Acknowledgements

Lorem Ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum

Contents

Introduction

To provide a comprehensive background for the following chapters that focus on the interaction of human mesenchymal stromal cells (hMSCs) with multiple myeloma (MM) cells, this

Aims

This project defines these aims:

- Characterise the interaction between myeloma cells and mesenchymal stromal cells
- Aim 2
- Aim 3

Chapter 2: Semi-Automation of Data Analysis

Abstract

`plotastic` addresses the challenges of transitioning from exploratory data analysis to hypothesis testing in Python's data science ecosystem. Bridging the gap between `seaborn` and `pingouin`, this library offers a unified environment for plotting and statistical analysis. It simplifies the workflow with user-friendly syntax and seamless integration with familiar `seaborn` parameters (y, x, hue, row, col). Inspired by `seaborn`'s consistency, `plotastic` utilizes a `DataAnalysis` object to intelligently pass parameters to `pingouin` statistical functions. Hence, statistics and plotting are performed on the same set of parameters, so that the strength of `seaborn` in visualizing multidimensional data is extended onto statistical analysis. In essence, `plotastic` translates `seaborn` parameters into statistical terms, configures statistical protocols based on intuitive plotting syntax and returns a `matplotlib` figure with known customization options and more. This approach streamlines data analysis, allowing researchers to focus on correct statistical testing and less about specific syntax and implementations.

Introduction

The reproducibility crisis in research highlights a significant challenge in contemporary biosciences, where a substantial portion of studies faces reproducibility issues (Begley & Ioannidis, 2015). One critical yet often overlooked aspect contributing to this crisis is data management. The literature most often refers to *big data* as the main challenge (Gomez-Cabrero et al., 2014). However, these challenges are also present in smaller datasets, which the author refers to as *semi-big data*. This term describes datasets that, while not extensive enough to necessitate advanced computational tools typically reserved for *big data*, are sufficiently large to render manual analysis very time-intensive. *Semi-big data* is often generated by methods like automated microscopy or multiplex qPCR, which produce volumes of data that are manageable on a surface level, but pose substantial barriers for in-depth, manual reproducibility (Bustin, 2014). This is further complicated by the complexity inherent in multidimensional datasets. For example, the qPCR experiment from Chapter 1, Fig. 4 involves the analysis of 19 genes across in three subpopulations, including eleven biological and three technical replicates, resulting in a total of 1881 data points that are all assigned to a complex set of experimental variables. Without a clearly documented data analysis protocol and standardized data formats, the reproduction of such analysis becomes extremely challenging, if not impossible (Bustin, 2014).

The evolving standards in data analysis advocate for the standardization of analytical pipelines, rationalization of sample sizes, and enhanced infrastructure for data storage, address-

ing some of these challenges (Goodman et al., 2016; Wilkinson et al., 2016). However, these advancements can place undue pressure on researchers, particularly those with limited training in statistics, underscoring the need for intuitive, user-friendly analytical tools (Gosselin, 2021; Armstrong, 2014; Gómez-López et al., 2019)

In this context, `plotastic` emerges as a tool designed to democratize access to sophisticated statistical analysis, offering a user-centric interface that caters to researchers across varying levels of statistical proficiency. By integrating robust statistical methodologies within an accessible framework, `plotastic` aims to contribute to enhancing the reproducibility and integrity of research in the biosciences (Gomez-Cabrero et al., 2014).

initially, the need to develop `plotastic` arose during this project. The first is to address the author's need for a tool that could handle the complex, multidimensional data generated by e.g. qPCR experiments. These experiments typically involve the analysis of multiple genes across several time points and biological replicates, resulting in datasets that are challenging to analyze manually. The author's experience with traditional statistical software, such as Prism, revealed that these tools required extensive manual input, making them unsuitable for the efficient analysis of complex, multidimensional data. - The second was to increase speed. THis is required for developing methods

Since `plotastic` optimizes the analysis of *semi-big data*, we introduce the term *semi-automation* to distinguish itself from the fully automated pipelines used for *big data*. Semi-automation is defined as the following aspects:

1. **Semi-big input:** The input size is oriented towards *semi-big data*, which is characterized as being manageable by manual analysis, yet highly time inefficient, and probably impossible to re-analyse by someone else than the researcher.

2. **Standardized input** The input follows a standardized format (e.g. long-format)

3. **Minimize user configuration:** User configuration is strictly minimized. The user is never asked to pass the same parameters twice. This reduces the risk of human error and time spent on configuration.

4. **Default configuration provides acceptable results:** If the user does not provide any manual configuration, the pipeline should provide acceptable results. Options should be provided to allow a level of flexibility to adapt the pipeline to the user's needs.

5. **Small Reviewable Processing Steps:** The analysis steps are structured into small processes that can be combined to form a complete analysis pipeline. That way, each step can act as a stage for quality control to improve error detection and troubleshooting. For

a statistical analysis, that means the processing steps are separated into 3 steps, those being assumption testing, factor analysis and post-hoc testing.

6. **Isolated Steps:** Processing steps should work independently from another, in the best case only depending on the raw data input. If a processing step depends on the output from other steps, the software should tell the user what exact steps it expects.

7. **Human readable outputs:** Every processing step may provide an output that is not necessarily standardized, but is required to be human readable to ensure reviewability.

Challenges: - Reproducibility crisis? - Data is exploding - Demands for rigorous statistical analysis are increasing - Biologists are not trained in statistics

The demands are rising: (Moreno-Indias et al., 2021)

As laid out in the introduction, one can doubt if a PhD student without coding skillsis at its max efficiency.

Why does Biomedicine need plotastic?: - Thorough analysis has become a standard, with assumption testing, omnibus tests and post-hoc analyses for every experiment. - But data is increasing - Example of my data? - The number of dedicated statisticians is limited - The know-how of statistics in biology is limited, for example, Some authors ignored the problem of multiple testing while others used the method uncritically with no rationale or discussion (Perneger, 1998; Armstrong, 2014)

Why did I need plotastic?

Why do biologists need plotastic? - Assays output more data in shorter time, e.g. multiplex qPCR - example: 20 genes, 3 timepoints, 11 biological replicates, (all 3 technical replicates already averaged) - 20 * 3 * 11 = 660 data points

this is multidimensional data: 660 data points spread across two dimensions: time and gene

-in manual analysis e.g. in Excel, the user has to manually select the data, copy it, paste it into a new sheet, and then perform the statistical test. In Prism, the user has to select the data, click on the statistical test, and then select the data again. This is not only time-consuming, but also prone to

- Re-Analysis: The user has to repeat the process for every gene and timepoint. This is not only time-consuming, but also prone to errors.

shortly Describe Main Packages in more detail: - seaborn: It multidimensional data - pingouin: It's a statistical package

manuscripts into the formatting of the thesis

Statement of Need

Python's data science ecosystem provides powerful tools for both visualization and statistical testing. However, the transition from exploratory data analysis to hypothesis testing can be cumbersome, requiring users to switch between libraries and adapt to different syntaxes. `seaborn` has become a popular choice for plotting in Python, offering an intuitive interface. Its statistical functionality focuses on descriptive plots and bootstrapped confidence intervals (Waskom, 2021). The library `pingouin` offers an extensive set of statistical tests, but it lacks integration with common plotting capabilities (Vallat, 2018). `statannotations` integrates statistical testing with plot annotations, but uses a complex interface and is limited to pairwise comparisons (Charlier et al., 2022).

`plotastic` addresses this gap by offering a unified environment for plotting and statistical analysis. With an emphasis on user-friendly syntax and integration of familiar `seaborn` parameters, it simplifies the process for users already comfortable with `seaborn`. The library ensures a smooth workflow, from data import to hypothesis testing and visualization.

Example

The following code demonstrates how `plotastic` analyzes the example dataset "fmri", similar to Waskom (2021) (Figure 1).

```
### IMPORT PLOTASTIC
import plotastic as plst

# IMPORT EXAMPLE DATA
DF, _dims = plst.load_dataset("fmri", verbose = False)

# EXPLICITLY DEFINE DIMENSIONS TO FACET BY
dims = dict(
      y = "signal",    # y-axis, dependent variable
      x = "timepoint", # x-axis, independent variable (within-subject factor)
      hue = "event",   # color,  independent variable (within-subject factor)
      col = "region"   # axes,   grouping variable
)
# INITIALIZE DATAANALYSIS OBJECT
DA = plst.DataAnalysis(
      data=DF,           # Dataframe, long format
      dims=dims,         # Dictionary with y, x, hue, col, row
      subject="subject", # Datapoints are paired by subject (optional)
      verbose=False,     # Print out info about the Data (optional)
)
# STATISTICAL TESTS
DA.check_normality()   # Check Normality
DA.check_sphericity()  # Check Sphericity
```

```
DA.omnibus_rm_anova()  # Perform RM-ANOVA
DA.test_pairwise()     # Perform Posthoc Analysis

# PLOTTING
(DA
.plot_box_strip()     # Pre-built plotting function initializes plot
.annotate_pairwise(   # Annotate results from DA.test_pairwise()
      include="__HUE" # Use only significant pairs across each hue
      )
)
```

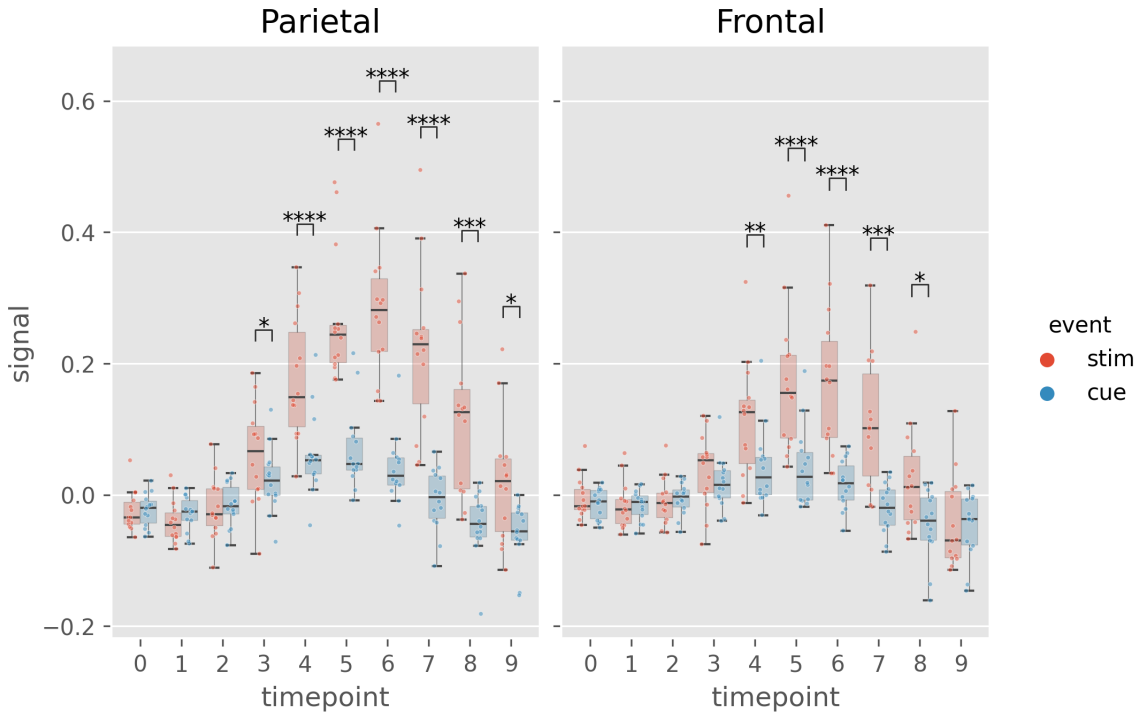

Figure 1: Example figure of `plotastic` (version 0.1). Image style was set by `plt.style.use("ggplot")`

Table 1: Results from `DA.check_sphericity()`. `plotastic` assesses sphericity after grouping the data by all grouping dimensions (hue, row, col). For example, `DA.check_sphericity()` grouped the 'fmri' dataset by "region" (col) and "event" (hue), performing four subsequent sphericity tests for four datasets.

'region', 'event'	spher	W	chi2	dof	pval	group count	n per group
'frontal', 'cue'	True	3.26e+20	-462.7	44	1	10	[14]
'frontal', 'stim'	True	2.45e+17	-392.2	44	1	10	[14]
'parietal', 'cue'	True	1.20e+20	-452.9	44	1	10	[14]
'parietal', 'stim'	True	2.44e+13	-301.9	44	1	10	[14]

Table 2: Results of `DA.omnibus_rm_anova()`. `plotastic` performs one two-factor RM-ANOVA per axis (grouping the data by row and col dimensions) using x and hue as the within-factors. For this example, `DA.omnibus_rm_anova()` grouped the 'fmri' dataset by "region" (col), performing two subsequent two-factor RM-ANOVAs. Within-factors are "timepoint" (x) and "event" (hue). For conciseness, GG-Correction and effect sizes are not shown.

'region'	Source	SS	ddof1	ddof2	MS	F	p-unc	stars
'parietal'	timepoint	1.583	9	117	0.175	26.20	3.40e-24	****
'parietal'	event	0.770	1	13	0.770	85.31	4.48e-07	****
'parietal'	timepoint * event	0.623	9	117	0.069	29.54	3.26e-26	****
'frontal'	timepoint	0.686	9	117	0.076	15.98	8.28e-17	****
'frontal'	event	0.240	1	13	0.240	23.44	3.21e-4	***
'frontal'	timepoint * event	0.242	9	117	0.026	13.031	3.23e-14	****

Overview

The functionality of `plotastic` revolves around a seamless integration of statistical analysis and plotting, leveraging the capabilities of `pingouin`, `seaborn`, `matplotlib` and `statannotations` (Vallat, 2018; Waskom, 2021; Hunter, 2007; Charlier et al., 2022). It utilizes long-format `pandas DataFrames` as its primary input, aligning with the conventions of `seaborn` and ensuring compatibility with existing data structures (Wickham, 2014; Team, 2020; McKinney, 2010).

`plotastic` was inspired by `seaborn` using the same set of intuitive and consistent parameters (y, x, hue, row, col) found in each of its plotting functions (Waskom, 2021). These parameters intuitively delineate the data dimensions plotted, yielding 'facetted' subplots, each presenting y against x. This allows for rapid and insightful exploration of multidimensional relationships. `plotastic` extends this principle to statistical analysis by storing these `seaborn` parameters (referred to as dimensions) in a `DataAnalysis` object and intelligently passing them to statistical functions of the `pingouin` library. This approach is based on the impression that most decisions during statistical analysis can be derived from how the user decides to arrange the data in a plot. This approach also prevents code repetition and streamlines statistical analysis. For example, the subject keyword is specified only once during `DataAnalysis` initialisation, and `plotastic` selects the appropriate paired or unpaired version of the test. Using `pingouin` alone requires the user to manually pick the correct test and to repeatedly specify the subject keyword in each testing function.

In essence, `plotastic` translates plotting parameters into their statistical counterparts. This translation minimizes user input and also ensures a coherent and logical connection between plotting and statistical analysis. The goal is to allow the user to focus on choosing the correct statistical test (e.g. parametric vs. non-parametric) and worry less about specific implementations.

At its core, `plotastic` employs iterators to systematically group data based on various dimensions, aligning the analysis with the distinct requirements of tests and plots. Normality

testing is performed on each individual sample, which is achieved by splitting the data by all grouping dimensions and also the x-axis (hue, row, col, x). Sphericity and homoscedasticity testing is performed on a complete sampleset listed on the x-axis, which is achieved by splitting the data by all grouping dimensions (hue, row, col) (Table 1). For omnibus and posthoc analyses, data is grouped by the row and col dimensions in parallel to the `matplotlib` axes, before performing one two-factor analysis per axis using x and hue as the within/between-factors. (Table 2).

`DataAnalysis` visualizes data through predefined plotting functions designed for drawing multi-layered plots. A notable emphasis within `plotastic` is placed on showcasing individual datapoints alongside aggregated means or medians. In detail, each plotting function initializes a `matplotlib` figure and axes using `plt.subplots()` while returning a `DataAnalysis` object for method chaining. Axes are populated by `seaborn` plotting functions (e.g., `sns.boxplot()`), leveraging automated aggregation and error bar displays. Keyword arguments are passed to these `seaborn` functions, ensuring the same degree of customization. Users can further customize plots by chaining `DataAnalysis` methods or by applying common `matplotlib` code to override `plotastic` settings. Figures are exported using `plt.savefig()`.

`plotastic` also focuses on annotating statistical information within plots, seamlessly incorporating p-values from pairwise comparisons using `statannotations` (Charlier et al., 2022). This integration simplifies the interface and enables options for pair selection in multidimensional plots, enhancing both user experience and interpretability.

For statistics, `plotastic` integrates with the `pingouin` library to support classical assumption and hypothesis testing, covering parametric/non-parametric and paired/non-paired variants. Assumptions such as normality, homoscedasticity, and sphericity are tested. Omnibus tests include two-factor RM-ANOVA, ANOVA, Friedman, and Kruskal-Wallis. Posthoc tests are t-tests, Wilcoxon, and Mann-Whitney-U.

To sum up, `plotastic` stands as a unified and user-friendly solution catering to the needs of researchers and data scientists, seamlessly integrating statistical analysis with the power of plotting in Python. It streamlines the workflow, translates `seaborn` parameters into statistical terms, and supports extensive customization options for both analysis and visualization.

Discussion

Is plotastic tested? Coverage? Does it cover every feature? What is not covered

Is plotastic USABLE for biologists? - Yes but use is limited by minimal knowledge of Python - However, that is subject to change as Python is becoming more popular in biology and AI assisted coding decreased the barrier to entry significantly. Tools like github copilot are able to

generate code, fix bugs and suggest improvements. This is a game changer for biologists that are not familiar with programming. - Furthermore, installing and using plotastic for biologists is overestimated. These steps re needed: - Install anaconda from the internet - Open the terminal - Type `pip install plotastic` - Check Rea

The evaluation of plotastic within this thesis reflects its potential to address key challenges in the field of data analysis. The software integrates a comprehensive suite of statistical tests, such as ANOVA and t-tests, designed for adaptability and ease of use, leveraging the functionalities of pingouin.

In the context of the reproducibility crisis in scientific research, plotastic offers noteworthy contributions, though it is not positioned as a universal remedy. The tool's unique approach to integrating statistical analysis with visual representation establishes a new paradigm, promoting methodological transparency. By mandating that statistical analyses accompany relevant graphical outputs, plotastic ensures that analyses are not only conducted with proper scientific rigor but also documented in a manner that facilitates replication, provided the user possesses proficiency in Python.

Usability is a critical attribute of analytical software, particularly as researchers confront increasingly complex datasets. While the developer's intimate familiarity with plotastic may bias perceptions of its ease of use, it is recognized that novices may initially encounter challenges. Nevertheless, plotastic is distinguished by its user-friendly interface, enabling users with minimal statistical training to perform sophisticated analyses by intuitively mapping plotting concepts to statistical operations.

The transition to a new analytical framework, especially one that incorporates coding, presents a learning curve. However, the advantages of plotastic in terms of analytical clarity, speed, and depth are anticipated to outweigh these initial challenges. Support mechanisms, such as assistance from advanced AI like ChatGPT, are available to mitigate these hurdles, supporting users across varying levels of expertise.

In conclusion, plotastic is posited as a valuable tool in the landscape of scientific research, offering a means to enhance the reproducibility and efficiency of data analysis. Its development ethos emphasizes simplifying complex analytical tasks, thereby contributing to the broader goal of fostering transparent and reproducible research practices.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016, March). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems* (No. arXiv:1603.04467). arXiv. Retrieved 2024-03-07, from http://arxiv.org/abs/1603.04467 doi: 10.48550/arXiv.1603.04467

Alcorta-Sevillano, N., Macías, I., Infante, A., & Rodríguez, C. I. (2020, December). Deciphering the Relevance of Bone ECM Signaling. *Cells*, *9*(12), 2630. Retrieved 2023-12-20, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7762413/ doi: 10.3390/cells9122630

Armstrong, R. A. (2014, September). When to use the Bonferroni correction. *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists)*, *34*(5), 502–508. doi: 10.1111/opo.12131

Barzilay, R., Ben-Zur, T., Bulvik, S., Melamed, E., & Offen, D. (2009, May). Lentiviral delivery of LMX1a enhances dopaminergic phenotype in differentiated human bone marrow mesenchymal stem cells. *Stem cells and development*, *18*(4), 591–601. doi: 10.1089/scd.2008.0138

Begley, C. G., & Ioannidis, J. P. A. (2015, January). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*(1), 116–126. doi: 10.1161/CIRCRESAHA.114.303819

Bianco, P. (2014). "Mesenchymal" stem cells. *Annual review of cell and developmental biology*, *30*, 677–704. doi: 10.1146/annurev-cellbio-100913-013132

Bladé, J., Beksac, M., Caers, J., Jurczyszyn, A., von Lilienfeld-Toal, M., Moreau, P., … Richardson, P. (2022, March). Extramedullary disease in multiple myeloma: A systematic literature review. *Blood Cancer Journal*, *12*(3), 1–10. Retrieved 2023-03-24, from https://www.nature.com/articles/s41408-022-00643-3 doi: 10.1038/s41408-022-00643-3

Bondi, A. B. (2000, September). Characteristics of scalability and their impact on performance. In *Proceedings of the 2nd international workshop on Software and performance* (pp. 195–203). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-03-07, from https://dl.acm.org/doi/10.1145/350391.350432 doi: 10.1145/350391.350432

Boswell, D., & Foucher, T. (2011). *The Art of Readable Code: Simple and Practical Techniques for Writing Better Code*. "O'Reilly Media, Inc.".

Brooke, J. (1996, January). SUS – a quick and dirty usability scale. In (pp. 189–194).

Bustin, S. A. (2014, December). The reproducibility of biomedical research: Sleepers awake! *Biomolecular Detection and Quantification*, *2*, 35–42. Retrieved 2024-03-18, from https://www.sciencedirect.com/science/article/pii/S2214753515000030 doi: 10.1016/j.bdq.2015.01.002

Caplan, A. (1991). Mesenchymal stem cells. *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, *9*(5), 641–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1870029 doi: 10.1002/jor.1100090504

Caplan, A. I. (1994, July). The mesengenic process. *Clinics in plastic surgery*, *21*(3), 429–435.

Chacon, S., & Straub, B. (2024, March). *Git - Book*. Retrieved 2024-03-07, from https://git-scm.com/book/de/v2

Charlier, F., Weber, M., Izak, D., Harkin, E., Magnus, M., Lalli, J., … Repplinger, S. (2022, October). *Trevis-md/statannotations: V0.5*. Zenodo. Retrieved 2023-11-16, from https://zenodo.org/record/7213391 doi: 10.5281/ZENODO.7213391

Cooper, G. M. (2000). The Cell: A Molecular Approach. 2nd Edition. *Sinauer Associates*, Proliferation in Development and Differentiation. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK9906/

da Silva Meirelles, L., Chagastelles, P. C., & Nardi, N. B. (2006, June). Mesenchymal stem cells reside in virtually all post-natal organs and tissues. *Journal of cell science*, *119*(Pt 11), 2204–2213. doi: 10.1242/jcs.02932

Dominici, M., Le Blanc, K., Mueller, I., Slaper-Cortenbach, I., Marini, F., Krause, D., … Horwitz, E. (2006). Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. *Cytotherapy*, *8*(4), 315–317. doi: 10.1080/14653240600855905

Duvall, P., Matyas, S., & Glover, A. (2007). *Continuous integration: Improving software quality and reducing risk*. Pearson Education. Retrieved from https://books.google.de/books?id=PV9qfEdv9L0C

Ekmekci, B., McAnany, C. E., & Mura, C. (2016, July). An Introduction to Programming for Bioscientists: A Python-Based Primer. *PLOS Computational Biology*, *12*(6), e1004867. Retrieved 2024-03-10, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004867 doi: 10.1371/journal.pcbi.1004867

Fazeli, P. K., Horowitz, M. C., MacDougald, O. A., Scheller, E. L., Rodeheffer, M. S., Rosen, C. J., & Klibanski, A. (2013, March). Marrow Fat and Bone—New Perspectives. *The Journal of Clinical Endocrinology and Metabolism*, *98*(3), 935–945. Retrieved 2023-12-20, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3590487/ doi: 10.1210/jc.2012-3634

Friedenstein, A., & Kuralesova, A. I. (1971, August). Osteogenic precursor cells of bone marrow in radiation chimeras. *Transplantation*, *12*(2), 99–108.

Friedenstein, A. J., Piatetzky-Shapiro, I. I., & Petrakova, K. V. (1966, December). Osteogenesis in transplants of bone marrow cells. *Journal of embryology and experimental morphology*, *16*(3), 381–390.

Gabr, M. M., Zakaria, M. M., Refaie, A. F., Ismail, A. M., Abou-El-Mahasen, M. A., Ashamallah, S. A., … Ghoneim, M. A. (2013). Insulin-producing cells from adult human bone marrow mesenchymal stem cells control streptozotocin-induced diabetes in nude mice. *Cell transplantation*, *22*(1), 133–145. doi: 10.3727/096368912X647162

García-Ortiz, A., Rodríguez-García, Y., Encinas, J., Maroto-Martín, E., Castellano, E., Teixidó, J., & Martínez-López, J. (2021, January). The Role of Tumor Microenvironment in Multiple Myeloma Development and Progression. *Cancers*, *13*(2). Retrieved 2021-02-02, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7827690/ doi: 10.3390/cancers13020217

Glavey, S. V., Naba, A., Manier, S., Clauser, K., Tahri, S., Park, J., … Ghobrial, I. M. (2017, November). Proteomic characterization of human multiple myeloma bone marrow extracellular matrix. *Leukemia*, *31*(11), 2426–2434. Retrieved 2023-09-05, from https://www.nature.com/articles/leu2017102 doi: 10.1038/leu.2017.102

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., … Tegnér, J. (2014, March). Data integration in the era of omics: Current and future challenges. *BMC Systems Biology*, *8*(2), I1. Retrieved 2024-03-18, from https://doi.org/10.1186/1752-0509-8-S2-I1 doi: 10.1186/1752-0509-8-S2-I1

Gómez-López, G., Dopazo, J., Cigudosa, J. C., Valencia, A., & Al-Shahrour, F. (2019, May). Precision medicine needs pioneering clinical bioinformaticians. *Briefings in Bioinformatics*, *20*(3), 752–766. doi: 10.1093/bib/bbx144

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016, June). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12-341ps12. Retrieved 2024-03-18, from https://www.science.org/doi/10.1126/scitranslmed.aaf5027 doi: 10.1126/scitranslmed.aaf5027

Gosselin, R.-D. (2021, February). Insufficient transparency of statistical reporting in preclinical research: A scoping review. *Scientific Reports*, *11*(1), 3335. Retrieved 2024-03-11, from https://www.nature.com/articles/s41598-021-83006-5 doi: 10.1038/s41598-021-83006-5

Gronthos, S., Graves, S. E., Ohta, S., & Simmons, P. J. (1994, December). The STRO-1+ fraction of adult human bone marrow contains the osteogenic precursors. *Blood*, *84*(12), 4164–4173.

Hunter, J. D. (2007, May). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*,

9(3), 90–95. Retrieved 2023-11-15, from https://ieeexplore.ieee.org/document/4160265 doi: 10.1109/MCSE.2007.55

Incerti, D., Thom, H., Baio, G., & Jansen, J. P. (2019, May). R You Still Using Excel? The Advantages of Modern Software Tools for Health Technology Assessment. *Value in Health*, *22*(5), 575–579. Retrieved 2024-03-11, from https://www.sciencedirect.com/science/article/pii/S1098301519300506 doi: 10.1016/j.jval.2019.01.003

Jansen, B. J. H., Gilissen, C., Roelofs, H., Schaap-Oziemlak, A., Veltman, J. A., Raymakers, R. A. P., … Adema, G. J. (2010, April). Functional differences between mesenchymal stem cell populations are reflected by their transcriptome. *Stem cells and development*, *19*(4), 481–490. doi: 10.1089/scd.2009.0288

Kazman, R., Bianco, P., Ivers, J., & Klein, J. (2020, December). *Maintainability* (Report). Carnegie Mellon University. Retrieved 2024-03-07, from https://kilthub.cmu.edu/articles/report/Maintainability/12954908/1 doi: 10.1184/R1/12954908.v1

Kibler, C., Schermutzki, F., Waller, H. D., Timpl, R., Müller, C. A., & Klein, G. (1998, June). Adhesive interactions of human multiple myeloma cell lines with different extracellular matrix molecules. *Cell Adhesion and Communication*, *5*(4), 307–323. doi: 10.3109/15419069809040300

Krekel, H., Oliveira, B., Pfannschmidt, R., Bruynooghe, F., Laugher, B., & Bruhin, F. (2004). *Pytest.* Retrieved from https://github.com/pytest-dev/pytest

McKinney, W. (2010, January). Data Structures for Statistical Computing in Python. In (pp. 56–61). doi: 10.25080/Majora-92bf1922-00a

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., … Zomer, A. L. (2021, February). Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Frontiers in Microbiology*, *12*. Retrieved 2024-03-18, from https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.635781/full doi: 10.3389/fmicb.2021.635781

Muruganandan, S., Roman, A. A., & Sinal, C. J. (2009, January). Adipocyte differentiation of bone marrow-derived mesenchymal stem cells: Cross talk with the osteoblastogenic program. *Cellular and molecular life sciences : CMLS*, *66*(2), 236–253. doi: 10.1007/s00018-008-8429-z

Myers, G. J., Sandler, C., & Badgett, T. (2011). *The art of software testing* (3rd ed.). Wiley Publishing. Retrieved from https://malenezi.github.io/malenezi/SE401/Books/114-the-art-of-software-testing-3-edition.pdf

Narzt, W., Pichler, J., Pirklbauer, K., & Zwinz, M. (1998, January). A Reusability Concept for Process Automation Software..

Nowotschin, S., & Hadjantonakis, A.-K. (2010, August). Cellular dynamics in the early mouse embryo: From axis formation to gastrulation. *Current opinion in genetics & development*, *20*(4), 420–427. doi: 10.1016/j.gde.2010.05.008

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019, December). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (No. arXiv:1912.01703). arXiv. Retrieved 2024-03-07, from http://arxiv.org/abs/1912.01703 doi: 10.48550/arXiv.1912.01703

Peng, R. D. (2011, December). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. Retrieved 2024-03-18, from https://www.science.org/doi/10.1126/science.1213847 doi: 10.1126/science.1213847

Perneger, T. V. (1998, April). What's wrong with Bonferroni adjustments. *BMJ : British Medical Journal*, *316*(7139), 1236–1238. Retrieved 2021-11-24, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112991/

Pittenger, M. F., Mackay, A. M., Beck, S. C., Jaiswal, R. K., Douglas, R., Mosca, J. D., … Marshak, D. R.

(1999). Multilineage Potential of Adult Human Mesenchymal Stem Cells. , *284*(April), 143–148. doi: 10.1126/science.284.5411.143

The Python Language Reference. (n.d.). Retrieved 2024-03-07, from https://docs.python.org/3/reference/index.html

R Core Team. (2018). *R: A language and environment for statistical computing* [Manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.. Retrieved 2024-03-07, from https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Rajkumar, S. V., & Kumar, S. (2020, September). Multiple myeloma current treatment algorithms. *Blood Cancer Journal*, *10*(9), 94. Retrieved 2023-07-03, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7523011/ doi: 10.1038/s41408-020-00359-2

Rayhan, A., & Gross, D. (2023). *The Rise of Python: A Survey of Recent Research.* doi: 10.13140/RG.2.2.27388.92809

Sacchetti, B., Funari, A., Remoli, C., Giannicola, G., Kogler, G., Liedtke, S., … Bianco, P. (2016). No identical "mesenchymal stem cells" at different times and sites: Human committed progenitors of distinct origin and differentiation potential are incorporated as adventitial cells in microvessels. *Stem Cell Reports*, *6*(6), 897–913. Retrieved from http://dx.doi.org/10.1016/j.stemcr.2016.05.011 doi: 10.1016/j.stemcr.2016.05.011

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013, October). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, *9*(10), e1003285. Retrieved 2024-03-07, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3812051/ doi: 10.1371/journal.pcbi.1003285

Shenghui, H., Nakada, D., & Morrison, S. J. (2009). Mechanisms of Stem Cell Self-Renewal. *Annual Review of Cell and Developmental Biology*, *25*(1), 377–406. Retrieved from https://doi.org/10.1146/annurev.cellbio.042308.113248 doi: 10.1146/annurev.cellbio.042308.113248

Smith, A. M., Niemeyer, K. E., Katz, D. S., Barba, L. A., Githinji, G., Gymrek, M., … Vanderplas, J. T. (2018). Journal of Open Source Software (JOSS): Design and first-year review. *PeerJ Preprints*, *4*, e147. doi: 10.7717/peerj-cs.147

Stock, P., Bruckner, S., Winkler, S., Dollinger, M. M., & Christ, B. (2014, April). Human bone marrow mesenchymal stem cell-derived hepatocytes improve the mouse liver after acute acetaminophen intoxication by preventing progress of injury. *International journal of molecular sciences*, *15*(4), 7004–7028. doi: 10.3390/ijms15047004

Tam, P. P., & Beddington, R. S. (1987, January). The formation of mesodermal tissues in the mouse embryo during gastrulation and early organogenesis. *Development (Cambridge, England)*, *99*(1), 109–126.

Tanavalee, C., Luksanapruksa, P., & Singhatanadgige, W. (2016, June). Limitations of Using Microsoft Excel Version 2016 (MS Excel 2016) for Statistical Analysis for Medical Research. *Clinical Spine Surgery*, *29*(5), 203. Retrieved 2024-03-11, from https://journals.lww.com/jspinaldisorders/fulltext/2016/06000/limitations_of_using_microsoft_excel_version_2016.5.aspx doi: 10.1097/BSD.0000000000000382

Team, T. P. D. (2020, February). *Pandas-dev/pandas: Pandas.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3509134 doi: 10.5281/zenodo.3509134

Terpos, E., Ntanasis-Stathopoulos, I., Gavriatopoulou, M., & Dimopoulos, M. A. (2018, January). Pathogenesis of bone disease in multiple myeloma: From bench to bedside. *Blood Cancer Journal*, *8*(1), 7. doi: 10.1038/s41408-017-0037-4

Ullah, I., Subbarao, R. B., & Rho, G. J. (2015). Human mesenchymal stem cells - current trends and future prospective Bioscience Reports.
doi: 10.1042/BSR20150025

Vallat, R. (2018, November). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. Retrieved 2023-05-29, from https://joss.theoj.org/papers/10.21105/joss.01026 doi: 10.21105/joss.01026

van Rossum, G., Lehtosalo, J., & Langa, L. (2014). *PEP 484 – Type Hints | peps.python.org.* Retrieved 2024-03-08, from https://peps.python.org/pep-0484/

Viguet-Carrin, S., Garnero, P., & Delmas, P. D. (2006, March). The role of collagen in bone strength. *Osteoporosis International*, *17*(3), 319–336. Retrieved 2023-12-20, from https://doi.org/10.1007/s00198-005-2035-9 doi: 10.1007/s00198-005-2035-9

Waskom, M. L. (2021, April). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. Retrieved 2023-03-26, from https://joss.theoj.org/papers/10.21105/joss.03021 doi: 10.21105/joss.03021

Wickham, H. (2014, September). Tidy Data. *Journal of Statistical Software*, *59*, 1–23. Retrieved 2023-11-15, from https://doi.org/10.18637/jss.v059.i10 doi: 10.18637/jss.v059.i10

Wilkins, A., Kemp, K., Ginty, M., Hares, K., Mallam, E., & Scolding, N. (2009, July). Human bone marrow-derived mesenchymal stem cells secrete brain-derived neurotrophic factor which promotes neuronal survival in vitro. *Stem cell research*, *3*(1), 63–70. doi: 10.1016/j.scr.2009.02.006

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016, March). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. Retrieved 2024-03-18, from https://www.nature.com/articles/sdata201618 doi: 10.1038/sdata.2016.18

Xu, W., Zhang, X., Qian, H., Zhu, W., Sun, X., Hu, J., … Chen, Y. (2004, July). Mesenchymal stem cells from adult human bone marrow differentiate into a cardiomyocyte phenotype in vitro. *Experimental biology and medicine (Maywood, N.J.)*, *229*(7), 623–631.

Yang, A., Troup, M., & Ho, J. W. (2017, July). Scalability and Validation of Big Data Bioinformatics Software. *Computational and Structural Biotechnology Journal*, *15*, 379–386. Retrieved 2024-03-07, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5537105/ doi: 10.1016/j.csbj.2017.07.002