Analytical approaches using the Bland-Altman plot

John Stansfield, Mikhail Dozmorov

Bland–Altman plot

- A method of data **plotting** to **analyze agreement** between **two** sets of measures.
- Popularized by J. Martin Bland and Douglas G. Altman after a 1981 short publication by Staffan Eksborg.
- Also known as "Tukey mean-difference plot."

Altman DG, Bland JM (1983). "Measurement in medicine: the analysis of method comparison studies." The Statistician. 32: 307–317. doi:10.2307/2987937.

Bland JM, Altman DG (1986). "Statistical methods for assessing agreement between two methods of clinical measurement." Lancet. 327 (8476): 307–10. doi:10.1016/S0140-6736(86)90837-8. - Cited >42,000 times

Measure of agreement

- How to assess the degree of agreement between two measures?
- Case scenario:
  - Temperature measures over 15 days by a mercury-in-glass and an alcohol thermometer.
  - How similar are measures from the two thermometers?

```
       d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15
Therm1  1  5 10 20 50 40 50 60 70  80  90 100 150 200 250
Therm2  8 16 31 25 38 55 41 69 73  63 123  81 182 258 276
```

Measure of agreement

- How to assess the degree of agreement between two measures?
- Case scenario:
    - Expression of 15 genes is measured by two types of microarrays, Affymetrix and Illumina.
    - How similar are measures from the two arrays?

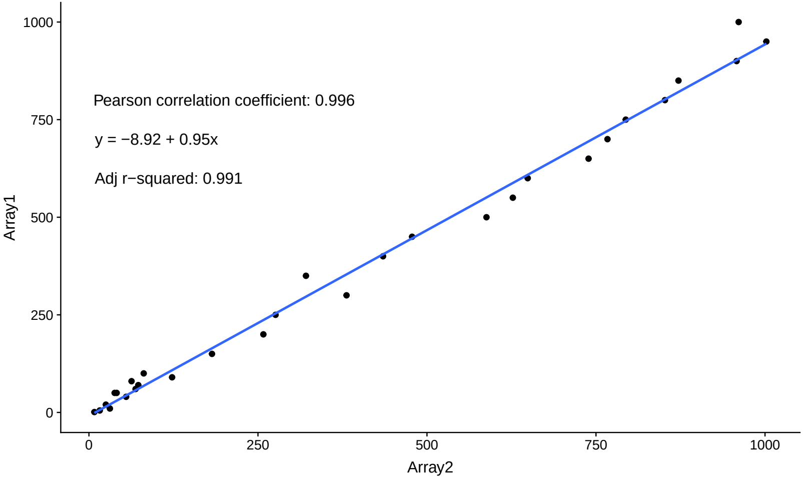
```
       g1 g2 g3 g4 g5 g6 g7 g8 g9 g10 g11 g12 g13 g14 g15
Array1  1  5 10 20 50 40 50 60 70  80  90 100 150 200 250
Array2  8 16 31 25 38 55 41 69 73  63 123  81 182 258 276
```

Measure of agreement

- How to assess the degree of agreement between two measures?
    - **Pearson or product-moment correlation** - whether, and how strongly, pairs of variables are related. r - the ratio of covariance between the variables to the product of their standard deviations.
    - **Linear regression** - finds the best line that predicts one variable from the other. r^2 - the coefficient of determination, the proportion of variance that the two variables have in common.
- The correlation coefficient and regression technique can be misleading when assessing agreement because they evaluate only the linear association of two sets of observations.

Udovicic M, Bazdaric K, Bilic-Zulle L, Petrovecki M. What we need to know when calculating the coefficient of correlation? Biochem Med (Zagreb) 2007;17:10-5. http://dx.doi.org/10.11613/BM.2007.002.

Measure of agreement



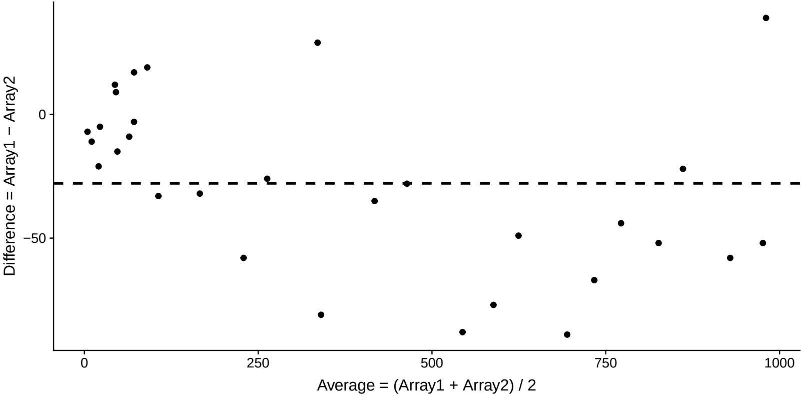
Bland-Altman plot for measuring agreement

- Plot every **difference** between two pairs of the measurements against the **average** of the measurement.

```
    Array1 Array2 Difference Average
g1       1      8         -7     4.5
g2       5     16        -11    10.5
g3      10     31        -21    20.5
g4      20     25         -5    22.5
g5      50     38         12    44.0
g6      40     55        -15    47.5
g7      50     41          9    45.5
g8      60     69         -9    64.5
g9      70     73         -3    71.5
g10     80     63         17    71.5
```

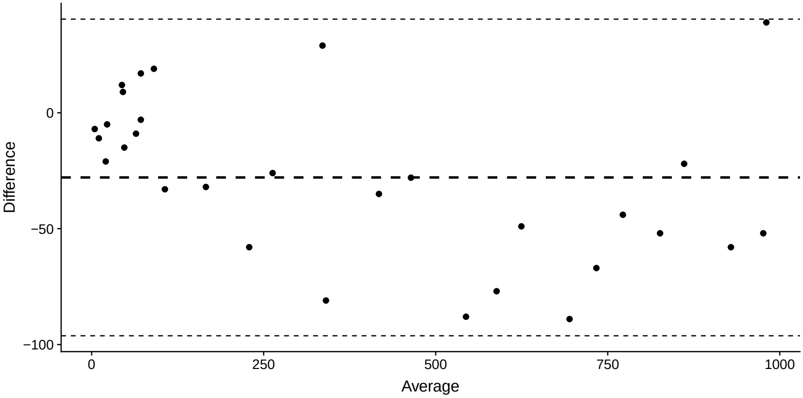
Bland-Altman plot for measuring agreement

- Allows investigation of the relationship between measurement error and the true value (assumed to be the mean).



Bland-Altman plot for measuring agreement

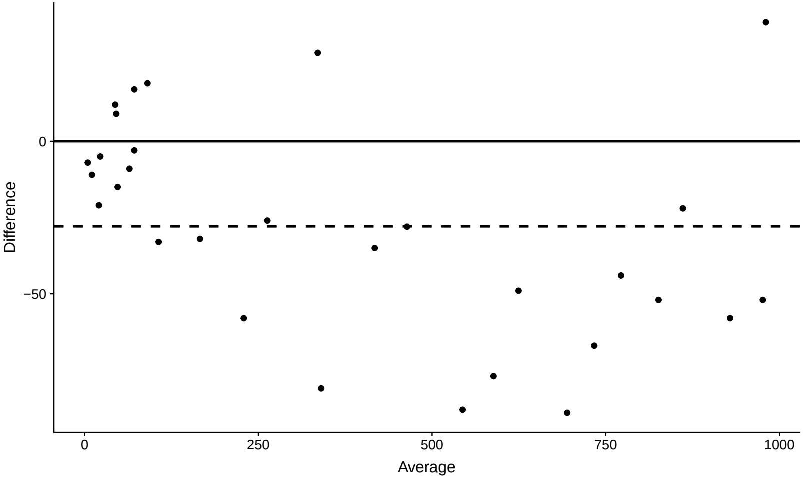
- Measure agreement as the proportion of the differences within *a priori* defined limits (e.g., $\pm 1.96SD$ of the differences around the mean)



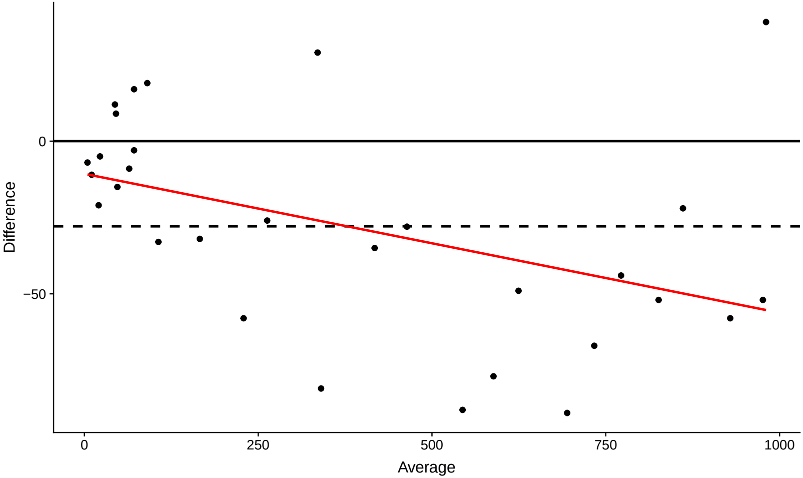
Bland-Altman plot for data normalization

- It is expected that measures from the two assays would be identical, or very similar.
- The difference between each pair of measures should be near zero.
- On the Bland-Altman plot, the differences should form a cloud around zero line.
  - Any deviation from the zero line indicates that one assay consistently over- or underestimates the measures.

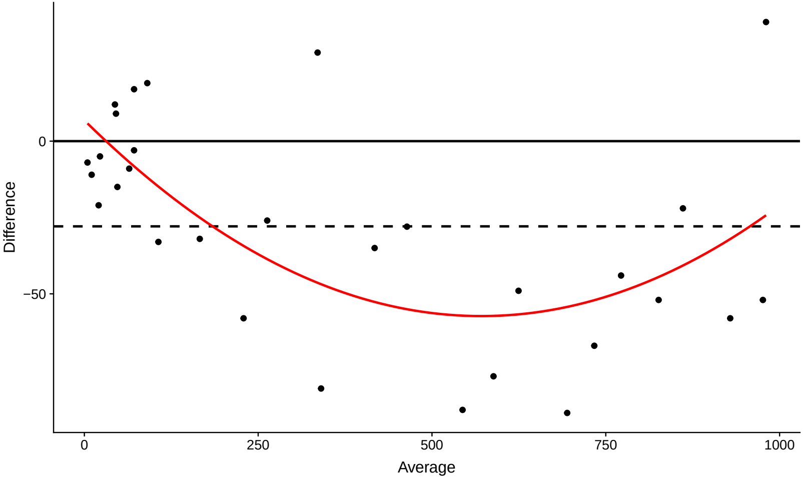
Bland-Altman plot for data normalization



Bland-Altman plot for data normalization



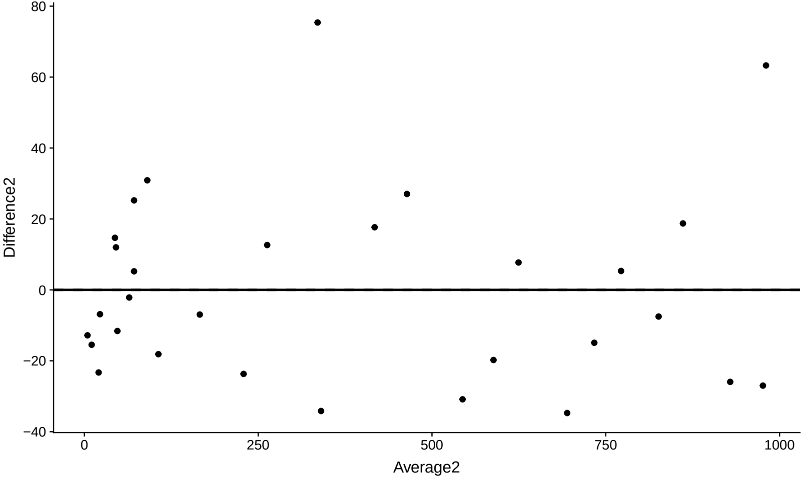
Bland-Altman plot for data normalization



Bland-Altman plot for data normalization

- When the two measures differ in some systematic way, we see deviations from $\mathit{Difference} = 0$.
- We don't know which measure is correct; hence, the goal is to adjust each measure in such a way that the differences between them become minimal (centered around $\mathit{Difference} = 0$).
    - Fit a curve through the Bland-Altman plot, L are the fitted values of the curve.
    - Subtract half of the fit from measures for Assay 1 and add half of the fit to measures for Assay 2.
        - $\mathit{Array}1 = \mathit{Array}1 - L/2$
        - $\mathit{Array}2 = \mathit{Array}2 + L/2$
    - Recalculate Difference and Average using the adjusted measures.
    - Plot them on the Bland-Altman plot.

Bland-Altman plot for data normalization



Scaling up the Bland-Altman plot

Adaptation of the Bland-Altman plot for omics data

- The Bland-Altman plot has become commonly used in omics data
- Typically it is used in the form of an MA plot (Minus vs. Average plot)
- MA plots are on the log scale
- $M = log(X) - log(Y) = log(X/Y)$
- $A = \frac{1}{2}log(X) + \frac{1}{2}log(Y) = \frac{1}{2}log(XY)$

Normalizing omics data

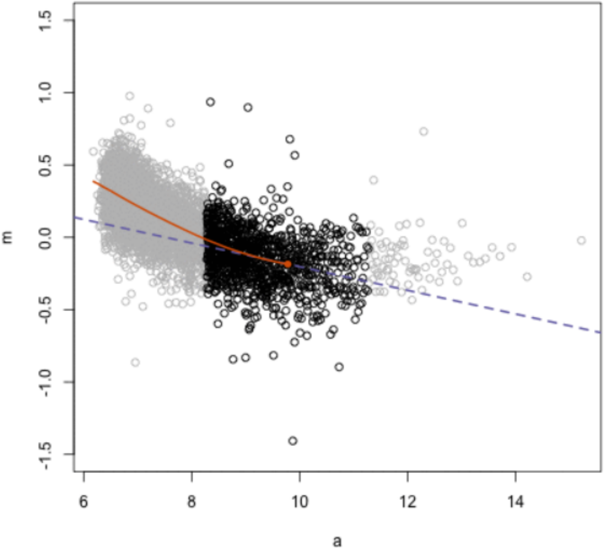
- It is assumed that the majority of genes (or genomic features) do not change expression between samples; thus, the average difference between samples should be 0.
- Average gene expression can be used as a common reference for normalization on the MA plot
- The logarithmic scale is important in genomics data due to a large range of possible values in the data

Loess normalization on the MA plot

- Loess - locally estimated scatterplot smoothing.
- A form of local regression where only a subset of the data is used to fit a curve - **non-parametric**, **data-driven**.
- Loess is the most common method of fitting a model to an MA plot for normalization.
- Allows for the removal of between sample biases.

William S. Cleveland, Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association Vol. 74, No. 368 (Dec., 1979). DOI: 10.2307/2286407

Loess

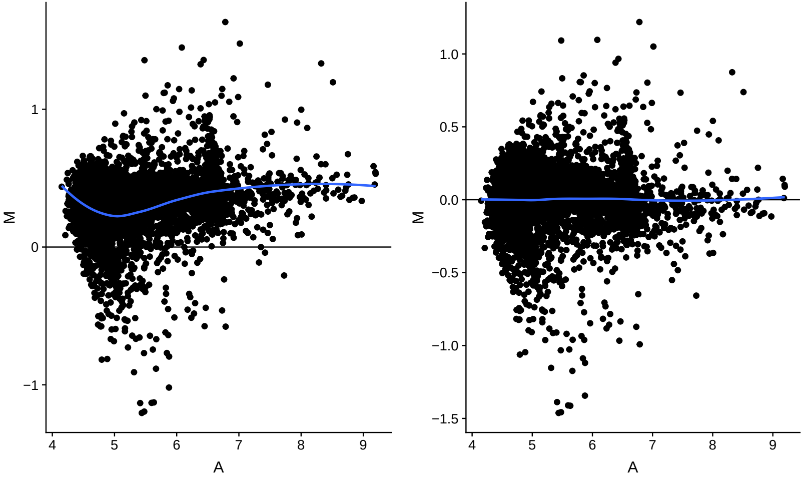


- `loess(), fANCOVA::loess.as()`

Steps of loess normalization

- Calculate M and A for the two datasets to be compared.
- Fit loess curve to MA plot.
- $log(X_N) = log(X) - L/2$
- $log(Y_N) = log(Y) + L/2$
- L are the fitted values of the loess curve, X_N, Y_N are the normalized values for the two datasets.

Example Normalization



Cyclic loess

When there are more than 2 samples to normalize cyclic loess can be used

1. Choose two out of the N total samples then generate an MA plot.
2. Fit a loess curve $f(x)$ to the MA plot.
3. Subtract $f(x)/2$ from the first dataset and add $f(x)/2$ to the second.
4. Repeat until all unique pairs have been compared.
5. Repeat until convergence.

Packages

- Several R packages have support for MA plotting and normalization of genomics data
- `limma`, `edgeR`, `DESeq2`
- Built-in functions for plotting and normalizing RNA-seq and microarray data

http://bioconductor.org/packages/limma/

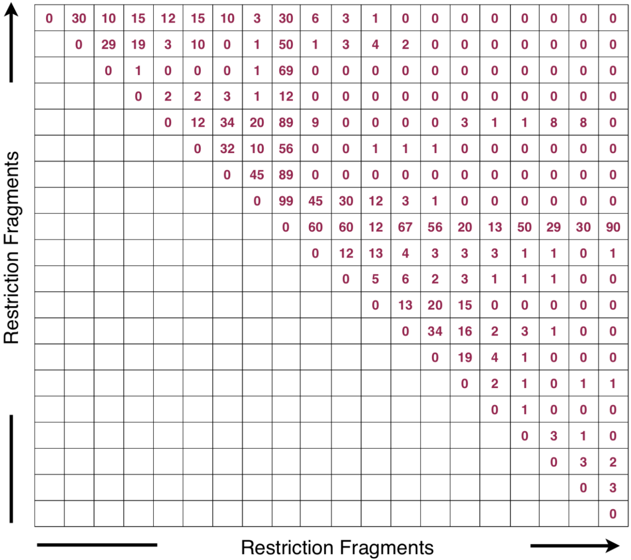
http://bioconductor.org/packages/edgeR/

https://github.com/mikelove/DESeq2

Hi-C data

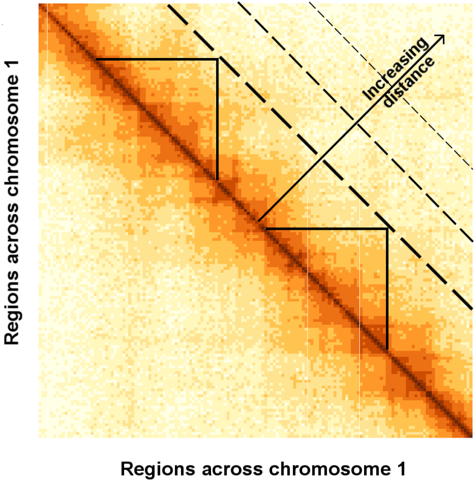
- Hi-C data is a sequencing technique that captures the 3D structure of the DNA.
- Symmetric matrix with values indicating the strength of interaction between two regions

Hi-C data


Restriction Fragments
Restriction Fragments

0	30	10	15	12	15	10	3	30	6	3	1	0	0	0	0	0	0	0	0
	0	29	19	3	10	0	1	50	1	3	4	2	0	0	0	0	0	0	0
		0	1	0	0	0	1	69	0	0	0	0	0	0	0	0	0	0	0
			0	2	2	3	1	12	0	0	0	0	0	0	0	0	0	0	0
				0	12	34	20	89	9	0	0	0	0	3	1	1	8	8	0
					0	32	10	56	0	0	1	1	1	0	0	0	0	0	0
						0	45	89	0	0	0	0	0	0	0	0	0	0	0
							0	99	45	30	12	3	1	0	0	0	0	0	0
								0	60	60	12	67	56	20	13	50	29	30	90
									0	12	13	4	3	3	3	1	1	0	1
										0	5	6	2	3	1	1	1	0	0
											0	13	20	15	0	0	0	0	0
												0	34	16	2	3	1	0	0
													0	19	4	1	0	0	0
														0	2	1	0	1	1
															0	1	0	0	0
																0	3	1	0
																	0	3	2
																		0	3
																			0

Hi-C data



- MA normalization has been applied to Hi-C data
- Our work proposed a modification of MA normalization

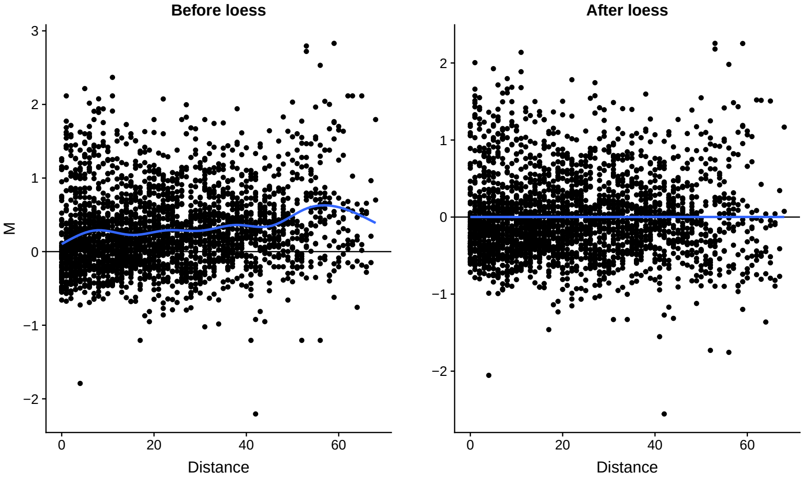
Our modification: MD plot

- Minus vs. Distance (MD) plot.
- Used for normalizing Hi-C data in the `HiCcompare` and `multiHiCcompare` R packages.
- Unit distance (D) between interacting genomic regions is used instead of A.
- This is better suited to the distance-centric nature of Hi-C data.

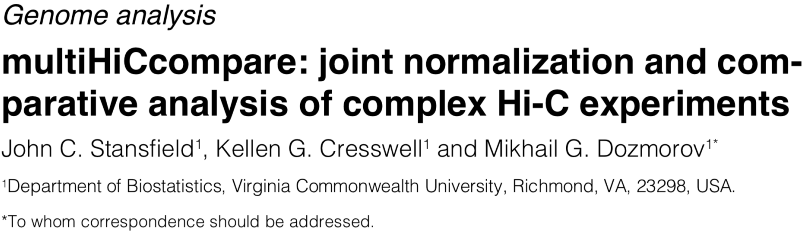
http://bioconductor.org/packages/HiCcompare/

http://bioconductor.org/packages/multiHiCcompare/

MD plot example



multiHiCcompare



Genome analysis

multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments

John C. Stansfield[1], Kellen G. Cresswell[1] and Mikhail G. Dozmorov[1*]

[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA.

*To whom correspondence should be addressed.

- Normalization utilizes cyclic loess on the MD plot.
- Accepted for publication in *Bioinformatics* journal.

DOI: 10.1093/bioinformatics/btz048 (in press)

Summary: Application of Bland-Altman plot

- Used to measure the agreement of a new measure with a gold standard.
- Compare two measures/check reproducibility.
- Remove bias (normalize) datasets in a data-driven manner.

Giavarina, Davide. “Understanding Bland Altman Analysis.” Biochemia Medica 25, no. 2 (2015): 141–51. https://doi.org/10.11613/BM.2015.015.

Thank you

Questions?

https://github.com/jstansfield0/Talk_Biostats_Grand_Rounds_2019