Reading Bioinformatics and Genomics Files in Power BI

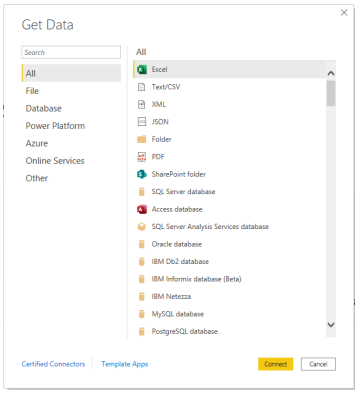
Colby Ford

For many users in the finance, insurance, retail, manufacturing, and even healthcare industries, Power BI has become a staple in their business intelligence plan. From interactive visualizations to advanced data wrangling, Power BI offers a one-stop shop for gaining insights from your data.

However, for any of us that are researchers in the bioinformatics and genomics space, we know that our files can be a bit difficult to work with. From FASTA to BAM, working with files in bioinformatics add a layer of uniqueness that requires some special care.

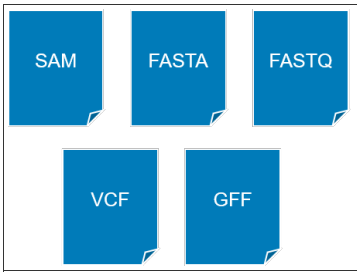
```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328    0   chr1    10145   25  36M *   0   0   AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT    hhhhHcWhhHTghcKA_ONhAAEEBZE?H?CBC?DA    NM:i:1  X1:i:1  MD:Z:33A2
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278    0   chr1    10148   25  36M *   0   0   CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC    hbFhhhXUYhT_ULZdLRTKNIMIKGLJCHFFJQJN    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685    16  chr1    10149   25  36M *   0   0   CCAAACACTAACCCTAACCCTAACCCTAACCTAACC     ><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO    NM:i:1  X1:i:1  MD:Z:29A3A2
SOLEXA-1GA-2_2_FC20EMB:5:35:583:827 0   chr1    10150   25  36M *   0   0   CTAACCCTAAACCTAACCCTAACCCTAACCTAACCA   hhW_X]MXNOHQQWMILHGIFMJGJLCFGGJAKIEH    NM:i:1  X1:i:1  MD:Z:10A24A0
SOLEXA-1GA-2_2_FC20EMB:5:248:130:724    0   chr1    10152   25  36M *   0   0   AACCCTAACCCTAACCCCAACCCTAACCTAACCCTA    hchPhc___cWS[VROb$XDTOUSJLXOA@LGFMC@    NM:i:1  X1:i:1  MD:Z:17C18
SOLEXA-1GA-2_2_FC20EMB:5:236:644:107    16  chr1    10154   25  36M *   0   0   CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC    ICF=A@BHGEGGDFLIKKYYGD^CahMaShfhNhhh    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:165:628:70 16  chr1    10155   25  36M *   0   0   CCTAACCCTAACCCTAACCTTTACATAACCCTAACC     >D?BAEAA?E=JGGBK>FKDGFJPVHSFTT\QQMch    NM:i:1  X1:i:1  MD:Z:11A2T1T19
SOLEXA-1GA-2_2_FC20EMB:5:108:485:455    16  chr1    10156   25  36M *   0   0   CTAACCCTAACCCTAACCCTAACATAACCCTAACCC     CFVIUIIIBGPROGhRhhhIhhhhhhhhhhhhhfhhh    NM:i:1  X1:i:1  MD:Z:12A23
SOLEXA-1GA-2_2_FC20EMB:5:240:501:237    0   chr1    10158   25  36M *   0   0   AACCCTAACCCTAACCCTAACCTAACCCTAACCATA     hhhchg_ORNbX[]RMZLREQMNTNFLDPMDDDEDKL    NM:i:1  X1:i:1  MD:Z:33A2
SOLEXA-1GA-2_2_FC20EMB:5:258:882:389    16  chr1    10215   25  36M *   0   0   CTAACCCTAAACCTAACCCCTAACCCTAACCCTAAA     IMZO>?EGDRRhhdXKVNRZKhhhhhhhhNhhhhhhhh    NM:i:1  X1:i:1  MD:Z:25A10
SOLEXA-1GA-2_2_FC20EMB:5:197:509:870    16  chr1    10216   25  36M *   0   0   TAAACCGAACCCGAACCCCTAACCCTAACCCTAACC     <IJElC@SCUMY?R?D\UPW^@d]hW2N\hhhhhhLh    NM:i:1  X1:i:1  MD:Z:23G5G2A3
SOLEXA-1GA-2_2_FC20EMB:5:160:880:612    0   chr1    10217   25  36M *   0   0   AACCCTAACCCTAACTCCTAACCCTAACCCTAAACA     hhhNh\O[WUNYQKOWNSIIDNFQPHJHIAI@GDJG    NM:i:1  X1:i:1  MD:Z:15T19A0
SOLEXA-1GA-2_2_FC20EMB:5:249:922:808    16  chr1    10217   25  36M *   0   0   AACACTAACCCCAACCCCTAACCCTAACCCTAAACA     hKWQRBhhB``Ah`hhhhhhhhhhhhhhhhhhhhhh    NM:i:1  X1:i:1  MD:Z:0A23C7A3
SOLEXA-1GA-2_2_FC20EMB:5:13:922:731 0   chr1    10236   25  36M *   0   0   GACCCAAACCCTAAACCCTAAACCCTAAACCCTAACC    hMhWZaPTSZHfUMMSOJTJEKLKELKKDJ>JPCEI    NM:i:1  X1:i:1  MD:Z:0G4A22A7
SOLEXA-1GA-2_2_FC20EMB:5:62:877:892 16  chr1    10237   25  36M *   0   0   ACCCCAACCCTACACCCTACACCCTAACCCTAACCC     CHLKEQGOUcCQ`GYdhF[hXhhhWhhhhhWhhhhh    NM:i:1  X1:i:1  MD:Z:16C6C7C4
SOLEXA-1GA-2_2_FC20EMB:5:172:417:550    16  chr1    10242   25  36M *   0   0   AACCCTAAACCCTAAACCCTAACCCTAACCCTAACC     B@BA>=EG>B@CANEDBEKBMGRCMGVTTORShKhh    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:129:615:872    0   chr1    10243   25  36M *   0   0   ACCCTAAACCCTAAACCCTAACCCTAACCCTAACCC     ghhhhhZ^h]hWeh\^SZ_QOUVL\IQWLTQJOMII    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:223:388:145    0   chr1    10244   25  36M *   0   0   CCCTAAACCCTAAACCCTAACCCTAACCCTAACCCT     g[WhhhUMMOdWTIRNJSIGHHIILLJJBFMB>?B?F    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:196:950:794    16  chr1    10247   25  36M *   0   0   AAAACCCAAAACCCTAACCCTAACCCTAACCCTAACC    LV_KJ@c^PfOOghThKahh`hhhhhhhhhhhhhhhh    NM:i:1  X1:i:1  MD:Z:28A6A0
SOLEXA-1GA-2_2_FC20EMB:5:6:541:676  16  chr1    10316   25  36M *   0   0   ATCACCAACCCTAACCCCTAACCCTAACCCTAACCC     P?X1X?UGCKEHDCJeU[hDVNa[hVQGhhhVhIhhh    NM:i:1  X1:i:1  MD:Z:32A1T1
SOLEXA-1GA-2_2_FC20EMB:5:32:129:784 0   chr1    10316   25  36M *   0   0   CACCCCAACCCTAACCCCTAACCCTCACCCTAACCC     hhhhhhhhhhhhcchhhhhhhhhhZh]HhF^ghQ^THY    NM:i:1  X1:i:1  MD:Z:0C24C10
SOLEXA-1GA-2_2_FC20EMB:5:216:419:993    16  chr1    47098   25  36M *   0   0   TAGAGAAACAGTTTTGAGCCCTTTTATTTTGCTCAA     ETQPP\JJZUVG^HR[hf^hJeReYhhahhhhhhhhh    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:55:82:273  0   chr1    133164  25  36M *   0   0   CCGGAGACCAAGAAGGCGCTGTCCTATTCCCTTCAT     hhhhhhhhVdhhhhZhhRhRhhfTG\?FhI@CGLMCO    NM:i:1  X1:i:1  MD:Z:25A3C2T3
SOLEXA-1GA-2_2_FC20EMB:5:243:393:525    0   chr1    136045  25  36M *   0   0   CTCCGGCTGGAGAAGGGTGTCTGCAGACCCCGCCGG     hhhhhh_hMRKTGBKNPEPGGADE?@BDBCE>=>=?    NM:i:1  X1:i:1  MD:Z:33C0C0G0
SOLEXA-1GA-2_2_FC20EMB:5:296:129:493    16  chr1    137807  25  36M *   0   0   TGTGGACCGACCAAGGCCGCCAGGAGATGGGTAGGC     FD?AD>@?KHD?DIR@KQSNPJSUL[^_UTbS\T]    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:27:706:481 0   chr1    157231  25  36M *   0   0   GGCGGGTAGACCTTGAGATCAGGAGTTCGAGACCAC    hULWMA?J?IF>DE=?E@?ECAEA=>=D=A@AACBA    NM:i:1  X1:i:1  MD:Z:10C24C0
SOLEXA-1GA-2_2_FC20EMB:5:2:433:256  16  chr1    159000  25  36M *   0   0   ACCCACAAACACACACACACACACACACACACCCC      >?@>>>=?=>???C>B@@@C>AGFDIGJDODLMMWS    NM:i:1  X1:i:1  MD:Z:33C2
SOLEXA-1GA-2_2_FC20EMB:5:60:308:161 0   chr1    159000  25  36M *   0   0   ACACACAAACACACACACACACACACACACCCCC      hghb`YTQNUVQHOCGELJRGF@AIRFG@CA?=>CA    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:151:57:170 0   chr1    159000  25  36M *   0   0   ACACACAAACACACACACACACACACACACCCCC      SghhLhVhKdU]ZFEUPRHIDKDICNMICMC@DQJ@A    NM:i:1  X1:i:1  MD:Z:21G14
SOLEXA-1GA-2_2_FC20EMB:5:188:371:826    0   chr1    159000  25  36M *   0   0   ACACACAAACACACACACACACACACACACCCCC      hhhhh[_[hdOh_ePWUXQUX[M\LJJPUQHEEBFH    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:198:594:408    0   chr1    159000  25  36M *   0   0   ACACACAAACACACACACACACACACACACCCCC      hhhhhhhhhhhhh`h_fKeK]R\J_J^ERFSBM@NBV    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:199:941:984    16  chr1    159000  25  36M *   0   0   ACACACAAACACACACACATACACACACACACCCCC     G?=GDD@@AEEJDOHGHUZGMVQhJ[KgRfLhhhhh    NM:i:1  X1:i:1  MD:Z:16T19
SOLEXA-1GA-2_2_FC20EMB:5:230:333:392    0   chr1    159000  25  36M *   0   0   ACACCCAAACACACACACACACACACACACCCCC      UCTDHDFH@FFCIJFGDGJNCBCGBE@C>E@AHAFE    NM:i:1  X1:i:1  MD:Z:4C31
SOLEXA-1GA-2_2_FC20EMB:5:268:422:605    0   chr1    159000  25  36M *   0   0   ACACACAAACACACACACACACACACACACCCCC      WQWQQKJEIKHHEFEHFF@CDQ?DA@???C>?=>?>    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:78:151:802 0   chr1    159001  25  36M *   0   0   CACACAAACACACACACACACACACACACACCCCA      hhhhhhhhhUhRhdhhhKJLVQ\DY^@KECODEECA    NM:i:1  X1:i:1  MD:Z:18G17
SOLEXA-1GA-2_2_FC20EMB:5:81:523:367 0   chr1    159001  25  36M *   0   0   CACACAAACACACACACACACACACACACCCCA      STZOQRFIMCFDHCOEH?DAACB@FAF@C?>?D@E?    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:95:426:691 0   chr1    159001  25  36M *   0   0   CACACCAACACACACAAACACACACACACACCCCA      OWSHCGFJJJJDOLFAIEB>EDA@BB?C>A>@???E@    NM:i:1  X1:i:1  MD:Z:5C8A21
SOLEXA-1GA-2_2_FC20EMB:5:145:524:43 16  chr1    159001  25  36M *   0   0   CACAAAAACACACACACACACACACACACCCCCA      DQA@GFFXEJGR]VNT\hDhCHJhhhTbhhWhhhhhh    NM:i:1  X1:i:1  MD:Z:31A4
SOLEXA-1GA-2_2_FC20EMB:5:203:297:318    0   chr1    159001  25  36M *   0   0   CACACAAACAGACACACACACACACACACACCCCCA     HJLPRLKFIKDKIGIBD=GDDCBAFAF?E?D@@AA>    NM:i:1  X1:i:1  MD:Z:10G25
SOLEXA-1GA-2_2_FC20EMB:5:225:665:744    0   chr1    159001  25  36M *   0   0   CACACAAACACACACACACACACACACACACCCCA      h_PQ_NKTXHhKTUMDTJQDVKF?T@IEEBE@D>?@    NM:i:1  X1:i:1  MD:Z:22G13
SOLEXA-1GA-2_2_FC20EMB:5:251:786:562    0   chr1    159001  25  36M *   0   0   CACACCAACACACACACACACACACACACACCCCA      \hhhhLWShdZQ]NNGHRMFGIVJIGKFGAC?@A>F    NM:i:1  X1:i:1  MD:Z:5C30
SOLEXA-1GA-2_2_FC20EMB:5:3:835:634  0   chr1    159002  25  36M *   0   0   ACACAAACACACAAACACACACACACACACCCCAA      ^ThgYWLRQNMWNHLQGMDIHIFLFKGI?DHAEDA@    NM:i:1  X1:i:1  MD:Z:13A15A6
SOLEXA-1GA-2_2_FC20EMB:5:19:290:488 16  chr1    159007  25  36M *   0   0   AACACAGACACACACACACACCCCAATCCCCA     CC>F?D>C@C@HAF?F?@@CJMRLKHCIgVLI]hSh    NM:i:1  X1:i:1  MD:Z:2C8A17G6
SOLEXA-1GA-2_2_FC20EMB:5:230:771:422    0   chr1    159007  25  36M *   0   0   AACACACACAAAACACACACACCCCCCCGATCTCA      hhhhhhhhhhhhhhhhWhhhhhhZVHCVP>F`FDB?A@    NM:i:1  X1:i:1  MD:Z:10A12C5G6
SOLEXA-1GA-2_2_FC20EMB:5:244:887:293    16  chr1    159007  25  36M *   0   0   AACACACACACACACACACACACAACCCCCAATCCCA     ?`FhEJM?Dhhhhhhhh8OIShhShhhhhUhhhhhhhh    NM:i:1  X1:i:1  MD:Z:2C8A24
SOLEXA-1GA-2_2_FC20EMB:5:187:265:300    16  chr1    250474  25  36M *   0   0   AACAACACACACACACACACACACACACACACACC      C?@??>BDDADEGEIKJKFEGKKNHTYEKRFZP]Va    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:19:384:574 0   chr1    540627  25  36M *   0   0   AAGGGGCGGCCTGTGGCGTTTTCCTGTAAAGTTGGG     hhThhhVa\XOh_JVFDLLQZGCEDFE?B?DGAA>C    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:189:196:190    16  chr1    566744  25  36M *   0   0   AAACAACATAAGCTTCTGACTCTTACCCCCCTCTCC     HKGF0=FMDEPHHJHRNHBSLhKHMhhhhhWh\hhhh    NM:i:1  X1:i:1  MD:Z:0C35
SOLEXA-1GA-2_2_FC20EMB:5:201:774:831    0   chr1    569891  25  36M *   0   0   GGCCTACTAACCAACACACTAACCATATACCAATGA      ag`hhh]hPNPV\RMMXUMMUNKIMFMMPKBCAFDF    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:219:208:94 0   chr1    569894  25  36M *   0   0   CTACTAACCAACACACTAACCATATACCAATGATGG     YLeTeHLPNMMOSWEMIKBEJGHHIHE?>FAGDA<>    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:116:646:499    16  chr1    569895  25  36M *   0   0   TACTAACCAACACACTAACCATATACCAATGATGGC     >SFDhTUNhc[YRMhQhShhhhhdhhhhhhhhhhhhh    NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:244:108:756    16  chr1    569895  25  36M *   0   0   TACTAACCAACACACTAACCATATACCAATGATGGC     E?NIRHGMMWhUPbLK\FYThOh]MhhhhhPbhchhh    NM:i:0  X0:i:1  MD:Z:36
```

Today, if you take a look at Power BI Desktop's options for getting data, you'll see a ton of sources to which you can easily connect. One problem: none of these uniquely help us bioinformaticians.



Putting the "Power" in Power Query

In bioinformatics, there are a plethora of file types for every occasion. Among these are very popular ones such as FASTA (or FASTQ) and BAM and, more recently, GFF3 and BGEN. We can break these data sources down into three main types:

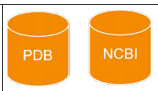


Text-Based Files

Files that are human readable and can be open using virtually any text editor.

Binary Files

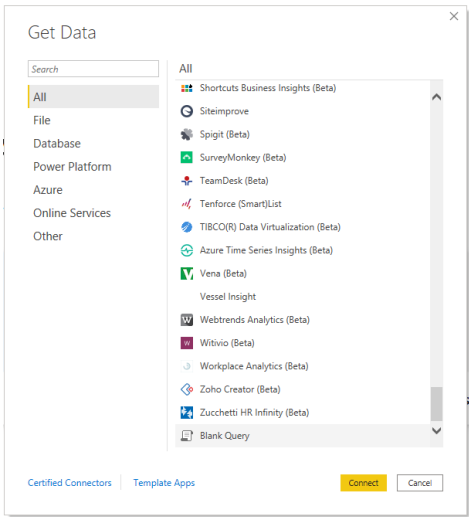
Files that are serialized and must be read by machines.



Online Sources

Databases, webpages, or FTP sites on the internet.

In Power BI, we can take advantage of Power Query to read in data and parse it appropriately. You'll notice that, while Power BI has tons of connects to everything from CSV files to Spark clusters, there are no built-in connectors for our beloved genomics file types (yet?). So, we'll have to use the Blank Query editor.



Text Files

In the query editor, we can write a custom M script to parse our files. For example, to parse a SAM file:

```
// Read SAM Files

let
    // Read in file
    Source = Table.FromColumns({Lines.FromBinary(File.Contents("C:\Users\Colby\Documents\GitHub\bioPowerBI\bam_and_sam\sample.sam"), null, null, 65001)}),
    // Skip @ lines
    #"Filtered Rows" = Table.SelectRows(Source, each not Text.StartsWith([Column1], "@")),
    // Split into columns by \t character and assign names
    // Note: This removes and values past the 11 standard columns
    #"Split Column by Delimiter" = Table.SplitColumn(#"Filtered Rows", "Column1", Splitter.SplitTextByDelimiter("#(tab)", QuoteStyle.Csv), {"QNAME","FLAG","RNAME","POS","MAPQ","CIGAR","RNEXT","PNEXT","TLEN","SEQ","QUAL"}),
    // Change data types
    #"Changed Type" = Table.TransformColumnTypes(#"Split Column by Delimiter",)

in
    #"Changed Type"
```

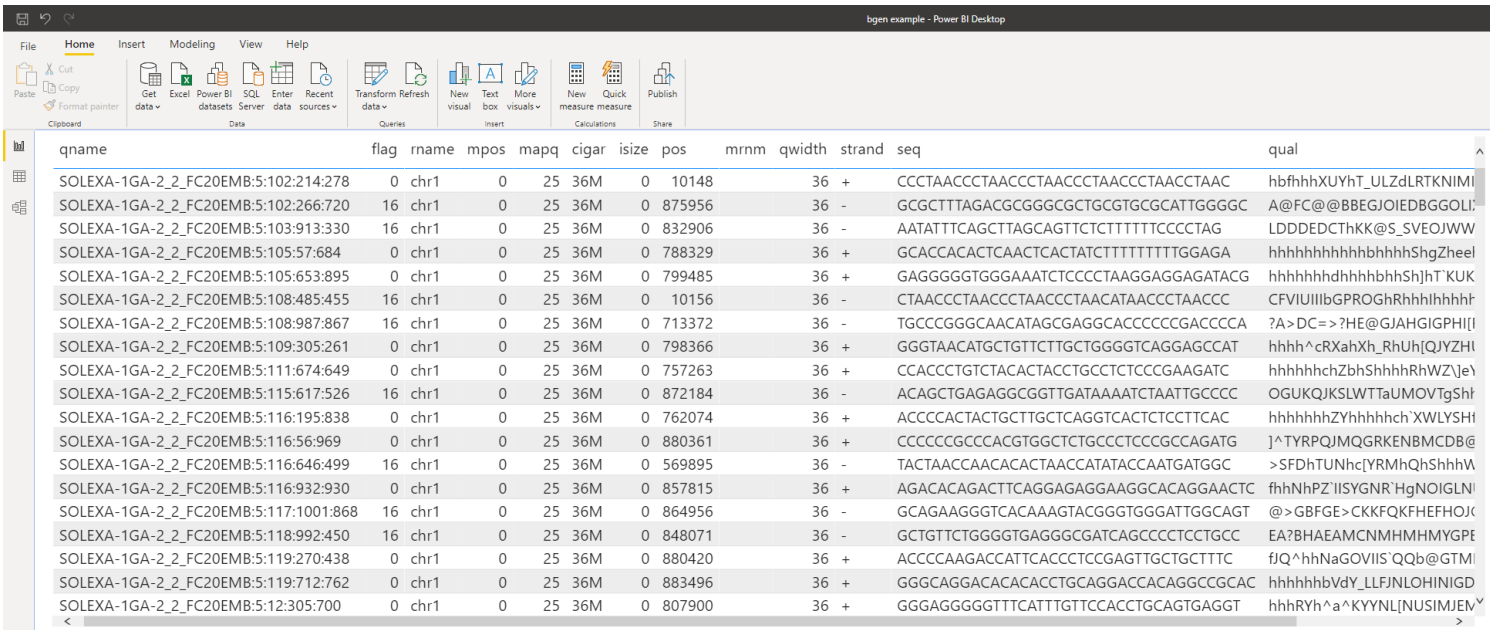
Binary Files

We can use R or Python to read in those pesky binary files as well. (This sometimes make even parsing text files simpler, too.) For example, to parse a BGEN file:

```
// Read BGEN Files

let
    // Use rbgen in an R Script to get the data from the .bgen file
    Source = R.Execute("library(rbgen)#(lf)file <- ""C:\\Users\\Colby Ford\\Desktop\\bioPowerBI\\bgen\\sample.bgen""#(lf)dataset <- as.data.frame(bgen.load(file))"),
    sample = Source{[Name="dataset"]}[Value]
in
    sample
```

...which easily makes the file available as a table.

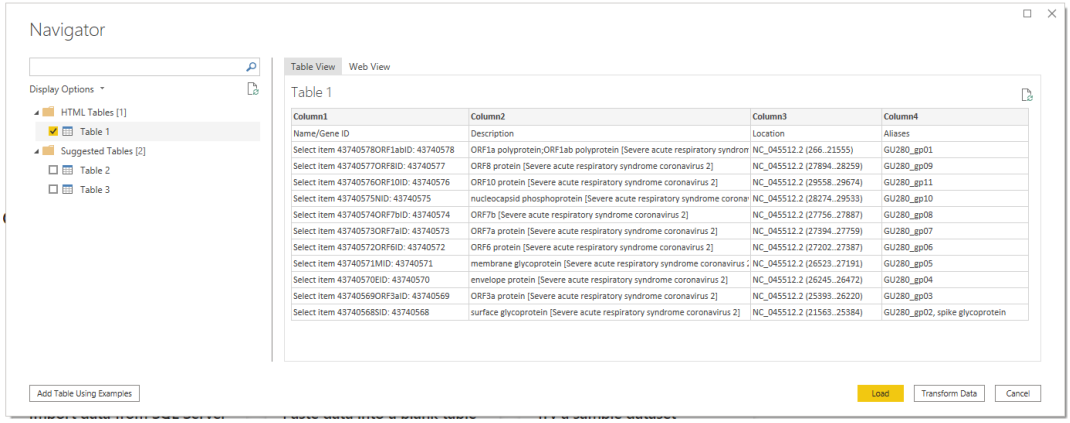


qname	flag	rname	mpos	mapq	cigar	isize	pos	mrnm	qwidth	strand	seq	qual
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278	0	chr1	0	25	36M	0	10148		36	+	CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC	hbfhhhXUYhT_ULZdLRTKNIMI
SOLEXA-1GA-2_2_FC20EMB:5:102:266:720	16	chr1	0	25	36M	0	875956		36	-	GCGCTTTAGACGCGGGCGCTGCGTGCGCATTGGGGC	A@FC@@BBEGJOIEDBGGOLI:
SOLEXA-1GA-2_2_FC20EMB:5:103:913:330	16	chr1	0	25	36M	0	832906		36	-	AATATTTCAGCTTAGCAGTTCTCTTTTTTCCCCTAG	LDDDEDCThKK@S_SVEOJWW
SOLEXA-1GA-2_2_FC20EMB:5:105:57:684	0	chr1	0	25	36M	0	788329		36	+	GCACCACACTCAACTCACTATCTTTTTTTTTGGAGA	hhhhhhhhhhhhbhhhhShgZheel
SOLEXA-1GA-2_2_FC20EMB:5:105:653:895	0	chr1	0	25	36M	0	799485		36	+	GAGGGGGTGGGAAATCTCCCCTAAGGAGGAGATACG	hhhhhhhdhhhhbhhShJhT`KUK
SOLEXA-1GA-2_2_FC20EMB:5:108:485:455	16	chr1	0	25	36M	0	10156		36	-	CTAACCCTAACCCTAACCCTAACATAACCCTAACCC	CFVIUIIIbGPROGhRhhhIhhhhh
SOLEXA-1GA-2_2_FC20EMB:5:108:987:867	16	chr1	0	25	36M	0	713372		36	-	TGCCCGGGCAACATAGCGAGGCACCCCCCGACCCCA	?A>DC=>?HE@GJAHGIGPHI[I
SOLEXA-1GA-2_2_FC20EMB:5:109:305:261	0	chr1	0	25	36M	0	798366		36	+	GGGTAACATGCTGTTCTTGCTGGGGTCAGGAGCCAT	hhhh^cRXahXh_RhUh[QJYZHI
SOLEXA-1GA-2_2_FC20EMB:5:111:674:649	0	chr1	0	25	36M	0	757263		36	+	CCACCCTGTCTACACTACCTGCCTCTCCCGAAGATC	hhhhhhchZbhShhhhRhWZ\JeV
SOLEXA-1GA-2_2_FC20EMB:5:115:617:526	16	chr1	0	25	36M	0	872184		36	-	ACAGCTGAGAGGCGGTTGATAAAATCTAATTGCCCC	OGUKQJKSLWTTaUMOVTgShI
SOLEXA-1GA-2_2_FC20EMB:5:116:195:838	0	chr1	0	25	36M	0	762074		36	+	ACCCCACTACTGCTTGCTCAGGTCACTCTCCTTCAC	hhhhhhhZYhhhhhch`XWLYSHI
SOLEXA-1GA-2_2_FC20EMB:5:116:56:969	0	chr1	0	25	36M	0	880361		36	+	CCCCCCGCCCACGTGGCTCTGCCCTCCCGCCAGATG	]^TYRPQJMQGRKENBMCDB@
SOLEXA-1GA-2_2_FC20EMB:5:116:646:499	16	chr1	0	25	36M	0	569895		36	-	TACTAACCAACACACTAACCATATACCAATGATGGC	>SFDhTUNhc[YRMhQhShhhW
SOLEXA-1GA-2_2_FC20EMB:5:116:932:930	0	chr1	0	25	36M	0	857815		36	+	AGACACAGACTTCAGGAGAGGAAGGCACAGGAACTC	fhhNhPZ`IISYGNR`HgNOIGLN
SOLEXA-1GA-2_2_FC20EMB:5:117:1001:868	16	chr1	0	25	36M	0	864956		36	-	GCAGAAGGGTCACAAAGTACGGGTGGGATTGGCAGT	@>GBFGE>CKKFQKFHEFHOJ(
SOLEXA-1GA-2_2_FC20EMB:5:118:992:450	16	chr1	0	25	36M	0	848071		36	-	GCTGTTCTGGGGTGAGGGCGATCAGCCCCTCCTGCC	EA?BHAEAMCNMHMHMYGPE
SOLEXA-1GA-2_2_FC20EMB:5:119:270:438	0	chr1	0	25	36M	0	880420		36	+	ACCCCAAGACCATTCACCCTCCGAGTTGCTGCTTTC	fJQ^hhNaGOVIIS`QQb@GTM
SOLEXA-1GA-2_2_FC20EMB:5:119:712:762	0	chr1	0	25	36M	0	883496		36	+	GGGCAGGACACACACCTGCAGGACCACAGGCCGCAC	hhhhhhbVdY_LLFJNLOHINIGD
SOLEXA-1GA-2_2_FC20EMB:5:12:305:700	0	chr1	0	25	36M	0	807900		36	+	GGGAGGGGGTTTCATTTGTTCCACCTGCAGTGAGGT	hhhRYh^a^KYYNL[NUSIMJEM

Online Sources

Lastly, Power BI makes it easy to grab data from the web. For example, let's say that I wanted to get a list of all annotated genes of *SARS-CoV-2* from NCBI.

I can quickly grab the URL from my browser and paste it into Power BI, which will then search the page for any tables of information.



This enables users to take advantage of data from virtually any site. Try it out on the Protein Data Bank, NCBI, PlasmoDB, and more!

Takeaway Messages

- Be mindful of memory. Bioinformatics files can be large and, if you're running on a machine with limited resources, you might bog it down.
- Check the defined specifications of any file format you're looking to parse.
- R or Python can be your BFF, especially for binary or really complex file types.

Demo Video

Resources

All code used in above demos and additional examples are available at: www.github.com/BlueGranite/bioPowerBI

If you'd like to learn more about Power BI and how it can help you, contact BlueGranite today.