**Supplementary Text 2: Metadata file example**

One of ARepA's main outputs are metadata files in the python pickle (pkl) format; the pickle module implements an algorithm for serializing and de-serializing a python object structure using a printable ASCII representation without restrictions imposed by external standards.

To read the metadata (pkl) format, ARepA provides a script to de-serialize the file which can be found in the source directly: arepa/src/unpickle.py

$ python arepa/src/unpickle.py < metadata.pkl

Example from the prostate cancer case study (Supplementary Table S1): GSE12348

$ python arepa/src/unpickle.py < GSE12348.pkl

Output:

|  |  |
| --- | --- |
|  | ['GSM310067', 'GSM310069', 'GSM310080', 'GSM310081', 'GSM310082', 'GSM310083', 'GSM310084', 'GSM310120', 'GSM310121'] |
| label\_ch1 | ['biotin', 'biotin', 'biotin', 'biotin', 'biotin', 'biotin', 'biotin', 'biotin', 'biotin'] |
| Series\_platform\_taxid | 9606 |
| label\_protocol\_ch1 | ["Affymetrix U133A manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol"] |
| data\_processing | ['MAS5.0', 'MAS5.0', 'MAS5.0', 'MAS5.0', 'MAS5.0', 'MAS5.0', 'MAS5.0', 'MAS5.0', 'MAS5.0'] |
| channels | 1 |
| contact\_address | ['1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53', '1650 Orleans Street, CRB1 Rm 1M53'] |
| Series\_contact\_name | Srinivasan,,Yegnasubramanian |
| Series\_status | Public on Nov 01 2008 |
| type | ['RNA', 'RNA', 'RNA', 'RNA', 'RNA', 'RNA', 'RNA', 'RNA', 'RNA'] |
| gloss | The primary goal of this study was to assess differences in gene expression between prostate cancer cell lines and normal prostate epithelial and stromal cells in primary culture. |
| contact\_name | ['Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian', 'Srinivasan,,Yegnasubramanian'] |
| source\_name\_ch1 | ['1s normal prostate stromal cells, basal conditions', '4ST prostate normal stromal cells', 'C42B prostate cancer cell line', 'CWR22Rv1 prostate cancer cell line', 'DU-145 prostate cancer cell line', 'LAPC4 prostate cancer cells', 'LNCaP prostate cancer cells', 'PC-3 prostate cancer cell line', 'PrEC normal prostate epithelial cells'] |
| supplementary\_file.1 | ['ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310067/GSM310067.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310069/GSM310069.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310080/GSM310080.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310081/GSM310081.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310082/GSM310082.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310083/GSM310083.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310084/GSM310084.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310120/GSM310120.CHP.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310121/GSM310121.CHP.gz'] |
| Series\_contact\_state | MD |
| Series\_contact\_address | 1650 Orleans Street, CRB1 Rm 1M53 |
| title | ['1S\_prostate\_normal\_stromacells-untreated', '4ST\_prostate\_normal\_stromalcells-untreated', 'C42B\_prostate\_cancer\_cell\_line-untreated', 'CWR22Rv1\_prostate\_cancer\_cell\_line-untreated', 'DU145\_prostate\_cancer\_cell\_line-untreated', 'LAPC4\_prostate\_cancer\_cell\_line-untreated', 'LNCaP\_prostate\_cancer\_cell\_line-untreated', 'PC3\_prostate\_cancer\_cell\_line-untreated', 'PREC\_prostate\_normal\_epithelial\_cells-untreated'] |
| scan\_protocol | ["Affymetrix U133A manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol"] |
| contact\_institute | ['Johns Hopkins University', 'Johns Hopkins University', 'Johns Hopkins University', 'Johns Hopkins University', 'Johns Hopkins University', 'Johns Hopkins University', 'Johns Hopkins University', 'Johns Hopkins University', 'Johns Hopkins University'] |
| Series\_contact\_city | Baltimore |
| platform | GPL96 |
| treatment\_protocol\_ch1 | ['No treatment', 'Untreated', 'Untreated', 'Untreated', 'Untreated', 'Untreated', 'Untreated', 'Untreated', 'Untreated'] |
| pmid | 18974140 |
| conditions | 9 |
| Series\_overall\_design | 9 samples were analyzed with 1 replicate each. Samples included 6 prostate cancer cell lines, 1 normal prostate epithelial cell system in primary culture, and 2 normal prostate stromal cell systems in primary culture. |
| submission\_date | ['Aug 05 2008', 'Aug 05 2008', 'Aug 05 2008', 'Aug 05 2008', 'Aug 05 2008', 'Aug 05 2008', 'Aug 05 2008', 'Aug 05 2008', 'Aug 05 2008'] |
| extract\_protocol\_ch1 | ['Qiagen RNeasy kit', 'Qiagen RNeasy kit', 'Qiagen RNeasy kit', 'Qiagen RNeasy kit', 'Qiagen RNeasy kit', 'Qiagen RNeasy kit', 'Qiagen RNeasy kit', 'Qiagen RNeasy kit', 'Qiagen RNeasy kit'] |
| contact\_state | ['MD', 'MD', 'MD', 'MD', 'MD', 'MD', 'MD', 'MD', 'MD'] |
| description | ['None', 'None', 'None', 'None', 'None', 'None', 'None', 'None', 'None'] |
| taxid\_ch1 | ['9606', '9606', '9606', '9606', '9606', '9606', '9606', '9606', '9606'] |
| Series\_contributor | Srinivasan,,Yegnasubramanian  William,G,Nelson |
| biomaterial\_provider\_ch1 | ['John T. Isaacs laboratory', 'John T. Isaacs Laboratory', 'John T. Isaacs Laboratory', 'ATCC', 'ATCC', 'John T. Isaacs laboratory', 'ATCC', 'ATCC', 'Cambrex'] |
| Series\_submission\_date | Aug 05 2008 |
| channel\_count | ['1', '1', '1', '1', '1', '1', '1', '1', '1'] |
| characteristics\_ch1 | ['These are a batch of normal prostate stromal cells propagated in primary culture. These cells are not immortalized.', 'Normal stromal cells from a the prostate gland grown in primary culture. Not transformed or immortalized.', 'This is a prostate cancer cell line that is an androgen independent variant of the LNCaP prostate cancer cell line. It is directly derived from the LNCaP cell line.', 'CWR22Rv1 prostate cancer cell line', 'Androgen receptor negative prostate cancer cell line', 'Androgen sensitive', 'Androgen sensitive prostate cancer cell line', 'PC3 androgen receptor negative prostate cancer cell line', 'Normal prostate epithelial cells grown in primary culture'] |
| geo\_accession | ['GSM310067', 'GSM310069', 'GSM310080', 'GSM310081', 'GSM310082', 'GSM310083', 'GSM310084', 'GSM310120', 'GSM310121'] |
| Series\_contact\_zip/postal\_code | 21231 |
| last\_update\_date | ['Sep 24 2008', 'Sep 24 2008', 'Sep 24 2008', 'Sep 24 2008', 'Sep 24 2008', 'Sep 24 2008', 'Sep 24 2008', 'Sep 24 2008', 'Sep 24 2008'] |
| growth\_protocol\_ch1 | ['Cells were propagated in Iscove\xe2\x80\x99s media (Life Technologies, Inc.) containing 10% fetal bovine serum and 1 nM R1881', "Grown/propagated in Iscove's media with 10% FBS and 1nM R1881", 'Grown/propagated in RPMI 1640 with 10% FBS.', 'Grown/propagated in RPMI 1640 with 10% FBS.', 'Grown/propagated in RPMI 1640 with 10% FBS.', 'Grown/propagated in RPMI 1640 with 10% FBS and 1 nM R1881', 'Grown/propagated in RPMI 1640 with 10% FBS', 'Grown/propagated in RPMI 1640 with 10% FBS', 'Grown/propagated in RPMI 1640 with 10% FBS'] |
| Series\_supplementary\_file | ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/series/GSE12348/GSE12348\_RAW.tar |
| Series\_relation | BioProject: http://www.ncbi.nlm.nih.gov/bioproject/113173 |
| molecule\_ch1 | ['total RNA', 'total RNA', 'total RNA', 'total RNA', 'total RNA', 'total RNA', 'total RNA', 'total RNA', 'total RNA'] |
| Series\_contact\_country | USA |
| data\_row\_count | ['22283', '22283', '22283', '22283', '22283', '22283', '22283', '22283', '22283'] |
| curated | ['', 'title', 'geo\_accession', 'status', 'submission\_date', 'last\_update\_date', 'type', 'channel\_count', 'source\_name\_ch1', 'organism\_ch1', 'characteristics\_ch1', 'biomaterial\_provider\_ch1', 'treatment\_protocol\_ch1', 'growth\_protocol\_ch1', 'molecule\_ch1', 'extract\_protocol\_ch1', 'label\_ch1', 'label\_protocol\_ch1', 'taxid\_ch1', 'hyb\_protocol', 'scan\_protocol', 'description', 'data\_processing', 'platform\_id', 'contact\_name', 'contact\_institute', 'contact\_address', 'contact\_city', 'contact\_state', 'contact\_zip/postal\_code', 'contact\_country', 'supplementary\_file', 'supplementary\_file.1', 'data\_row\_count'] |
| contact\_city | ['Baltimore', 'Baltimore', 'Baltimore', 'Baltimore', 'Baltimore', 'Baltimore', 'Baltimore', 'Baltimore', 'Baltimore'] |
| Series\_geo\_accession | GSE12348 |
| Series\_sample\_id | GSM310067 GSM310069 GSM310080 GSM310081 GSM310082 GSM310083 GSM310084 GSM310120 GSM310121 |
| taxid | 9606 |
| platform\_id | ['GPL96', 'GPL96', 'GPL96', 'GPL96', 'GPL96', 'GPL96', 'GPL96', 'GPL96', 'GPL96'] |
| status | ['Public on Nov 01 2008', 'Public on Nov 01 2008', 'Public on Nov 01 2008', 'Public on Nov 01 2008', 'Public on Nov 01 2008', 'Public on Nov 01 2008', 'Public on Nov 01 2008', 'Public on Nov 01 2008', 'Public on Nov 01 2008'] |
| contact\_country | ['USA', 'USA', 'USA', 'USA', 'USA', 'USA', 'USA', 'USA', 'USA'] |
| mapped | TRUE |
| supplementary\_file | ['ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310067/GSM310067.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310069/GSM310069.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310080/GSM310080.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310081/GSM310081.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310082/GSM310082.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310083/GSM310083.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310084/GSM310084.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310120/GSM310120.CEL.gz', 'ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/samples/GSM310nnn/GSM310121/GSM310121.CEL.gz'] |
| Series\_last\_update\_date | Nov 14 2012 |
| organism\_ch1 | ['Homo sapiens', 'Homo sapiens', 'Homo sapiens', 'Homo sapiens', 'Homo sapiens', 'Homo sapiens', 'Homo sapiens', 'Homo sapiens', 'Homo sapiens'] |
| hyb\_protocol | ["Affymetrix U133A manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol", "Affymetrix hgU133a manufacturer's recommended protocol"] |
| Series\_contact\_institute | Johns Hopkins University |
| contact\_zip/postal\_code | ['21231', '21231', '21231', '21231', '21231', '21231', '21231', '21231', '21231'] |