`recountmethylation` cheatsheet

TERMINOLOGY

DNA methylation (DNAm) terms

CpG locus	:	Cytosine-guanine dinucleotide pair
DNA methylation (DNAm)	:	DNA with covalently bound methyl (CH_3) groups; here specify cytosine-bound methyl groups in a CpG locus.
CG Island	:	DNA regions enriched for cytosine, guanine, and CpG loci.

BeadArray terms

CpG probe	:	Class of DNAm array technology using bead-bounded probes to quantify DNAm at specific CpG loci.
BeadArray	:	Type of microarray using BeadArray probes, manufactured by Illumina, to quantify DNAm.
HM27k	:	Type of DNAm array, introduced around 2005, targeted roughly 27,000 CpG loci in humans.
HM450k	:	Type of DNAm array, introduced around 2011, targeting roughly 450,000 CpG loci in humans.
EPIC	:	Type of DNAm array, introduced around 2015, targeting roughly 855,000 CpG loci in humans. Shares 93% (about 453,000) of probes with the HM450k platform.

Data object classes

`HDF5` / `h5`	:	Hierarchial database format 5, a type of database syntax implementing compression and chunking.
`se`	:	Short for `SummarizedExperiment` , a multifaceted object class in R/Bioconductor containing slots for assay measurements and metadata for the platform, samples, and experiment.
`h5se`	:	Short for `HDF5-SummarizedExperiment` , a hybrid object class that uses `DelayedArray` for caching.
`RGChannelSet` / `rg`	:	Type of `se` object containing dual color channel data for red and green channels on Illumina's BeadArray DNAm array platforms.
`GenomicMethylSet` / `gm`	:	Type of `se` object containing the methylated (a.k.a. 'M') and unmethylated (a.k.a. 'U') signals calculated from dual channel intensity data.
`GenomicRadioSet` / `gr`	:	Type of `se` object containing the DNAm fractions (a.k.a. Beta-values) and/or logit2-transformed fractions (a.k.a. M-values) calculated from the M and U fractions.

Download DNAm compilations

The following command line options show how to download database files from the server (http://www.recount.bio/data (http://www.recount.bio/data)), or you may try to `right-click -> download` from within your browser. Note, it may help to increase your timeout period for long downloads.

- Use `recountmethylation` : e.g. `getdb_h5se_rg("hm450k")` to download the HM450k `RGChannelSet` data as an `HDF5-SummarizedExperiment` object (see `?getdb` for more info).
- Use `wget` : From command line, enter `wget -r <filepath>` , replacing `<filepath>` with an address from `https://www.recount.bio/data` . Note the `-r` is needed for `h5se` objects, which are directories.

`DelayedArray` operations and pipes

The following operations make use of `DelayedArray` caching.

- Perform summary operations with `dim()` / `nrow()` / `ncol()`
- Rapidly update the sample metadata in an `h5se` object with `quickResaveHDF5SummarizedExperiment()` .
- Automatically pipe data chunks between `DelayedArray` objects, e.g.

```
library(HDF5Array)
h5 <- loadHDF5SummarizedExperiment(h5se.path)
bval <- getBeta(h5) # this is delayed
saveHDF5Array(bval, file = "db.h5") # executes with chunking
```

RECAST DATASET OBJECTS

If operations on datasets throw errors due to their class, you may attempt to recast and rerun them without the `DelayedArray` backend.

Recast an `RGChannelSet`

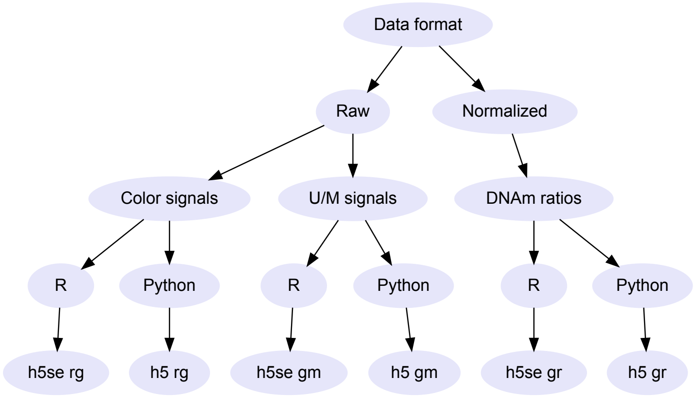
```
library(minfi)
rg <- loadHDF5SummarizedExperiment(rg.path)
green.matrix <- as.matrix(getGreen(rg))
red.matrix <- as.matrix(getRed(rg))
anno <- annotation(rg)
metadata <- DataFrame(pData(rg))
rg.new <- RGChannelSet(Green = green.matrix,
                       Red = red.matrix,
                       annotation = anno,
                       colData = metadata)
```

Recast a `GenomicRatioSet`

```
library(minfi)
gr <- loadHDF5SummarizedExperiment(gr.path)
bval.matrix <- as.matrix(getBeta(gr))
anno <- annotation(gr)
metadata <- DataFrame(pData(gr))
ranges <- granges(gr)
gr.new <- GenomicRatioSet(gr = ranges,
                          Beta = bval.matrix,
                          annotation = anno,
                          colData = metadata)
```

DECISION TREE

Once you've selected a DNAm array platform (e.g. HM450k or EPIC), you can use the following decision tree to determine which database compilation object type to download.


Data format
Raw
Normalized
Color signals
U/M signals
DNAm ratios
R
Python
R
Python
R
Python
h5se rg
h5 rg
h5se gm
h5 gm
h5se gr
h5 gr