI</td></tr>
<tr><td rowspan="2">35JK</td></tr>
<tr><td colspan="2">34C</td><td colspan="2">34F</td><td colspan="2">34I</td><td rowspan="2">4L</td><td rowspan="2">4M</td></tr>
<tr><td>2B</td><td colspan="2">2CD</td><td>2EF</td><td>2GH</td><td colspan="2">2IK</td></tr>
</table>

File sizes are as follows:

Filename	**Compressed size (GB)**	**Uncompressed size (GB)**
2B.7z	3.04	4.27
2CD.7z	2.50	3.42
2EF.7z	2.50	3.42
2GH.7z	2.50	3.42
2IK.7z	4.33	5.98
34C.7z	5.37	7.48
34F.7z	5.37	7.48
34I.7z	5.37	7.48
35JK.7z	4.31	5.98
4L.7z	2.75	3.84
4M.7z	2.74	3.84
56AC.7z	5.38	7.48
56DF.7z	5.43	7.57
56GI.7z	5.41	7.48
57M.7z	2.75	3.74
58L.7z	2.68	3.84
610M.7z	3.35	4.72
68JK.7z	4.34	6.02
78AC.7z	5.17	7.48
78DF.7z	5.24	7.48
78GI.7z	5.28	7.56
910AC.7z	5.23	7.48
910DF.7z	5.33	7.48
910GI.7z	5.30	7.48
910JL.7z	5.36	7.57
1112AC.7z	5.36	7.48
1112DF.7z	5.40	7.48
1112GI.7z	5.41	7.48
1112JL.7z	5.30	7.48
1114M.7z	2.69	3.87
1314AC.7z	5.34	7.48
1314DF.7z	5.43	7.58
1314GI.7z	5.40	7.48
1314JL.7z	5.25	7.52
1516AC.7z	5.37	7.48
1516DE.7z	5.42	7.48
1516DF.7z	5.35	7.48
1516GI.7z	5.42	7.56
BR20832-RawData.7z	2.67	4.44
Total	**176.84**	**248.28**