# Meta-analysis

Levi Waldron, CUNY School of Public Health

July 14, 2016

# Outline

- Preparation for meta-analysis
- Fixed and Random Effects Synthesis
- Assessing Heterogeneity
- Leave-one-dataset-in and Leave-one-dataset-out Validation of Prediction Models
- `curatedMetagenomicData` and `MultiAssayExperiment`

# Scope: what is meta-analysis?

- Broad definition: the full scope of among-study analysis
- Narrow definition: a synthesis of per-study estimates
- Not: pooling of per-patient data ("mega-analysis")

> *"We understand meta-analysis as being the use of statistical techniques to combine the results of studies addressing the same question into a summary measure."*
>
> *Villar et al. (2001)*

# Preparation: finding datasets

- Systematic literature review
- Gene Expression Omnibus (GEO)
  - *web page (filter by species, disease, sample size)*
  - *GEOmetadb Bioconductor package (requires SQL knowledge)*
- ArrayExpress
  - *also includes many GEO datasets*
  - *Bioconductor package has search features*
- InSilicoDB
  - *better curation, lower coverage*

# Preparation: downloading datasets

- `GEOquery::getGEO()` is a workshorse
  - *maximum coverage, minimum frills*
  - *all metadata included, most is irrelevant*
  - *large studies limited to 256 patients per list element*
  - *processed data as uploaded by authors -> list of ExpressionSets*
  - *no probeset to gene mapping*

# Preparation: downloading datasets (cont'd)

- A couple helpful functions from LeviRmisc
  - *`getGEO2()`: consolidate and simplify `getGEO()` output*
  - *`geoPmidLookup()`: look up experiment and publication data from GEO and Pubmed, put in dataframe*

```
## BiocLite("lwaldron/LeviRmisc")
library(LeviRmisc)
df <- geoPmidLookup(c("GSE26712", "PMID18593951"))
```

```
## [1] "WARNING: please set your email using Sys.setenv(email='name@email.com')"
```

```
df[, c(1:3, 15, 16)]
```

```
##              pubMedIds platform_accession platform_summary     journal
## GSE26712      18593951              GPL96          hgu133a Cancer Res.
## PMID18593951  18593951               <NA>             <NA> Cancer Res.
##              volume
## GSE26712         68
## PMID18593951     68
```

# Preparation: curation

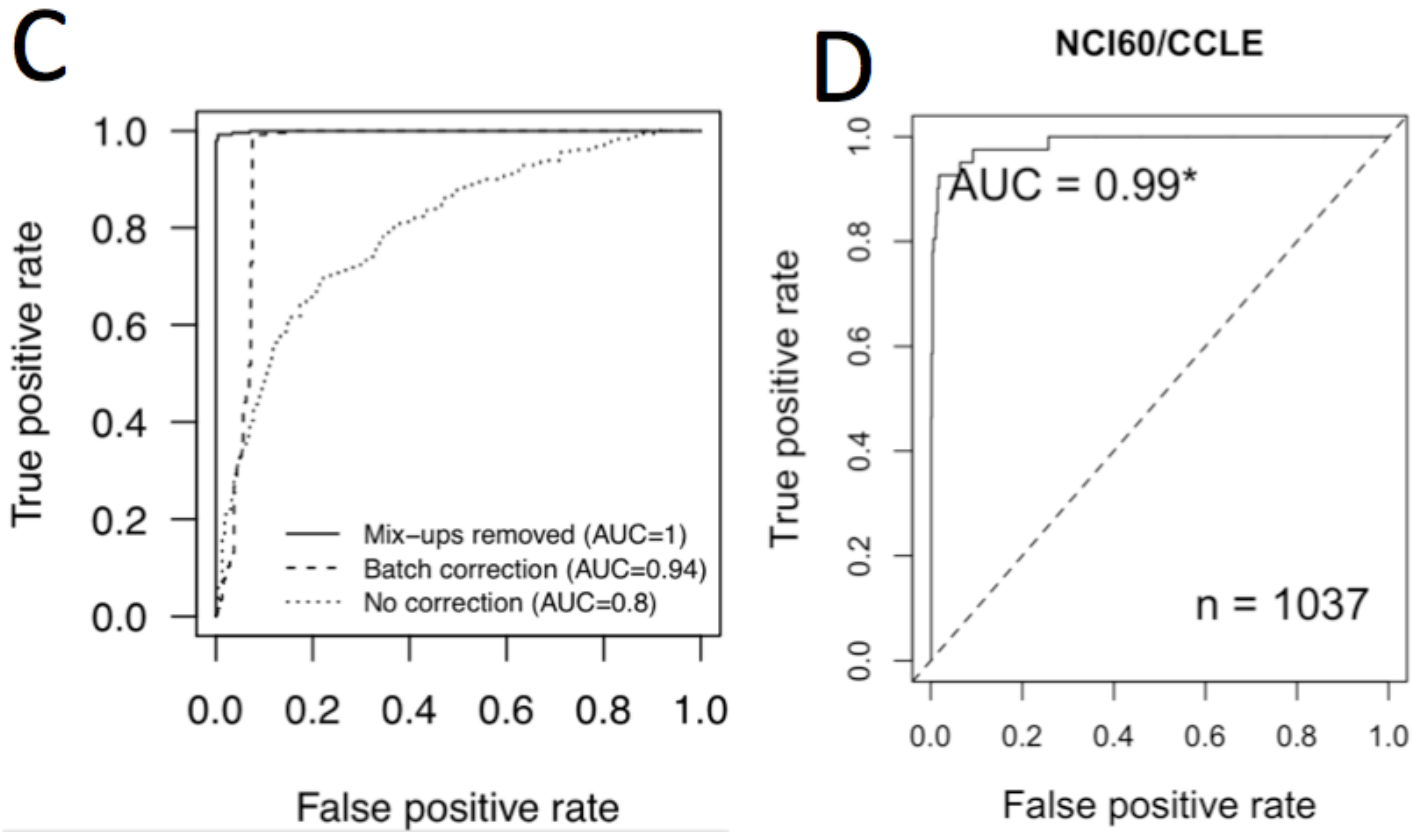
- per-sample metadata must be standardized across studies
- process is error-prone, template-based syntax checking recommendable
  - *e.g. using the template and checker for curatedOvarianData.*

# Preparation: preprocessing and gene mapping

- it is possible and desirable to synthesize across array platforms
  - *or spanning array and RNA-seq*
- common preprocessing is desirable but not necessary
  - *deal with non-standardized preprocessing through gene scaling, e.g. z-score*
- must map probeset IDs to common gene identifiers:
  - *if using a representative probeset for a gene, best to use the same one in each dataset*
  - *alternatively, simply average redundant probesets*

# Preparation: duplicate checking

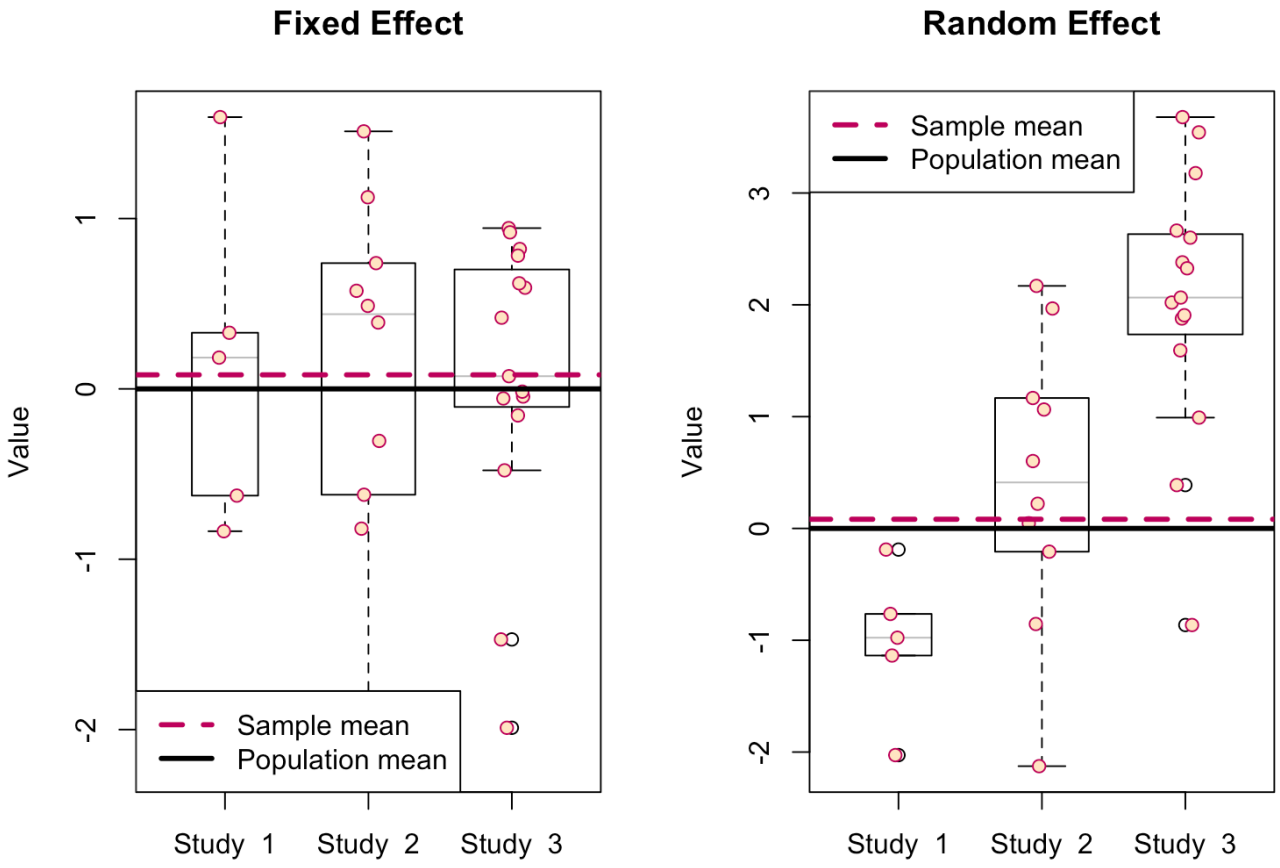
- duplicate samples bias meta-analysis
  - *doppelgangR Bioconductor package for high-throughput duplicate checking*




C: Matching RNAseq to microarray, D: matching cell lines between CCLE and NCI-60

Waldron L, *et al.*: The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. J. Natl. Cancer Inst. 2016, 108.

# Fixed and Random Effects Synthesis




- Fixed effect: population mean of all studies is $\theta$
- Random effect: population mean of study $k$ is $\theta + \mu_k; \mu_k \overset{iid}{\sim} N(0, \tau^2)$

# Assessing Heterogeneity

- Q-test: Under the null hypothesis of no heterogeneity between studies ($\tau = 0$),

$$Q \sim \chi^2_{K-1}$$

- Standard descriptions of heterogeneity:
  - $\tau^2$: *estimate of total amount of heterogeneity*
  - $I^2$: *fraction of total variability due to heterogeneity*
- For further info:
  - *Viechtbauer W: Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 2010.*

# Example 1: Is CXCL12 gene a prognostic factor for ovarian cancer?

Load the curatedOvarianData package, look at available datasets:

```
library(curatedOvarianData)
data(package="curatedOvarianData")
```

Load (and check out) rules defined in default configuration file:

```
downloader::download("https://bitbucket.org/lwaldron/ovrc4_sigvalidation/raw/tip/input/patientselect
                     destfile="patientselection.config")
source("patientselection.config")
impute.missing <- TRUE
keep.common.only <- TRUE
```

# Example 1 (cont'd)

- Calculate "effect size" log(HR) and S.E. for one dataset:

```
runCox <- function(eset, probeset="CXCL12"){
  library(survival)
  eset$y <- Surv(eset$days_to_death, eset$vital_status == "deceased")
  if(probeset %in% featureNames(eset)){
    obj <- coxph(eset$y ~ scale(t(exprs(eset[probeset, ]))[, 1]))
    output <- c(obj$coefficients, sqrt(obj$var))
    names(output) <- c("log.HR", "SE")
  }else{output <- NULL}
    output}
runCox(esets[[1]])
```

```
##    log.HR        SE
## 0.1080378 0.1167063
```

# Example 1 (cont'd)

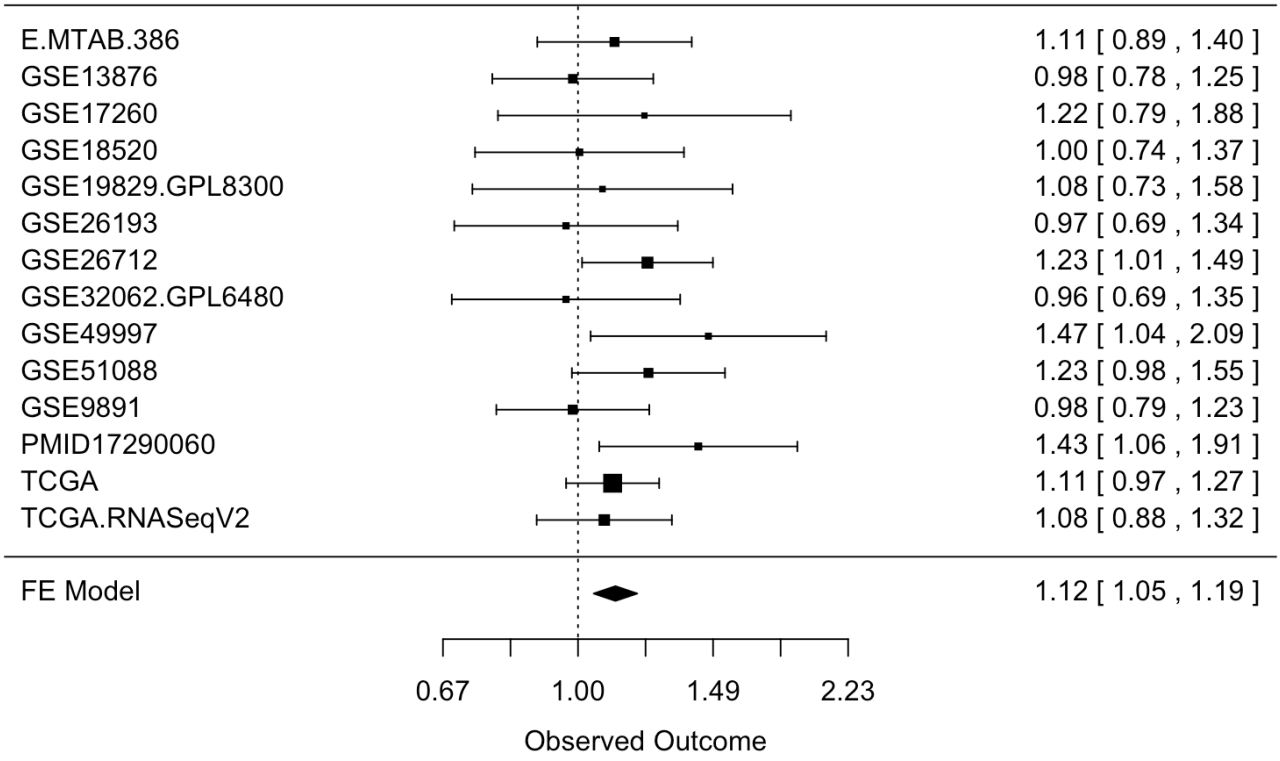
- Calculate "effect size" (HR) and Standard Error for all datasets:

```
(study.coefs <- t(sapply(esets, runCox)))
```

```
##                              log.HR         SE
## E.MTAB.386_eset         0.108037829 0.11670634
## GSE13876_eset          -0.015533625 0.12165106
## GSE17260_eset           0.196604844 0.22132140
## GSE18520_eset           0.004334577 0.15785733
## GSE19829.GPL8300_eset   0.072413433 0.19658498
## GSE26193_eset          -0.035518891 0.16886806
## GSE26712_eset           0.205703027 0.09889057
## GSE32062.GPL6480_eset -0.035661806 0.17253159
## GSE49997_eset           0.386074941 0.17795245
## GSE51088_eset           0.208534008 0.11565319
## GSE9891_eset           -0.015481600 0.11555760
## PMID17290060_eset       0.356194786 0.14969168
## TCGA_eset               0.102434252 0.07029190
## TCGA.RNASeqV2_eset      0.077791413 0.10215517
```

# Example 1 (cont'd): forest plot

```
library(metafor)
res.fe <- metafor::rma(yi=study.coefs[, 1], sei=study.coefs[, 2], method="FE")
forest.rma(res.fe, slab=gsub("_eset$","",rownames(study.coefs)), atransf=exp)
```



# Example 1 (cont'd): FE vs. RE

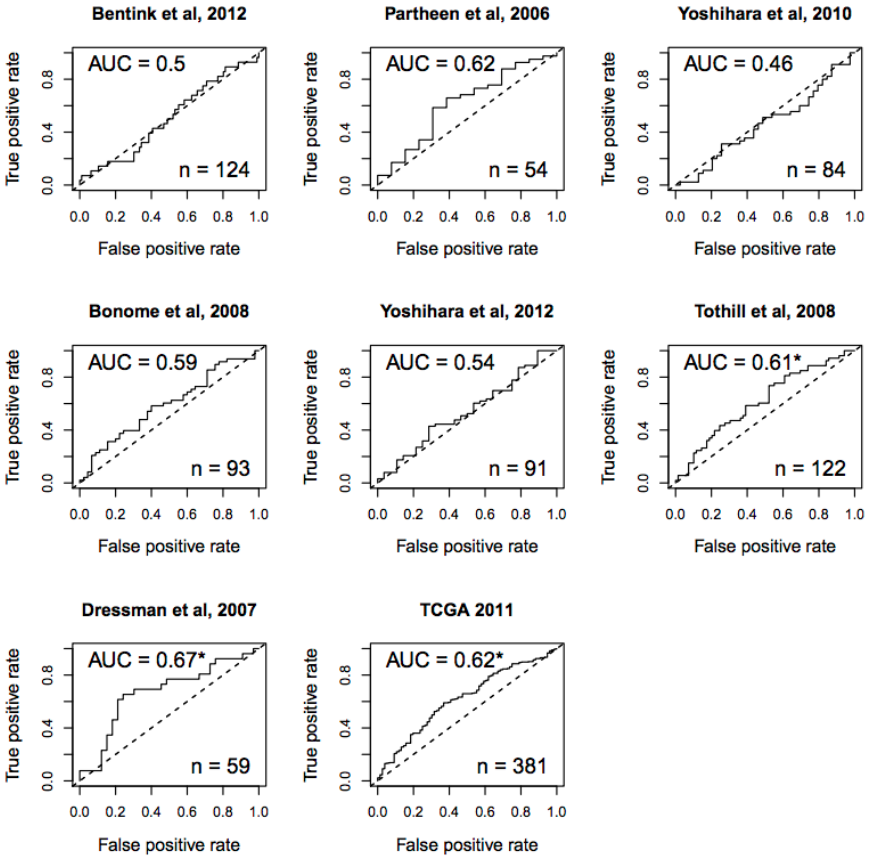
```
(res.re <- metafor::rma(yi=study.coefs[, 1], sei=study.coefs[, 2], method="DL"))
```

```
##
## Random-Effects Model (k = 14; tau^2 estimator: DL)
##
## tau^2 (estimated amount of total heterogeneity): 0 (SE = 0.0062)
## tau (square root of estimated tau^2 value):      0
## I^2 (total heterogeneity / total variability):   0.00%
## H^2 (total variability / sampling variability):  1.00
##
## Test for Heterogeneity:
## Q(df = 13) = 11.2219, p-val = 0.5922
##
## Model Results:
##
## estimate       se     zval     pval    ci.lb    ci.ub
##   0.1108   0.0329   3.3664   0.0008   0.0463   0.1754      ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Example 1 (cont'd): closing comments

- Replace simple univariate regression with multivariate regression to correct for known clinical factors (e.g. see Ganzfried *et. al.* 2013)
- Replace HR with any coefficient + S.E.
- Replace single probeset or gene with any score or classifier

# Example 2: Leave-one-dataset-out validation



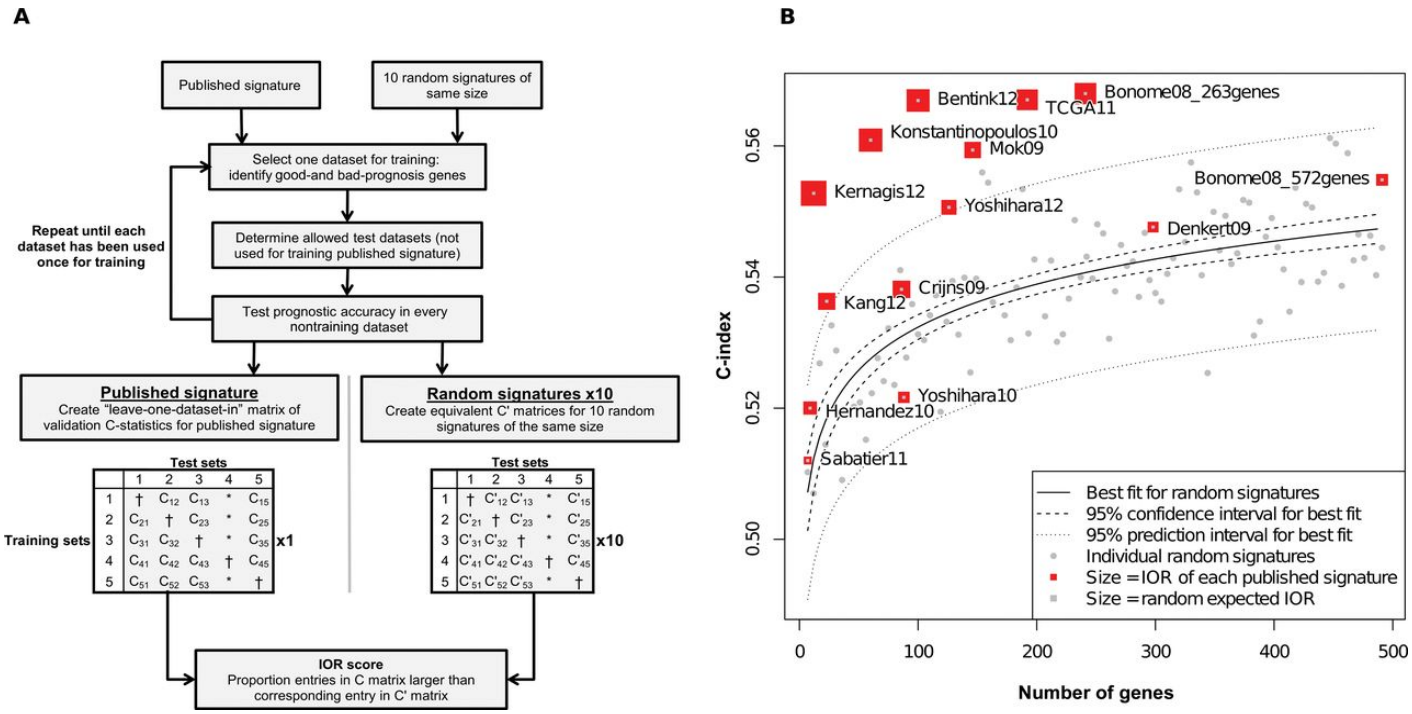

Leave-one-dataset-out validation of a survival signature. (Riester *et al.* JNCI 2014)

# Example 3: Leave-one-dataset-in validation

- Independent datasets for evaluation of prediction models or gene signatures
- Train and test using all dataset pairs (Waldron *et al.* JNCI 2014, Bernau *et al.* Bioinformatics 2014, Zhao *et al.* Bioinformatics 2014)

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | CV | $Z_{12}$ | $Z_{13}$ | $Z_{14}$ | $Z_{15}$ |
| 2 | $Z_{21}$ | CV | $Z_{23}$ | $Z_{24}$ | $Z_{25}$ |
| 3 | $Z_{31}$ | $Z_{32}$ | CV | $Z_{34}$ | $Z_{35}$ |
| 4 | $Z_{41}$ | $Z_{42}$ | $Z_{43}$ | CV | $Z_{45}$ |
| 5 | $Z_{51}$ | $Z_{52}$ | $Z_{53}$ | $Z_{54}$ | CV |

# Leave-one-dataset-in validation (cont'd)




"Improvement over random signatures (IOR)" score of gene signatures relative to random gene signatures, equalizing the influences of authors' algorithms for generating risk scores, quality of the original training data, and gene signature size (Waldron *et al.* JNCI 2014).

# Resources in Bioconductor

- Cancer gene expression data packages:
  - *curatedOvarianData, curatedCRCData, curatedBladderData*
- `curatedMetagenomicData`, available through `ExperimentHub` in bioc-devel
  - *taxonomic and metabolic profiles from whole-metagenome shotgun sequencing*
  - *~3,000 human microbiome samples from 13 datasets*
  - *manually curated metadata*
- Provides six `ExpressionSet` objects per dataset:
  - *1: species-level taxonomic profiles (convertible to `phyloseq`);*
  - *2-3: marker presence and abundance data; and*
  - *4-6: gene families, pathway coverage and pathway abundance.*
- Manual of datasets

# curatedMetagenomicData and ExperimentHub

```
library(ExperimentHub)
eh = ExperimentHub()
myquery = query(eh, "curatedMetagenomicData")
```

```
myquery
View(mcols(myquery))
```

```
taxabund = eh[["EH2"]]
taxabund
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 3302 features, 38 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: H10 H11 ... IT8 (38 total)
##   varLabels: dataset_name sampleID ... group (211 total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
##   pubMedIds: 25981789
## Annotation: NA
```

# Conclusions

- many alternatives for meta-analysis of genomics experiments have been proposed, none as flexible or well-understood as traditional approaches
- metafor R package is highly recommendable and well-documented (Viechtbauer 2010)
- data availability and curation are critical