Title
illuminaio: An open source IDAT parsing tool for Illumina microarrays.

Permalink
https://escholarship.org/uc/item/4r39c73q

Authors
Smith, Mike L
Baggerly, Keith A
Bengtsson, Henrik
et al.

Publication Date
2013

DOI
10.12688/f1000research.2-264.v1

Peer reviewed
WEB TOOL

**illuminaio**: An open source IDAT parsing tool for Illumina microarrays [v1; ref status: indexed, http://f1000r.es/2f5]

Mike L Smith¹, Keith A. Baggerly², Henrik Bengtsson³, Matthew E. Ritchie⁴,⁵, Kasper D. Hansen⁶,⁷

¹CRUK Cambridge Institute, Li Ka Shing Centre, The University of Cambridge, Cambridge, CB2 0RE, UK
²Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
³Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94107, USA
⁴Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia
⁵Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3052, Australia
⁶McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA
⁷Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

**Abstract**

The IDAT file format is used to store BeadArray data from the myriad of genomewide profiling platforms on offer from Illumina Inc. This proprietary format is output directly from the scanner and stores summary intensities for each probe-type on an array in a compact manner. A lack of open source tools to process IDAT files has hampered their uptake by the research community beyond the standard step of using the vendor’s software to extract the data they contain in a human readable text format. To fill this void, we have developed the illuminaio package that parses IDAT files from any BeadArray platform, including the decryption of files from Illumina’s gene expression arrays. illuminaio provides the first open-source package for this task, and will promote wider uptake of the IDAT format as a standard for sharing Illumina BeadArray data in public databases, in the same way that the CEL file serves as the standard for the Affymetrix platform.
**Introduction**

The DNA microarray field is dominated by the three manufacturers: Affymetrix, Illumina and Agilent. While the basic premise behind their competing products is the same (i.e. the measurement of hybridisation between sample and immobilised probes on arrays via fluorescence), the formats in which these data are presented to end users are quite different, with each manufacturer electing to use their own proprietary format. The most ubiquitous of these is the CEL file, which has been accepted as a standard format for publishing the raw data generated on the Affymetrix platform. A search of the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database finds over 90% of submissions of Affymetrix data include one or more CEL files as supplementary material. The format itself is well documented by the manufacturer, who also provides an open-source software development kit (SDK). As a result, in addition to Affymetrix’s own software suite, a large number of CEL parsing tools exist, including a parser implemented based on the file format documentation: affyio⁴ and a parser based on the SDK: affsparser⁵.

The same is not true of the primary IDAT format from Illumina, with only 1.5% (49 out of 3208) of the submissions in GEO that use Illumina BeadArrays including IDAT files as supplementary material. Given that IDATs are the standard file type generated during BeadArray processing, it seems reasonable to assume that the relative dearth of IDAT files in the public domain is due to the lack of widespread support for the format. The development of alternative parsing tools has proven more challenging for IDATs for a number of reasons. The foremost amongst these is a lack of public documentation, leaving tool developers to determine the file structure themselves. A further hurdle has been the encryption of IDAT files generated from expression chips. These barriers initially left researchers reliant on the output from Illumina’s GenomeStudio software to convert the data into a more convenient format. Existing open source tools, particularly those that focus on gene expression analysis such as beadarray⁶, limat⁷ and limma⁸, all require that the IDAT files have been processed using GenomeStudio to generate a plain-text ASCII file before any analysis can take place (Figure 1).

The GenePattern⁹ software suite includes support for reading expression IDAT files, although it is limited to extracting only a subset of the array information. GenomeStudio output also omits various information that is available from the IDAT, such as control probe intensities (for SNP and methylation platforms), so-called out-of-band probes (methylation 450k)⁹ and meta information including software versions and scan date (all platforms).

Here we introduce the Bioconductor⁴ package illuminaio that can handle IDAT files from any Illumina BeadArray platform, providing a simple unified interface to various low-level data extraction routines.

**Data format**

The IDAT file format varies depending upon the array platform (Table 1). IDATs generated during the scanning of genotyping and methylation BeadArrays are binary files (one for each of the red and green channels). The bulk of each file is comprised of four fields: the ID of each bead-type on the array, the mean and standard deviation of their intensities, and the number of beads of each type. Additionally, metadata including the date the array was scanned, specific software versions used and the type of BeadChip are also included. Once the structure of the file is understood these binary values can be read directly.

On the other hand, gene expression IDAT files are produced as encrypted XML files. Once decrypted the majority of the data are found as ten Base64 encoded strings. These ten fields include the ID, mean and standard deviation values as found in genotyping IDATs, as well as median and trimmed-mean intensity values, the mean and standard deviation of local background intensities, and the number of beads both before and after outliers have been excluded.

Each array type is also associated with a manifest file (with file extension BPM or BGX) that provides details of probe sequences, intended genomic targets and whether it is a control probe or not, information that is necessary to correctly interpret the data.

**Implementation**

 illuminaio is an R package⁵. The reading of IDAT files is achieved using the readIDAT function. This routine is able to determine
the type of IDAT file that has been passed and calls the appropriate
code to read the file and return the data as a R list object (Figure 1).
This not only contains intensity data, but also the meta information
such as scan date that are not routinely extracted and can be useful
for detecting batch effects10.

Decryption of expression IDATs is performed using the open-
source DES decryption routine available in Gnutlfib7. There is no
official documentation of this file format, but illumiaino includes
a document describing our findings in detail. Source code for the
appropriate routines has been adapted and included in illumiaino,
removing any requirement for specific external libraries to be
installed on a user’s computer. Thus the package can be built and run
on all three major operating systems (Linux, Windows and Mac).

The illumiaino package also supports the parsing of non-encrypted
IDAT files compressed by gzip and the reading of manifest files
describing the array design (readBGX and readBPM).

The summarised intensity values obtained by illumiaino are nearly
identical to those reported using GenomeStudio. Small discrepancies
related to rounding performed by GenomeStudio are observed. The
package vignette contains a detailed comparison. The time taken to
read an IDAT depends on the platform, with encrypted expression
arrays taking around 1 second per file (for 50,000 probes), and
methylation and SNP platforms between 1 to 6 seconds depending
on the chip density (which can range between a few hundred thou-
sand and several million probes).

Discussion
The availability of an open-source IDAT reader through illumiaino
that can read files from any of Illumina’s BeadArray technologies
will promote greater use of the IDAT file as a primary data format in
the analysis and sharing of results from BeadArray based profiling
studies. The illumiaino package is intended for use by developers
to efficiently extract the content of both IDAT and bead-manifest
files, thereby expanding the possibilities for conducting reproduc-
able research with these data.

One exception to the dearth of IDAT files noted in the introduc-
tion is the The Cancer Genome Atlas (TCGA, http://cancergenome.
nih.gov/). IDAT files from Illumina methylation and genotyping
arrays are available in large numbers as Tier 1 data from the TCGA
website (https://tcga-data.nci.nih.gov/tcga/). Of particular interest
is the Illumina 450k methylation array, for which Triche et al.7
has shown improvements in background correction by using out-
of-band probes, information that is only available through IDAT
files and not the GenomeStudio output. For this work Triche et al.
used illumiaino to access the out-of-band probes, which shows the
advantage of having access to low-level data.

illumiaino is currently used in the mingi12, methylumi13 and crimm14,15
packages for importing IDAT files from the Infinium methylation
and genotyping platforms respectively, demonstrating its utility.

Software availability
illumiaino is an R package available from the Bioconductor project
(http://www.bioconductor.org) and from 10.5281/zenodo.7588.

Author contributions
KAB developed the first version of the IDAT reader for unencrypted
IDAT files. This work was later improved by HB, MER and KDH.
MLS developed the IDAT reader for encrypted files. All authors
wrote and approved the manuscript.

Competing interests
No competing interests were disclosed.

Grant information
This work was supported by the European Community’s Seventh
Framework Programme under grant agreement No. 305626 (Pro-
ject RADIANT) (MLS), NHMRC Project grant 1050661, Victorian
State Government Operational Infrastructure Support and Australian
Government NHMRC IRIISS (MER) and Grant Number R01
GM103552 from the National Institute of General Medical Sciences,
National Institutes of Health (KDH).

The funders had no role in study design, data collection and analysis,
decision to publish, or preparation of the manuscript.

References
1. Bolstad BM: affyio: Tools for parsing Affymetrix data files. R package version 1.30.0. Reference Source
2. Bengtsson H, Bukard J, Hansen KD: affparsers: Affymetrix File Parsing SDK. R package version 1.34.0. 2013. Reference Source
3. Dunning MJ, Smith ML, Ritchie ME, et al.: beadarray: R classes and methods for Illumina bead-based data. Bioinformatics. 2007; 23(16): 2183–2184. PubMed Abstract | Publisher Full Text
4. Du P, Kibbe WA, Lin SM: lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008; 24(13): 1547–1548. PubMed Abstract | Publisher Full Text
5. Smyth GK: Limma: linear models for microarray data. In: R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, Bioinformatics and Computational Biology Solutions Using R and Bioconductor, 2005; 397–420. Springer, New York. Reference Source

Table 1. Summarising the file formats for various BeadArray platforms.

| Array type       | File format | No. data fields |
|------------------|-------------|-----------------|
| SNP genotyping   | Binary      | 4               |
| Methylation      | Binary      | 4               |
| Gene expression  | Encrypted XML | 10             |
6. Reich M, Liefield T, Gould J, et al.: GenePattern 2.0. Nat Genet. 2006; 38(5): 500–501. PubMed Abstract | Publisher Full Text
7. Triche TJ Jr, Weissenberger DJ, Van Den Berg D, et al.: Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013; 41(7): e90. PubMed Abstract | Publisher Full Text | Free Full Text
8. Gentleman RC, Carey VJ, Bates DM, et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5(10): R80. PubMed Abstract | Publisher Full Text | Free Full Text
9. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. Reference Source
10. Leek JT, Scharpf RB, Bravo HC, et al.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11(10): 733–739. PubMed Abstract | Publisher Full Text
11. GNU Project. GnuLib - The GNU Portability Library, 2013. Reference Source
12. Hansen KD, Aryee M: minfi: Analyze Illumina's 450k methylation arrays, 2013. R package version 1.8.3. Reference Source
13. Davis S, Du P, Bilke S, et al.: methylumi: Handle Illumina methylation data, R package version 2.8.0. 2013. Reference Source
14. Ritchie ME, Carvalho BS, Hetrick KN, et al.: R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. Bioinformatics. 2009; 25(19): 2621–2623. PubMed Abstract | Publisher Full Text | Free Full Text
15. Scharpf RB, Irizarry RA, Ritchie ME, et al.: Using the R Package crlmm for Genotyping and Copy Number Estimation. J Stat Softw. 2011; 40(12): 1–32. PubMed Abstract | Free Full Text
Referee Responses for Version 1

Tiffany Morris
UCL Cancer Institute, University College London, London, UK

Approved: 29 January 2014

Referee Report: 29 January 2014
This manuscript written by Smith et al. presents an R package to process IDAT files, the file format of Illumina BeadArrays. The package illuminaio fills an important gap in data processing by enabling the analysis of Illumina data without Illumina software. The gap is illustrated when the authors cite the disparity between the availability of CEL files (Affymetrix file format) and IDAT files in public data repositories. In addition to eliminating the need to use Illumina software, illuminaio extracts additional information from the IDAT files allowing improved quality control analyses and additional data correction algorithms. Figure 1 of the manuscript illustrates the simplified analysis pipeline enabled by illuminaio.

The authors briefly describe the IDAT file format that was not publically documented before illuminaio. Figure 2 of the manuscript summarises the details of the file format for each of the BeadArray platforms. As such this is not just the development of a new tool, as the authors have provided thorough documentation for a widely used platform. In addition, the vignette accompanying illuminaio clearly explains the packages methods and how to use them.

This package, illuminaio, has already led to the development of many new analysis packages that utilise IDAT files and the additional information they provide. In turn these packages are leading to the increased exchange and availability of the raw IDAT files in collaborations and data repositories.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

1 Comment

Author Response

Kasper Hansen, khansen@jhsph.edu, USA
Posted: 06 Feb 2014

Dr. Morris,

Thank you for your positive review. We are happy that you find our tool useful.

Best,
Jan Oosting  
Department of Pathology, Leiden University Medical Center, Leiden, Netherlands  

Approved: 27 January 2014  

Referee Report: 27 January 2014  
The article describes illuminaio, an R package to process the raw data files produced by the Illumina scanning software. This tool is valuable, because it enables researchers to use a completely open analysis workflow, without having to use a closed source, blackbox, analysis step.

I have tried the package on IDAT files from several chiptypes, and only files from the beginning of 2005 scanned with beadscan 2.2 were rejected. Files scanned with beadscan 2.3 (Nov 2005) and onwards were processed properly.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

1 Comment

Author Response  

Kasper Hansen, khansen@jhsph.edu, USA  
Posted: 06 Feb 2014  

Dr. Oosting,

Thank you for your review. I am especially happy that you have taken the effort to apply illuminaio on output from older versions of beadscan.

We have contacted Dr. Oosting offline and have received a number of files from old scans. We have started implementing support for these older files, and initial results are promising. We expect to release a new version of illuminaio shortly, which will read these files.

Best,
Kasper D. Hansen  

Competing Interests: No competing interests were disclosed.