# Reanalysis of mouse-rat comparative gene expression data towards batch effects

Final Project for Modeling of complex biological systems

Maciej Bielecki

# Table of contents

# 1. Introduction

Batch effects in molecular biology refer to artifacts in the data caused by various factors, seemingly unrelated to the variables being actively researched. These artifacts do not impact singular data points - rather, entire series of data points, as determined by the design of the study, may be affected by a common cause, potentially a very mundane one, such as atmospheric conditions, reagents or tools used for a given batch, or difference in personnel carrying out the experiments in separate series (Goh *et al*., 2017).

The unwanted variation caused by batch effects may be so significant so as to "overpower" the actual, underlying variation of the data. One publication to underscore both the potential impact of batch effects, and the importance of accounting for their presence in data was by Gilad and Mizrahi-Man (2015), wherein the authors reanalyze the dataset published by Lin *et al*. (2014). Notably, said dataset, which is a comparative RNA-Seq dataset of gene expression in various human (*Homo sapiens*) and mouse (*Mus musculus*) tissues, seemed to imply that gene expression is generally more similar among different tissues within each species, than between corresponding tissues of separate species (henceforth, such a tendency will be referred to as "*intra*-over-*inter* clustering").

These findings ran counter to the results of previous similar research, where the opposite tendency was observed (Chan *et al*., 2009), hence the interest in potential reanalysis. The results of Gilad and Mizrahi-Man's reanalysis seemed to suggest that this apparent tendency within Lin *et al*.'s dataset was in fact caused, at least in part, by batch effects, those in turn stemming from the study's design. These new findings cast doubt on the original conclusions.

More recently, a somewhat similar dataset to Lin *et al*.'s was released by Söllner *et al*. (2017). Like Lin *et al*.'s, this dataset is a comparative RNA-Seq dataset of gene expression in various tissues, though Söllner *et al*.'s dataset compiles mouse and rat (*Rattus norvegicus*) samples, rather than mouse and human ones. Also like Lin *et al*.'s, this dataset seems to exhibit *intra*-over-*inter* clustering. As mentioned, this tendency is unexpected, even more so with the species in question being related more closely than in Lin *et al*.'s case (Jansa and Weksler, 2004).

The aim of this research project is to reanalyze Söllner *et al*.'s dataset towards batch effects, like Gilad and Mizrahi-Man did Lin *et al*.'s dataset, in an attempt to determine whether the *intra*-over-*inter* clustering exhibited by said dataset is the result of batch effects.

# 2. Methods

All of the following analyses were carried out using R version 4.3.1 (R Core Team, 2023). The R code used to carry out the described analyses is freely available on GitHub in the form of an rmarkdown file, under `https://github.com/macbiel/cbs/tree/main/final_project`. The data for the analyses is retrieved remotely from the original sources and therefore not included in the repository.

## 2.1. Original study design

In total, there were 39 mouse samples and 40 rat samples, with each combination of tissue and species having three separate associated samples. This is in contrast to Lin *et al*.'s dataset only having 1 sample for each combination.

As part of their published dataset, Söllner *et al*. have provided metadata pertaining to the design of their study. Based on this metadata (Simon, 2017, files '`rat_design.txt`' and '`mouse_design.txt`'), individual samples were assigned to separate experimental batches, as shown in Table 1. Söllner *et al*.'s dataset contained an additional rat sample without a proper tissue annotation (labeled as "Unknown"). This sample was included in the analyses, out of curiosity whether it would be possible to determine what tissue said sample was of.

| Batch number | Samples present (**Bold**: rat samples, *italics*: mouse samples) | | | | | |
|---|---|---|---|---|---|---|
| 1. | **Kidneys** | **Quadriceps** | | | | |
| 2. | **Brain** | **Esophagus** | **Heart** | **Thymus** | | |
| 3. | *Colon*<br>*Pancreas* | *Duodenum*<br>*Stomach* | *Ileum* | *Jejunum* | *Kidney* | *Liver* |
| 4. | *Brain*<br>*Stomach*<br>**Kidneys** | *Esophagus*<br>*Thymus*<br>**Liver** | *Heart*<br>**Colon**<br>**Quadriceps** | *Liver*<br>**Duodenum**<br>**Stomach** | *Pancreas*<br>**Ileum** | *Quadriceps*<br>**Jejunum** |
| 5. | *Colon*<br>*Quadriceps*<br>**Liver** | *Duodenum*<br>*Thymus*<br>**Pancreas** | *Heart*<br>**Brain**<br>**Stomach** | *Ileum*<br>**Duodenum**<br>**Thymus** | *Jejunum*<br>**Esophagus** | *Kidney*<br>**Heart** |
| 6. | *Brain*<br>*Pancreas*<br>**Jejunum** | *Duodenum*<br>*Stomach*<br>**Kidneys** | *Esophagus*<br>**Colon**<br>**Quadriceps** | *Ileum*<br>**Esophagus**<br>**Thymus** | *Jejunum*<br>**Heart** | *Liver*<br>**Ileum** |
| 7. | *Brain*<br>*Thymus*<br>**Liver** | *Colon*<br>**Brain**<br>**Pancreas** | *Esophagus*<br>**Colon**<br>**Pancreas** | *Heart*<br>**Duodenum**<br>**Stomach** | *Kidney*<br>**Ileum**<br>**Unknown** | *Quadriceps*<br>**Jejunum** |

Table 1.: Correspondence between samples and batches.

A potential flaw in Lin *et al*.'s study design was that the sample-to-batch assignment was largely confounded with the samples' species annotation. This is less so the case in Söllner *et al*.'s dataset, with only batches 1-3 containing samples of a single species each. The tissue annotations are not batch-confounded either, with samples of each tissue being present in several batches.

Another point of difference between the two datasets is the distribution of tissue types. While Lin *et al*.'s dataset contained samples from every organ system sans the skeletal systems (for human and mouse samples both), Söllner *et al*.'s is largely dominated by gastrointestinal (GI) tract tissue and altogether lacks samples from the respiratory, integumentary, skeletal, endocrine and reproductive systems.

## 2.2. Preprocessing of dataset

### 2.2.1. Compilation of homologue pairs

Söllner *et al*.'s dataset was released as two separate gene expression matrices, each matrix containing each species' samples. However, the analyses to be carried out on the dataset require the samples compiled into a single matrix. To unify the matrices, mouse and rat genes present in the dataset had to be paired into homologue pairs.

Söllner *et al*. provided a short list of mouse-rat homologue pairs (Supplementary S3 in the original publication). This was used as the basis for building a more complete list of homologue pairs, and extended upon based on the following three gene databases:

1. ENSEMBL (Harrison *et al*., 2024):
   data taken from ENSEMBL's Biomart service,
   using the `biomaRt` Bioconductor package (Durinck *et al*., 2009);
2. Mouse Genome Informatics ("MGI"; Baldarelli *et al*., 2024):
   data extracted from MGI's downloadable "Homology Classes" dataset;
3. HUGO Gene Nomenclature Committee ("HGNC"; Seal *et al*., 2023):
   data extracted from HGNC's downloadable complete dataset.

All four lists were combined, and deduplicated such that each gene appears in the list at most once. This deduplication was carried out sequentially, with entries earlier in the list taking precedence in the finalized list over the later entries. As the individual lists were concatenated in the order described above (Söllner *et al*.'s list as the first, HGNC as the last), this meant that homologue pairs from specific sources were prioritized over other sources'. This was done to cope with what was considered to be different levels of uncertainty in the various sources: Söllner *et al*.'s list was given the highest priority, as it was provided alongside the original publication; ENSEMBL and MGI specifically provide data for mouse (and in ENSEMBL's case, also rat) genes, and as such were considered more "trustworthy" than HGNC, which primarily lists human genes and relates them to homologues present in other genomes.

### 2.2.2. Filtering and normalization

Prior to adjusting it for batch effects, the dataset was filtered and normalized. Both of these steps were carried out essentially identically to Gilad and Mizrahi-Man's process. 30% of genes with lowest overall sum expression were removed from the dataset, as were mitochondrial genes. The data was then lane- and depth-normalized, and log-transformed. Normalization was carried out using the `edgeR` and `EDASeq` Bioconductor packages (Chen *et al*., 2024; Risso *et al*., 2011). `EDASeq` was also used to retrieve the GC content of the genes in the dataset.

The starting dataset, as shared by Söllner *et al*., consisted of 47531 mouse and 32662 rat genes. The complete list of homologues comprised of 17998 wholly unique gene pairs, and by extension, that was the number of genes in the combined rat-mouse gene matrix. Filtering removed approx. 30% of the remaining genes, which reduced the gene count in the final dataset to 12578. As such, the final dataset made up only 26% and 39% of the original mouse and rat datasets, respectively.

## 2.3. Adjustment for batch effects

The data was adjusted for the presence of batch effects using the `ComBat` function of the `sva` Bioconductor package (Leek *et al*., 2024). The `ComBat` function takes a data matrix and batch variables as inputs, and outputs a matrix of the same dimensions as the input, with the contained measurements adjusted for batch variables, as based on an empirical Bayesian framework (Leek *et al*., 2012).

## 2.4. Assessing the emergence of *intra*-over-*inter* clustering

To verify whether or not the dataset truly exhibited *intra*-over-*inter* clustering, statistical analysis and visualizations of the dataset were carried out. This was done for data prior to and after the batch effect adjustment.

### 2.4.1. Statistical analysis

For statistical analysis, each sample's tissue and species annotations were one-hot encoded, and a linear model was fitted to the data, using the tissue and species annotations as predictors, the gene expression values as the dependent variable, and gene IDs as grouping factor. The model was fitted using the `lme4` R package (Bates *et al*., 2015). The coefficients of the model were then compared to see which annotation was a more meaningful predictor.

### 2.4.2. Visualizations

Two visualization methods were utilized: a scatter plot of the first 2 primary components (PCs) of the dataset's singular value decomposition (SVD), and a sample-pairwise correlation heatmap, with hierarchical clustering of the samples. The two types of plots were produced using the the `ggplot2` and `pheatmap` R packages, respectively (Wickham, 2016; Kolde, 2019).

# 3. Results

All three employed methods of assessing sample clustering appear to confirm that the data exhibits *intra*-over-*inter* clustering before and after the batch effect adjustment.

The coefficients of the fitted linear model were significantly higher for the predictors corresponding to the samples' species annotations than for the tissue annotations (Table 2), which implies that species was an overall stronger predictor of gene expression than tissue. This was the case for data both before and after adjustment.

Also notable is that the adjustment appears to have induced a reduction in the values of species annotation coefficients, and conversely, an increase in the tissue annotation coefficients. The model-fitting algorithm dropped the "Unknown" tissue annotation predictor due to insufficient data.

| Predictor | | Model coefficients | |
|---|---|---|---|
| | | Before adjustment | After adjustment |
| *Species* | Mouse | 10.196 | 9.826 |
| | Rat | 9.909 | 9.556 |
| *Tissue* | Pancreas | -2.049 | -1.649 |
| | Liver | -1.820 | -1.465 |
| | Stomach | -1.589 | -1.223 |
| | Duodenum | -1.629 | -1.267 |
| | Jejunum | -1.587 | -1.219 |
| | Ileum | -1.575 | -1.211 |
| | Colon | -1.601 | -1.249 |
| | Kidney | -1.619 | -1.282 |
| | Quadriceps | -1.806 | -1.413 |
| | Thymus | -1.494 | -1.134 |
| | Heart | -1.899 | -1.535 |
| | Esophagus | -1.509 | -1.141 |
| | Brain | -1.437 | -1.099 |
| | Kidneys | -1.704 | -1.251 |

Table 2.: Coefficients of linear models fitted to the data before and after adjustment.

The heatmaps (Figure 2) depict rat samples clustering entirely separately from mouse samples, both in the case of the data before and after adjustment. *Intra*-over-*inter* clustering was invariant of the clustering method employed.

As an exception to i*ntra*-over-*inter* clustering, the brain samples of both species cluster wholly separately from all other samples. This isn't consistent across different clustering methods, however.

The batch effect adjustment seems to have slightly reduced the disparity between individual samples, though only to a small extent. Within each species, the samples cluster according to tissue generally as expected, e.g. intestinal tissue samples cluster together, and this per-tissue within-species clustering is slightly more consistent after adjustment. The rat sample with the "Unknown" tissue annotation clusters next to "Pancreas" samples.

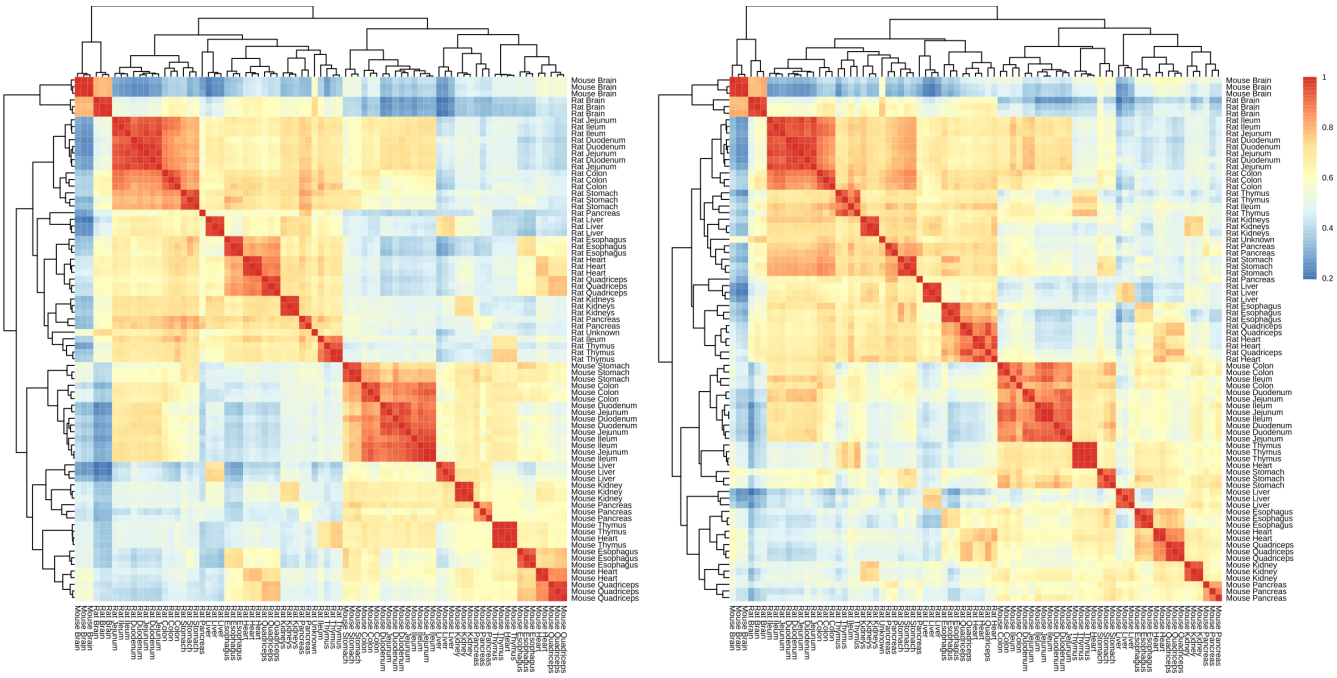

Figure 2.: Sample-pairwise correlation heatmaps, with hierarchical clustering of the samples, of data before (left) and after (right) adjustment. Pearson correlation coefficient, Euclidean distance and complete linkage were used as measure of correlation, measure of distance and clustering method respectively.

The first and second PCs of the SVD of the data closely align with the samples' species and tissue annotation respectively. This is the case both for the data before and after adjustment, though this is less so the case after adjustment, especially for first PC. The "Unknown" sample does not appear to cluster together with any specific sample on the SVD plots.

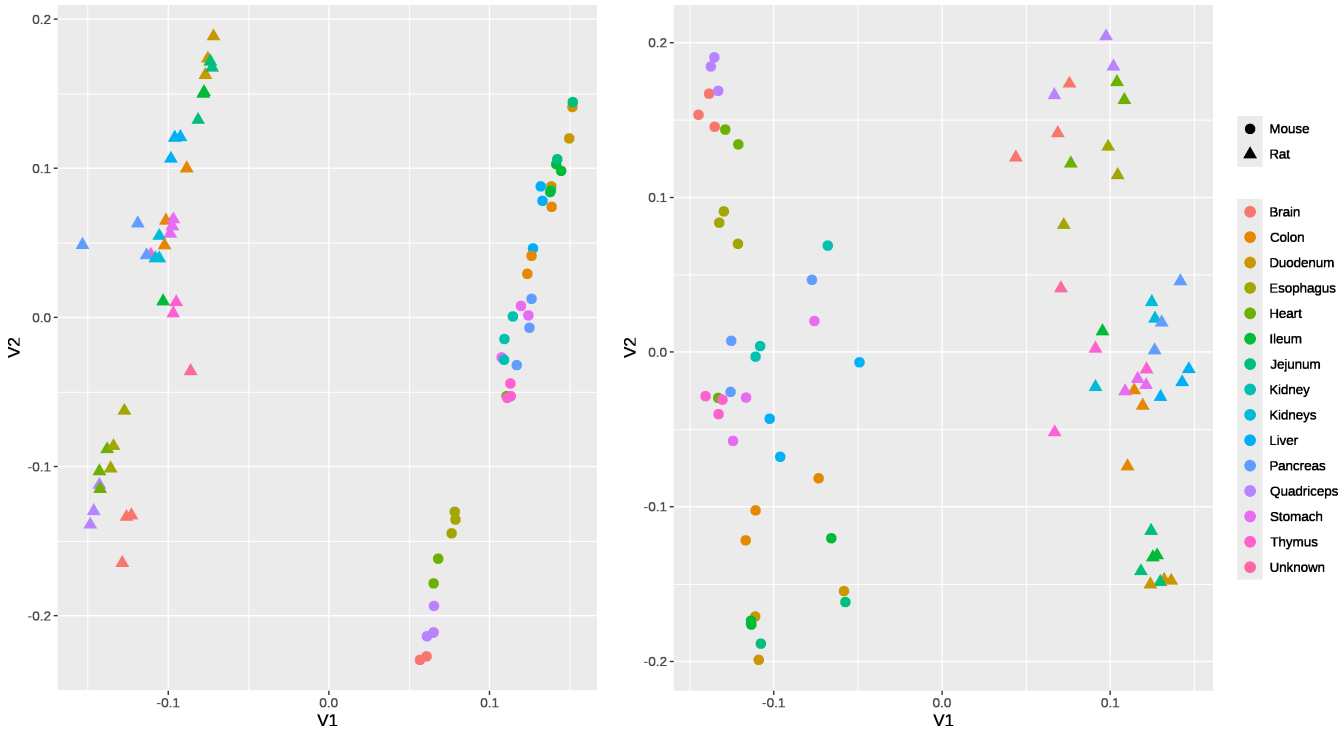


Figure 3.: Plots of the first two principal components of the singular value decomposition of the data before (left) and after (right) adjustment.

# 4. Discussion

Overall, the data was shown to exhibit *intra*-over-*inter* clustering which persists through the batch effect adjustment, which should sufficiently disprove that this mode of clustering was an artifact caused by batch effects. Whether the *intra*-over-*inter* clustering of the samples is an artifact caused by some other flaw in the study design, or a genuine feature of the data, however, remains inconclusive. Compared to Lin *et al*., Söllner *et al*.'s study design has the merit of including multiple samples for each combination of species and tissue, which lends credence to the possibility that the mode of clustering assessed here is genuine, but on the other hand, has lower variety in the types of tissues studied, which might have negatively impacted the study's results specifically with regards to the comparison of individual samples.

Söllner *et al*. call out a potential flaw in the study design in their publication (section "Expression variability across tissues" therein), namely that since samples were taken from complex organs, the specific gene expression patterns of the individual cell types present in the sample may mask the more general gene expression of the tissue/organ as a whole. Pancreatic beta cells, which have very high insulin expression while at the same time being the only type of cell to actually express insulin, are cited as a potential example. Such an effect would not be very likely to have caused *intra*-over-*inter* clustering encompassing the entire dataset, however, and would sooner cause within-species per-tissue clustering of samples to differ between species, which is not the case in this dataset.

Ultimately, the "Unknown" rat sample's missing tissue annotation could not have been determined. The heatmaps suggest that it might have been a pancreatic sample, but this was deemed as insufficient proof, given that the other results do not corroborate this being the case.

Broadly speaking, the batch effect adjustment appears to have weakened the samples' species and tissue annotation as predictors of the gene expression values, as evidenced by the change in the fitted model's coefficients and the comparatively less concrete clustering of samples in the SVD plot. The exact cause behind this is not presently understood.

The separate clustering of brain samples is also explicitly acknowledged by Söllner *et al*. in their publication, though no suitable reason for this is provided. This research project has failed to make this clearer.

# 5. References


1. Baldarelli R. M., Smith C. L., Ringwald M. *et al*.: Mouse Genome Informatics: an integrated knowledgebase system for the laboratory mouse. 2024, *Genetics*, **227**(1).
DOI: 10.1093/genetics/iyae031

2. Bates D., Mächler M., Bolker B. *et al*.: Fitting Linear Mixed-Effects Models Using lme4. 2015, *Journal of Statistical Software*, **67**(1), p. 1-48.
DOI: 10.18637/jss.v067.i01

3. Chan E. T., Quon G. T., Chua G. *et al*.: Conservation of core gene expression in vertebrate tissues. 2009, *Journal of Biology*, **8**(33).
DOI: 10.1186/jbiol130

4. Chen Y., Chen L., Lun A. T. L. *et al*.: edgeR 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. 2024, *bioRxiv*.
DOI: 10.1101/2024.01.21.576131

5. Durinck S., Spellman P., Birney E. *et al*.: Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. 2009, *Nature Protocols*, **4**, p. 1184–1191.
DOI: 10.18129/B9.bioc.biomaRt

6. Gilad Y. and Mizrahi-Man O.: A reanalysis of mouse ENCODE comparative gene expression data. 2015, *F1000Research*, **4**(121).
DOI: 10.12688/f1000research.6536.1

7. Goh W. W. B., Wang W. and Wong L.: Why Batch Effects Matter in Omics Data, and How to Avoid Them. 2017, *Trends in Biotechnology*, **35**(6), p. 498-507.
DOI: 10.1016/j.tibtech.2017.02.012

8. Harrison P. W., Amode M. R., Austine-Orimoloye O. *et al*.: Ensembl 2024. 2024, *Nucleic Acids Research*, **52**(D1), p. D891-D899.
DOI: 10.1093/nar/gkad1049

9. Jansa S. A. and Weksler M.: Phylogeny of muroid rodents: relationships within and among major lineages as determined by IRBP gene sequences. 2004, *Molecular Phylogenetics and Evolution*, **31**(1), p. 256-276.
DOI: 10.1016/j.ympev.2003.07.002

10. Kolde R.: pheatmap: Pretty Heatmaps. 2019.
DOI: 10.32614/CRAN.package.pheatmap

11. Leek J. T., Johnson W. E., Parker H. S. *et al*.: The sva package for removing batch effects and other unwanted variation in high-throughput experiments. 2012, *Bioinformatics,* **28**(6), p. 882-883.
DOI: 10.1093/bioinformatics/bts034

12. Leek J. T., Johnson W. E., Parker H. S. *et al*.: sva: Surrogate Variable Analysis. 2024, R package version 3.52.0.
DOI: 10.18129/B9.bioc.sva

13. Lin S., Lin Y., Nery J. R. *et al*.: Comparison of the transcriptional landscapes between human and mouse tissues. 2014, *Proceedings of the National Academy of Sciences of the United States of America,* **111**(48), p. 17224-17229.
DOI: 10.1073/pnas.1413624111

14. R Core Team: R: A Language and Environment for Statistical Computing. 2020.
https://www.R-project.org

15. Risso D., Schwartz K., Sherlock G. *et al*.: GC-content normalization for RNA-Seq data. 2011, *BMC Bioinformatics*. **12(**480).
DOI: 10.1186/1471-2105-12-480

16. Seal R. L., Braschi B., Gray K. *et al*.: Genenames.org: the HGNC resources in 2023. 2023, *Nucleic Acids Research*, **51**(D1), p. D1003-D1009.
DOI: 10.1093/nar/gkac888

17. Simon E.: An RNASeq normal tissue atlas for mouse and rat. 2017, *BioStudies*.
ArrayExpress accession: E-MTAB-6081

18. Söllner J., Leparc G., Hildebrandt T. *et al*.: An RNA-Seq atlas of gene expression in mouse and rat normal tissues. 2017, *Scientific Data*, **4**(170185).
DOI: 10.1038/sdata.2017.185

19. Wickham H.: ggplot2: Elegant Graphics for Data Analysis. 2016, Springer-Verlag New York.
DOI: 10.32614/CRAN.package.ggplot2