Journal Club

Deepak Tanwar

12 th October, 2018

Table of Contents

Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing

Keegan Korthauer, Sutirtha Chakraborty, Yuval Benjamini, Rafael A Irizarry

Biostatistics, kxy007, https://doi.org/10.1093/biostatistics/kxy007

Published: 22 February 2018

Applications of WGBS

- Detection of DMRs
- Detection of DMLs
- Detection of CNVs
- Detection of SNPs
- Detection of Cytosine methylation levels of TFBSs

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depends on large sample sizes.

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depends on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome have ~30 million CpGs [1].

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depends on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome have ~30 million CpGs [1].
- Measurements are spatially correlated across the genome [2] (which is not accounted in most of the methods). Measurements from all loci are treated independently.

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depends on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome have ~30 million CpGs [1].
- Measurements are spatially correlated across the genome [2] (which is not accounted in most of the methods). Measurements from all loci are treated independently.
- Muliple testing corrections without considering the spatial correlation can result in a loss of power.

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depends on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome have ~30 million CpGs [1].
- Measurements are spatially correlated across the genome [2] (which is not accounted in most of the methods). Measurements from all loci are treated independently.
- Muliple testing corrections without considering the spatial correlation can result in a loss of power.
- **Biological interpretability:** most individual CpG loci likely do not have a large impact on epigenetic function on their own, but rather through a biochemical modification that involves several loci. Regional DNA methylation levels are correlated with the expression levels of nearby genes. Differentially expressed genes were consistently more likely to be located near DMRs than DMLs [3].

....cont.

- DML approaches may construct DMRs by chaining together neighboring significant loci, this type of approach will not yield a proper assessment of the statistical significance of the constructed regions, nor will the FDR be properly controlled [4]. Controlling the FDR at the level of individual loci is not the same as controlling FDR of regions.

....cont.

- DML approaches may construct DMRs by chaining together neighboring significant loci, this type of approach will not yield a proper assessment of the statistical significance of the constructed regions, nor will the FDR be properly controlled [4]. Controlling the FDR at the level of individual loci is not the same as controlling FDR of regions.
- FDR correction at the level of individual loci means that the proportion of expected false positive (FP) loci is controlled, not the proportion of FP regions. Statistically, this is a critical point since FDR control of DMR detection is not guaranteed under the DML setting.

....cont.

- DML approaches may construct DMRs by chaining together neighboring significant loci, this type of approach will not yield a proper assessment of the statistical significance of the constructed regions, nor will the FDR be properly controlled [4]. Controlling the FDR at the level of individual loci is not the same as controlling FDR of regions.
- FDR correction at the level of individual loci means that the proportion of expected false positive (FP) loci is controlled, not the proportion of FP regions. Statistically, this is a critical point since FDR control of DMR detection is not guaranteed under the DML setting.
- To perform inference at the region level, many tools: perform analysis on predefined regions of interest or fixed sliding windows. Though useful in targeted settings such as reduced representation bisulfite sequencing (RRBS), or when we have prior knowledge of the DMR size, they are not applicable to identifying DMRs of arbitrary size from WGBS.

....cont.

- DML approaches may construct DMRs by chaining together neighboring significant loci, this type of approach will not yield a proper assessment of the statistical significance of the constructed regions, nor will the FDR be properly controlled [4]. Controlling the FDR at the level of individual loci is not the same as controlling FDR of regions.
- FDR correction at the level of individual loci means that the proportion of expected false positive (FP) loci is controlled, not the proportion of FP regions. Statistically, this is a critical point since FDR control of DMR detection is not guaranteed under the DML setting.
- To perform inference at the region level, many tools: perform analysis on predefined regions of interest or fixed sliding windows. Though useful in targeted settings such as reduced representation bisulfite sequencing (RRBS), or when we have prior knowledge of the DMR size, they are not applicable to identifying DMRs of arbitrary size from WGBS.
- Challenges for assessing DMRs:
    - Region boundaries
    - methods ignore correlation across loci
    - biological variability from sample to sample

Methods implementation challenges

Methods for identifying DMLs also need to properly account for the statistical properties of count data that do not conform to standard **Gaussian models**. This is in contrast to methylation array analysis, where Gaussian models performed well (Jaffe and others, 2012). One option is to assume that methylation proportions, defined as the number of methylated reads divided by the number of total reads covering a given CpG locus, follow a normal distribution (Hansen and others, 2012). However this assumption clearly does not hold when the total reads covering the CpG, referred to as the coverage, is small, a common occurrence in these data sets. The approach also ignores that variance of this proportion depends on the coverage. To overcome these limitations, DML approaches have also modeled WGBS count data using **Binomial models** (Saito and others, 2014). However, Binomial models on their own cannot account for biological variability within sample groups. In order to account for biological variability in count data, **Beta-Binomial models** (Park and others, 2014; Sun and others, 2014) are a natural extension. However, they come at the cost of increased computational burden when testing millions of loci.

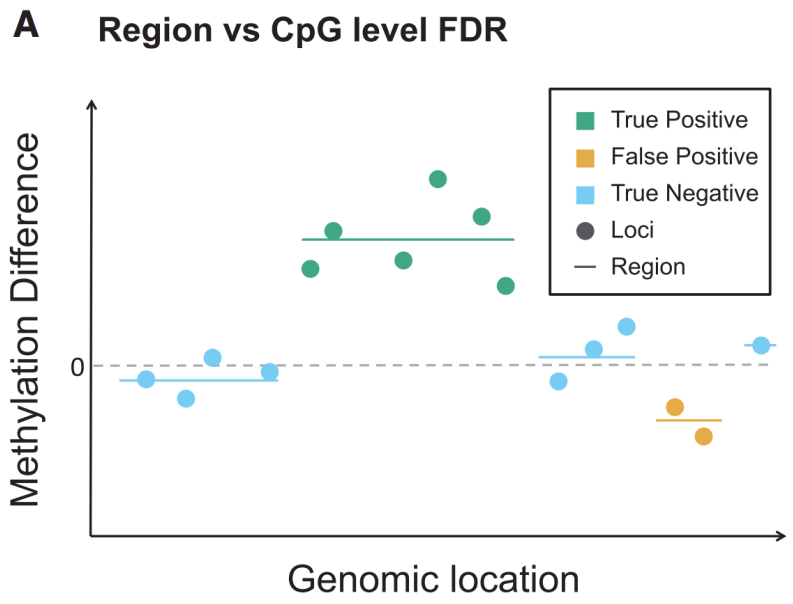
dmrseq

Aim:

1. To **maximize power** while **controlling the FDR** even with sample sizes as small as **two samples per condition**.

1. To develop a procedure to control FDR at the region level and provide an accurate measure of statistical significance for each region.

FDR at the loci level is not the same as FDR at the region level



Genomic location versus methylation difference estimates at several neighboring loci.

The individual CpGs (points) are shaded by whether they are a true or false positive. Regions are denoted by lines.

The loci FDR is:

FDRloci = (#False Positive Loci) / (Total # of Significant Loci)

The region FDR is:

FDRregion=(#False Positive Regions) / (Total # of Significant Regions)

Identification of DMRs by dmrseq

1. **de novo candidate regions:** detects candidate regions and then explicitly evaluates statistical significance at the region level while accounting for known sources of variability

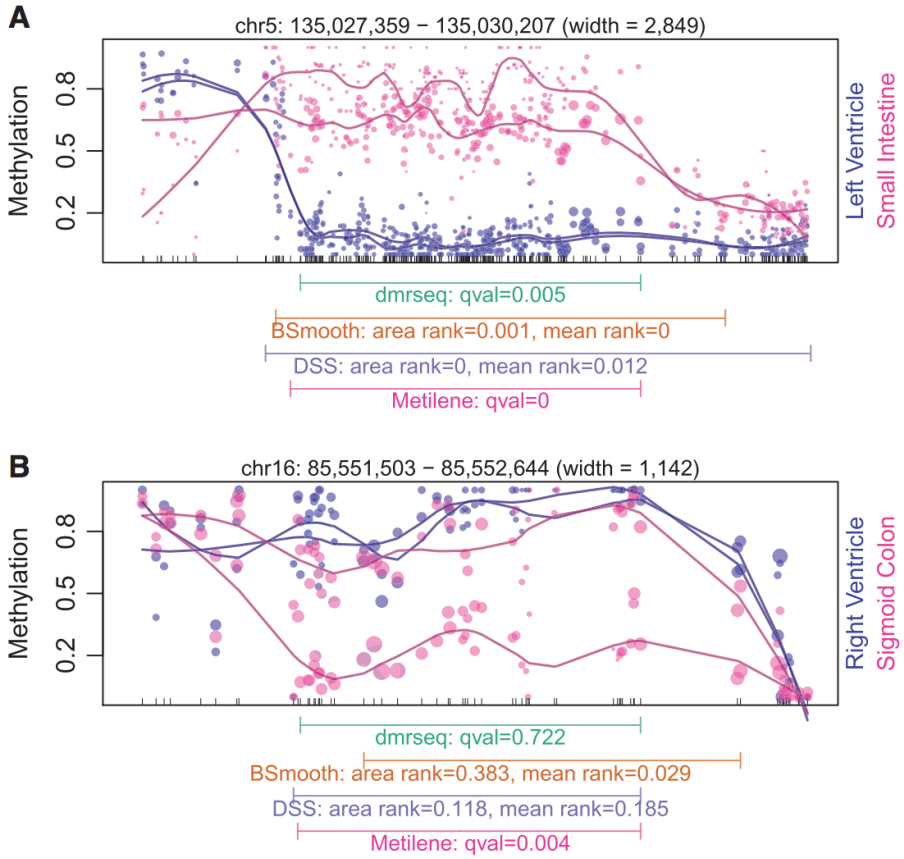
Identification of DMRs by dmrseq

1. **de novo candidate regions:** detects candidate regions and then explicitly evaluates statistical significance at the region level while accounting for known sources of variability

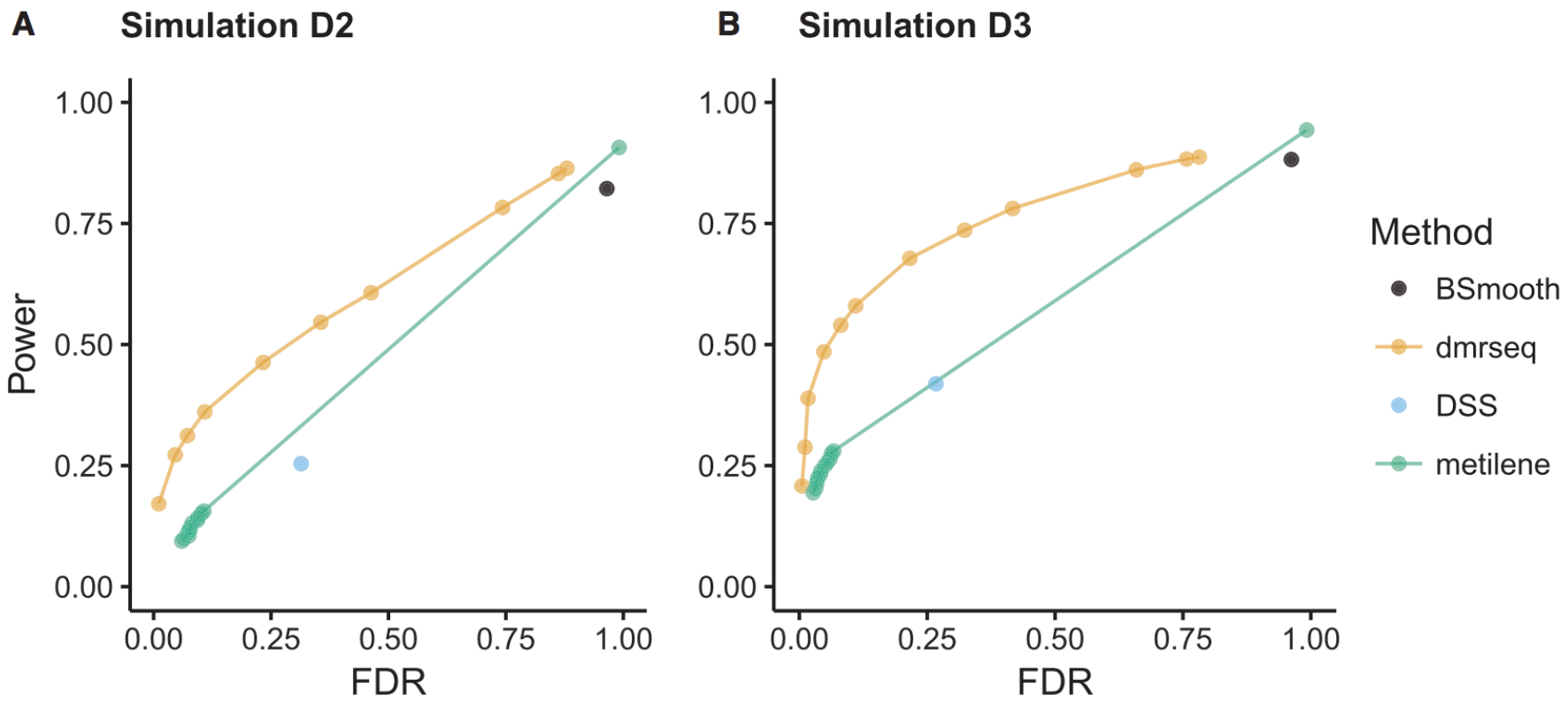
2. compute a statistic for each candidate DMR that takes into account variability between biological replicates and spatial correlation among neighboring loci. Significance of each region is assessed via a permutation procedure.

- Estimation of region statistics with generalized least squares models
- Covariance of methylation levels within regions
- Permutation to generate a null set of regions

dmrseq performs better



dmrseq is more powerful than other methods



Power is calculated as the proportion of simulated DMRs overlapped by at least one identified DMR. FDR is calculated as the proportion of DMRs identified that do not overlap with any of the simulated DMRs.

Summary

- Implementation in the case of 2 samples per group.

Summary

- Implementation in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.

Summary

- Implementation in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.

Summary

- Implementation in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.
- Implementation in **R**.

Summary

- Implementation in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.
- Implementation in **R**.
- Comparion of their method with the "famous" methods.

Summary

- Implementation in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.
- Implementation in **R**.
- Comparion of their method with the "famous" methods.
- Comparison is not done with **RnBeads**. Probably because RnBeads performs differential methylation analysis (only) at single nucleotide level.
- Comparison is not done with **methylKit**. Although methylKit performs differential methylation analysis at regional level.

References

[1] Smith Z. D. and Meissner A. (2013). DNA methylation: roles in mammalian development. Nature Reviews Genetics 14, 204–220.

[2] Leek J. T., Scharpf R. B., Bravo H. C., Simcha D., Langmead B., Johnson W. E., Geman D., Baggerly K. and Irizarry R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11, 733–739.

[3] Aryee M. J., Jaffe A. E., Corrada-Bravo H., Ladd-Acosta C., Feinberg A. P., Hansen K. D. and Irizarry R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369.

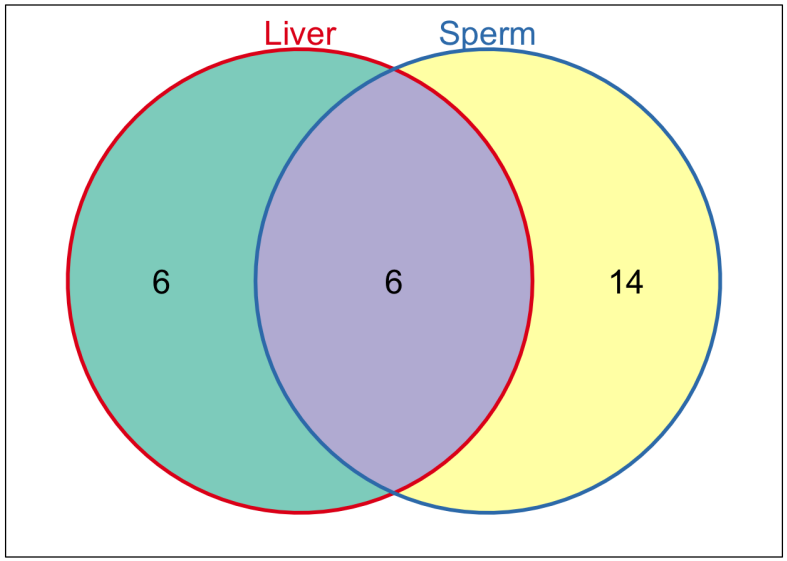
[4] Robinson M. D., Kahraman A., Law C. W., Lindsay H., Nowicka M., Weber L. M. and Zhou X. (2014). Statistical methods for detecting differentially methylated loci and regions. Frontiers in Genetics 5, 324.

Venn diagrams: Good or Bad

	Genes	Liver	Brain	Testis	Sperm	Stomach	Kidney	Lungs	Intestine	Heart
1	a	1	0	0	0	0	0	0	1	0
2	b	1	0	0	0	0	0	0	1	0
3	c	1	0	0	0	1	0	0	1	1
4	d	1	0	0	0	1	0	1	1	1
5	e	1	1	0	0	1	0	1	1	1
6	f	1	1	0	0	1	0	1	1	1

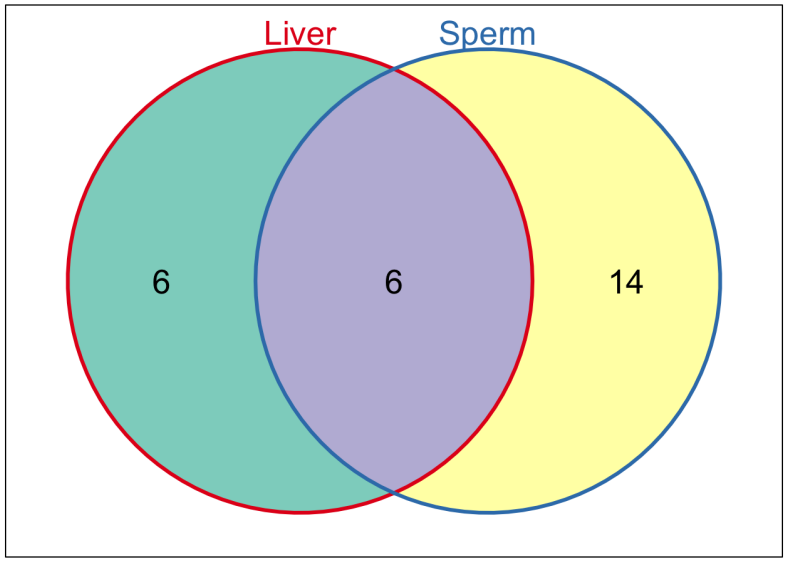
Venn and Euler

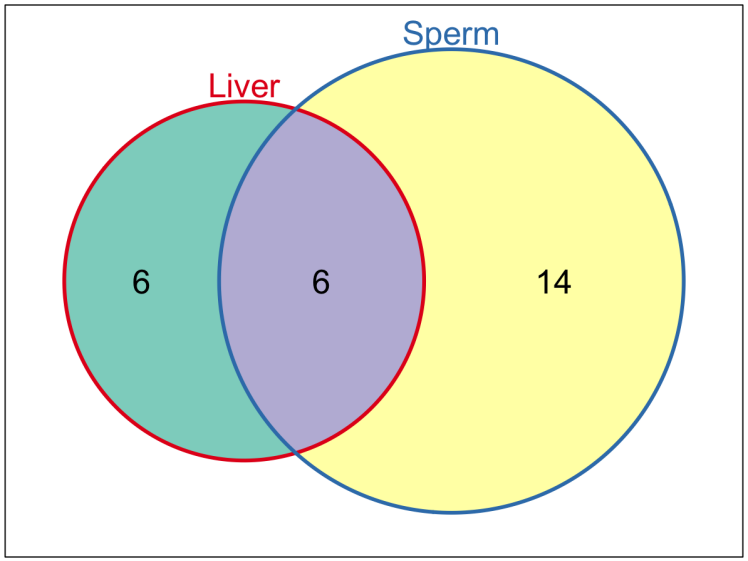
Two Groups



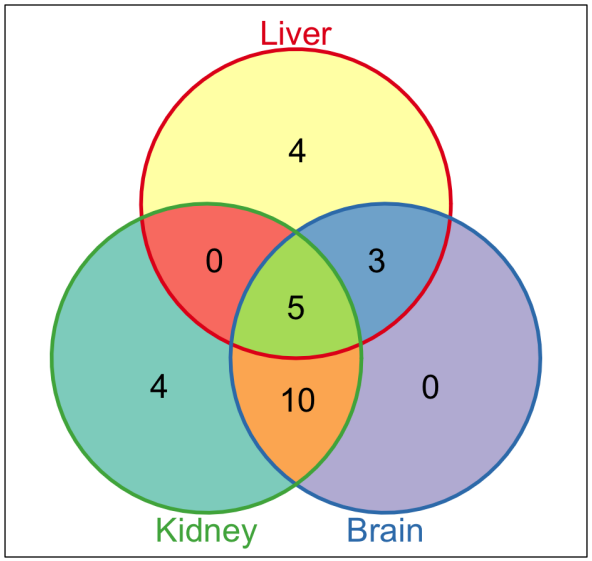
Venn and Euler

Two Groups

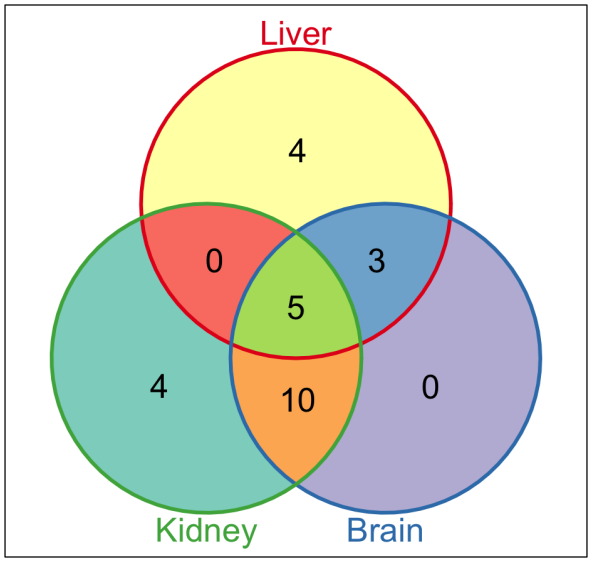


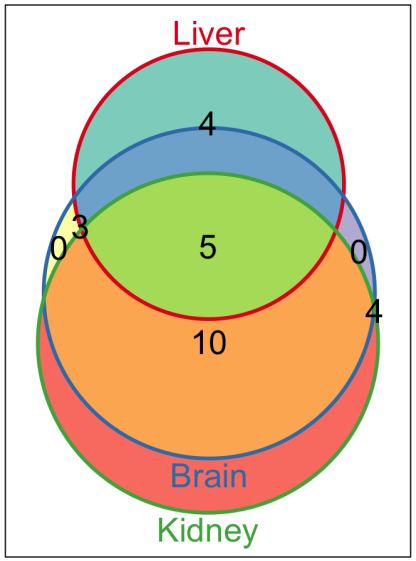


Three Groups

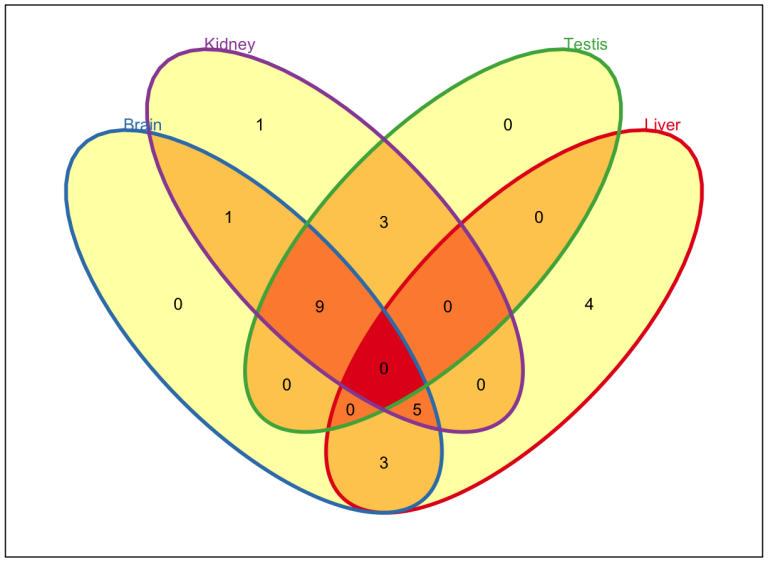


Three Groups

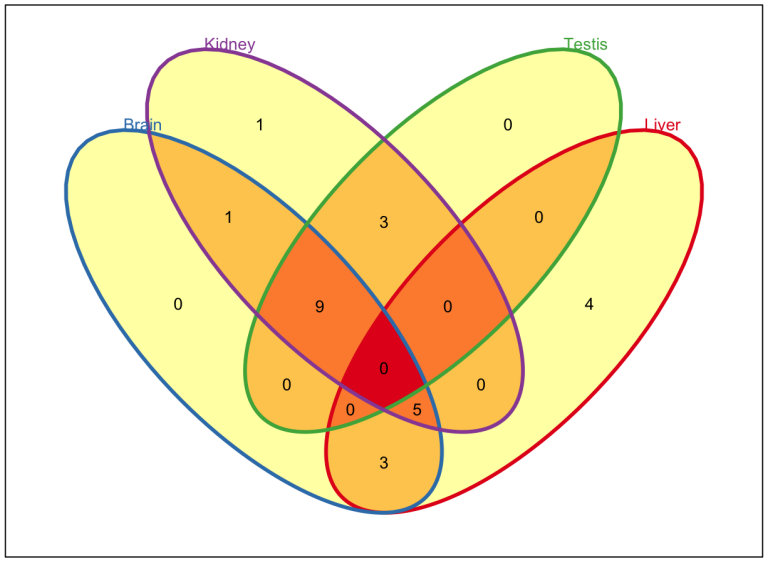


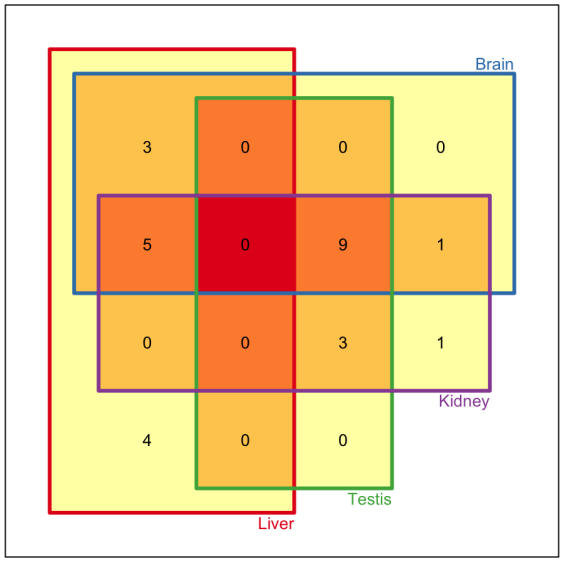


Four Groups

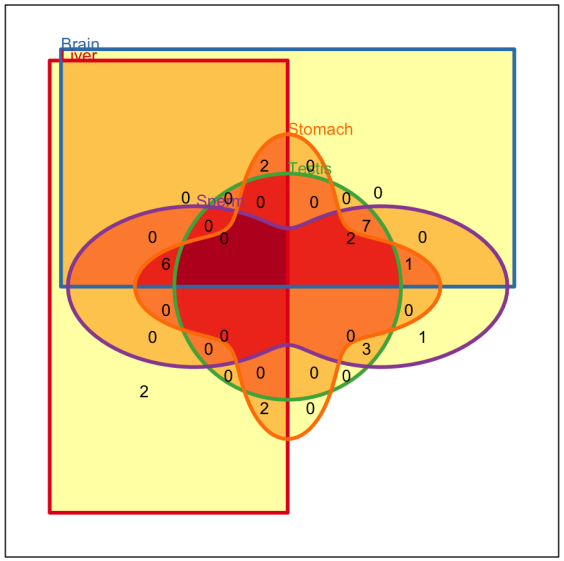


Four Groups

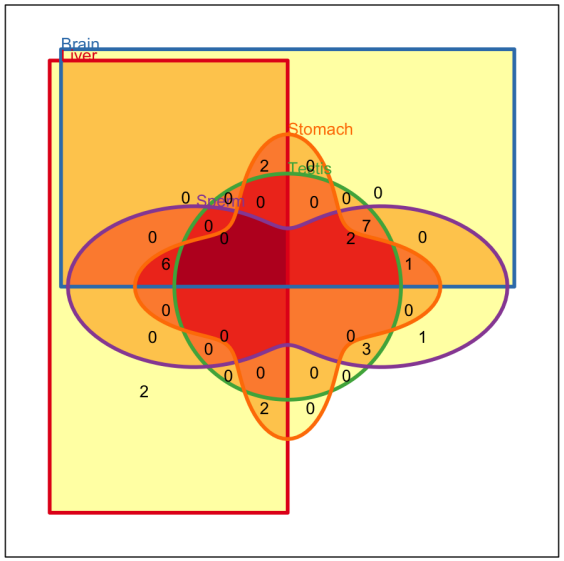

Kidney
Testis
Brain
Liver
1
0
1
3
0
0
9
0
4
0
0
0
0
5
3

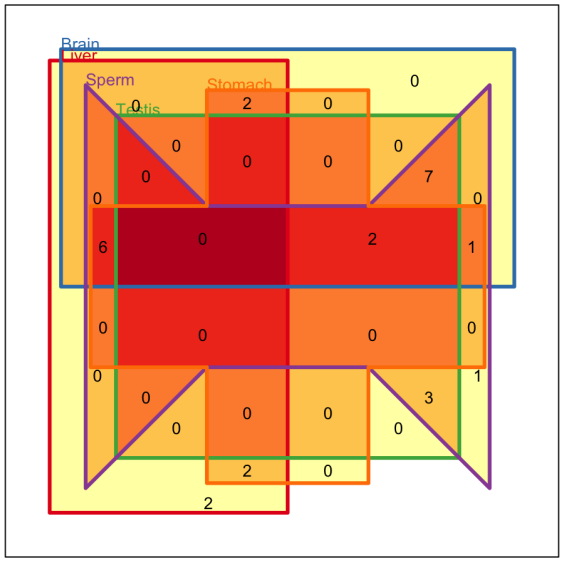

Brain
3
0
0
0
5
0
9
1
0
0
3
1
Kidney
4
0
0
Testis
Liver

Five groups

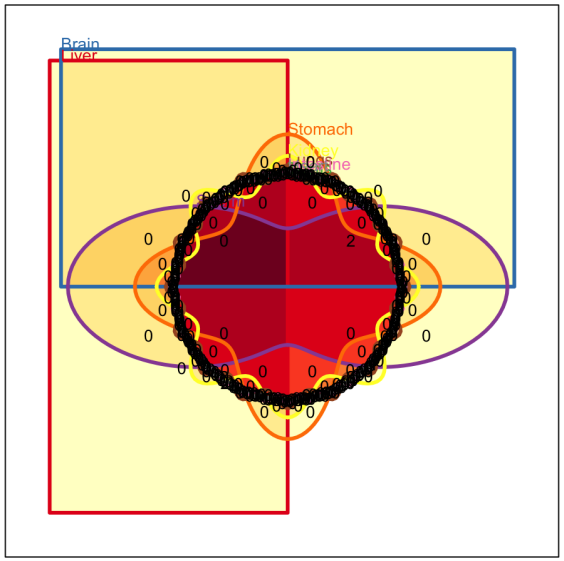

Brain
Liver
Stomach
Testis
Spleen
0
2
0
0
0
0
0
0
0
0
0
0
0
0
7
0
0
2
0
6
1
0
0
0
0
0
0
0
0
3
1
0
0
0
0
0
0
2
2
0

Five groups

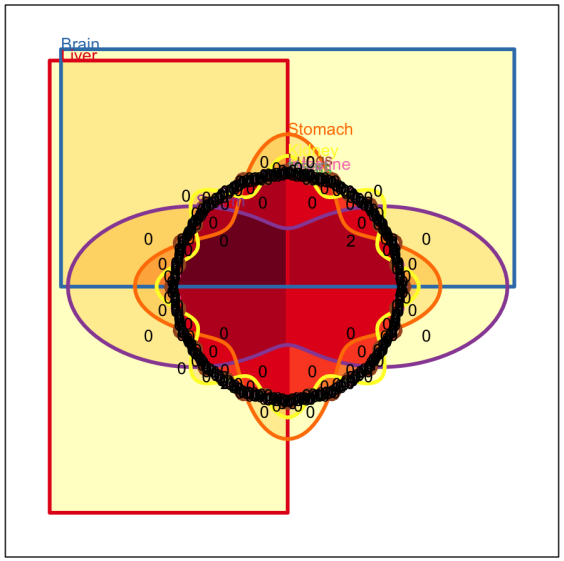

Brain
Liver
Stomach
Testis
Sperm

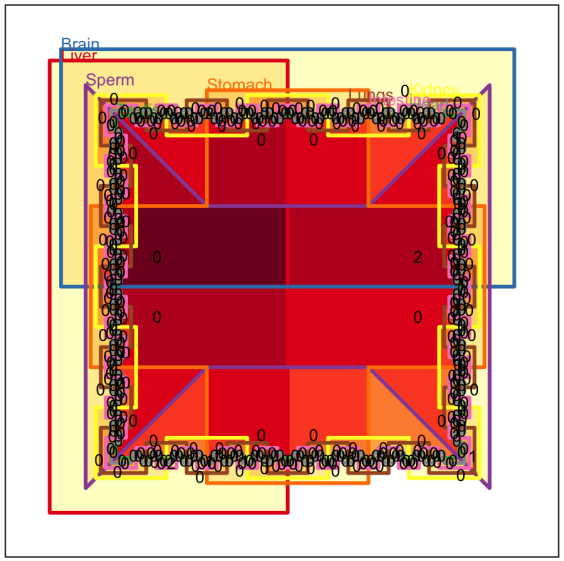

Brain
Liver
Sperm
Stomach
Testis

Nine groups

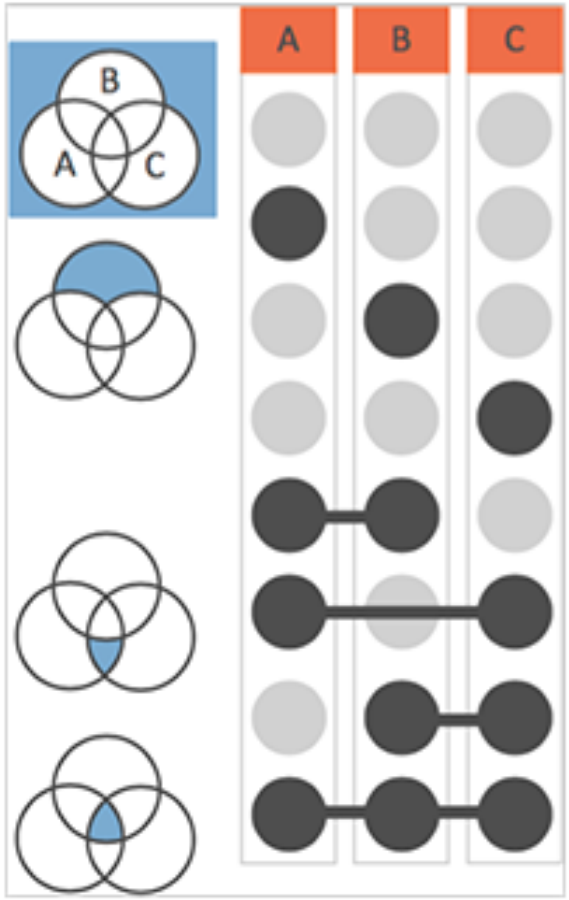


Nine groups

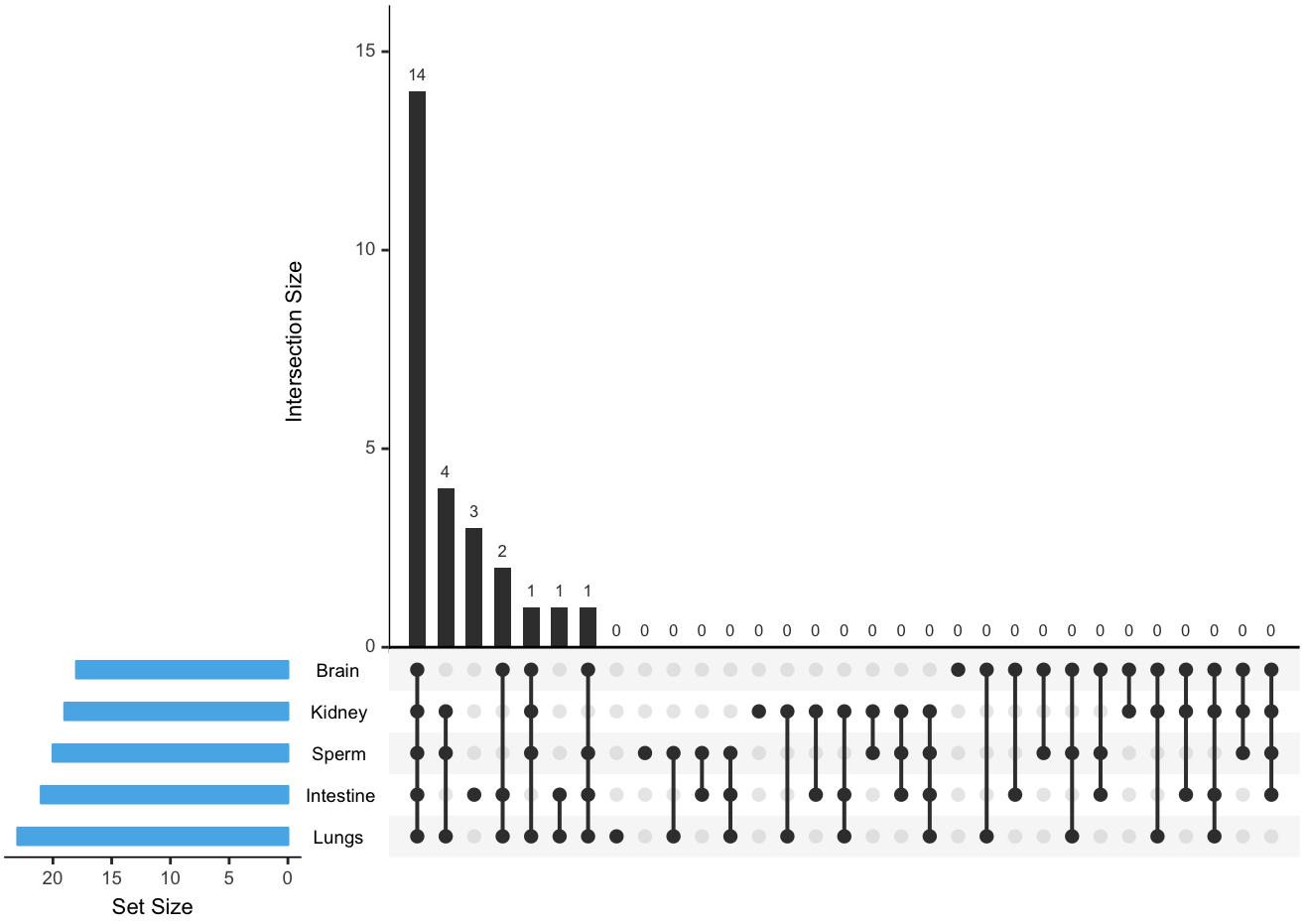

Brain
Liver
Stomach


Brain
Liver
Sperm
Stomach
Lungs
Kidney

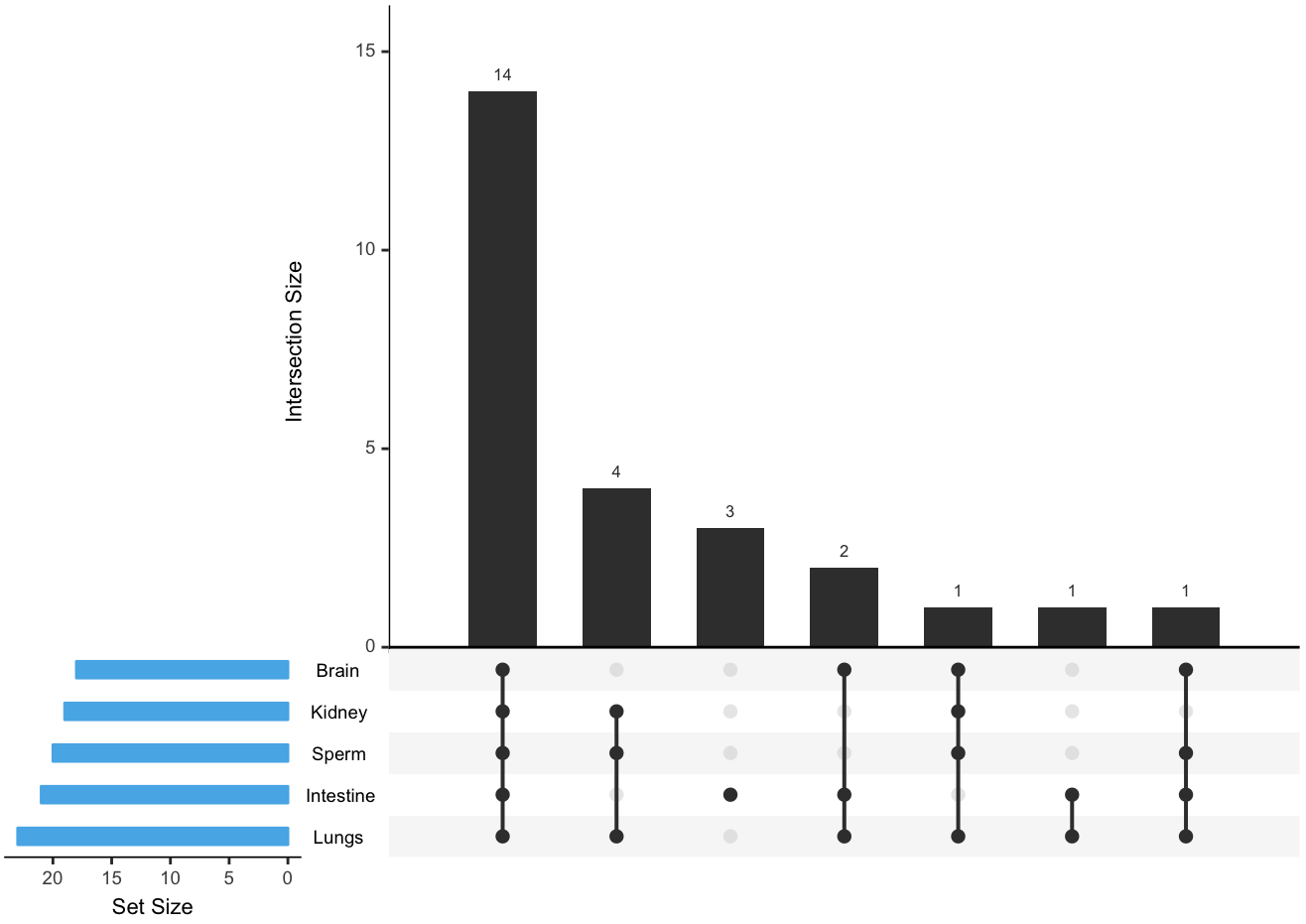
A better way



UpSet plots


Intersection Size
15
10
5
0
14
4
3
2
1
1
1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Brain
Kidney
Sperm
Intestine
Lungs
20
15
10
5
0
Set Size

UpSet plot (no empty)



Sets and intersection

POINTS OF VIEW

Sets and intersections

Complex relationships demand trade-offs.

https://intervene.shinyapps.io/intervene/

Thank you!