Encrypted IDAT Format

Mike L. Smith

September 24, 2014

The aim of this document is too briefly describe what's known about the decryption of Illumina's expression IDAT files and the format of the files once they have been decrypted.

Decryption

Most of the information provided here, including the decryption key, was determined by examining the IlluminaExpressionFileCreator module in GenePattern. The decryption key itself appears to be universal (I've not found an expression IDAT that it doesn't work on) and is given here:

```
127, 10, 73, -115, -47, -40, 25, -85
```

The format of each IDAT file follows the same basic layout. Each encrypted IDAT starts with the string `IDAT`, followed by five bytes with the values `1, 0, 0, 0, 8`. I've not seen anything different here, although code comments suggest the first four values may be a version number. The fifth value indicates the length in bytes of the next item to be read, a file specific decryption key. This must be decrypted using the universal key shown above. After this the universal key is no longer needed and the file specific key is used to decrypt the actual data. The data begins immediately after the file key and continues until the end of the file. *illuminaio* decrypts it in blocks of 8 bytes using the now decrypted file key. The output is an XML file.

XML Format

So far I've encountered two different forms for the XML file, although the underlying schema seems the same. On rough observations this appears to be related to date of scanning, but it might be chip type or scanner software related. Listed here are the consistencies and differences I've observed.

Consistencies

The various summarized bead data are stored as attributes within the `<IntensityFile>` tag. Each new attribute is indicated by `__NameOfEntry=` and the data follow in base64 format. Following the opening `<IntensityFile>` tag (and the bead summary data) are a number of tags containing meta data such as the chip type, various software versions, dates for decoding and scanning. Currently *illuminaio* ignores these and extracts on the summarized data.

Differences

The difference between the two types of file seen so far seem primarily to be how many characters are stored per line, and thus the total number of lines.

Older IDAT

- Line length for base64 entries is 76 characters
- Total number of lines $\approx$ 32000

Newer IDAT

- base64 entries stored as a single 'very long' string
- Total number of lines $<$ 100

Implementation

Decryption routines are included with the package and are based on code taken from *gnulib*. File processing is handled by a C function called using `.C` that reads the IDAT file and writes out XML. The XML file is then read into R and each of the data chunks are processed in turn. For each, the base64 text is written to file and then read by the `decode` function from the *base64* package. This writes back out the binary form of the data which is finally read back into R as a vector. The final output is a `data.frame` where each row corresponds to a bead-type and each column is an attribute.

Data

All the IDATs that I've decrypted have contained the same ten fields of summarized bead data. I don't think the format is documented anywhere so I've extrapolated what I think is in each. Users of GenomeStudio might have a better idea of what these are, but I've tried to work them out from looking at the bead-level data. In most cases the values appear to be derived after outliers have been removed.

- MeanBinData - Mean intensity of non-outlying beads.
- TrimmedMeanBinData - Clearly a trimmed mean, but I don't know at what threshold.
- DevBinData - Standard deviation for intensity of non-outlying beads.
- MedianBinData - Apparently the median intensity of non-outlying beads. Sometimes I can't reproduce the value, other times it's fine. Maybe there's some other outlier criteria used here.
- BackgroundBinData - Mean? Median? background value for all beads of this type.
- BackgroundDevBinData - Standard deviation background value for all beads.
- CodesBinData - ProbeID.
- NumBeadsBinData - The total number of beads of this type decoded for the array.
- NumGoodBeadsBinData - The number of beads that were not classed as outliers.
- IllumicodeBinData - ProbeID again. I've not seen this differ from the CodesBinData.