gcplyr: an R package for microbial growth curve data analysis

Michael Blazanin[1]

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA.

Correspondence to: Mike Blazanin, Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect St, New Haven, CT, 06520, USA; mike.blazanin@yale.edu

Abstract

Characterization of microbial growth is of both fundamental and applied interest. Modern platforms can automate collection of high-throughput microbial growth curves, necessitating the development of computational tools to handle and analyze these data to produce insights. However, existing tools are limited. Many use parametric analyses that require mathematical assumptions about the microbial growth characteristics. Those that use non-parametric or model-free analyses often can only quantify a few traits of interest, and none are capable of importing and reshaping all known growth curve data formats. To address this gap, here I present a newly-developed R package: gcplyr. gcplyr can flexibly import growth curve data in every known format, and reshape it under a flexible and extendable framework so that users can design custom analyses or plot data with popular visualization packages. gcplyr can also incorporate metadata and generate or import experimental designs to merge with data. Finally, gcplyr carries out model-free and non-parametric analyses, extracting a broad range of clinically and ecologically important traits, including initial density, lag time, growth rate and doubling time, carrying capacity, diauxie, area under the curve, extinction time, and more. gcplyr makes scripted analysis of growth curve data in R straightforward, streamlines common data wrangling and analysis steps, and easily integrates with common visualization and statistical analyses.

Availability

gcplyr is available from the central CRAN repository (https://CRAN.R-project.org/package=gcplyr), or from GitHub (https://github.com/mikeblazanin/gcplyr).

Supplementary information

Supplementary materials are available at *Bioinformatics* online.

Introduction

Characterization of microbial population growth dynamics has been of fundamental and applied interest since nearly the dawn of microbiology (Monod, 1949). From bacteria and archaea to yeasts, growth curves are a ubiquitous technique to study microbial growth. Indeed, modern automated platforms, including plate readers, can collect high-throughput growth data over time on hundreds of samples simultaneously. Yet, this data-generation capacity has outpaced the development of computational tools to handle and analyze microbial growth data, presenting new challenges.

First and foremost, data are rarely output in the ideal format for analysis, visualization, or publication. Reorganizing data manually can be tedious and fraught with the potential for introduction of errors. Moreover, since output files vary between different plate readers, scripted reorganization may require tailored code for each output format. Despite this, existing software tools provide limited utilities for streamlined data wrangling and reorganization (Table S1).

Once data are reorganized, scientists face the challenge of converting raw data into quantitative microbial traits. Many existing computational tools use parametric analyses of microbial growth curve data (Table S1) (Wirth and Funk, 2023; Coutin *et al.*, 2020; Bukhman *et al.*, 2015; Huang, 2014; Liu *et al.*, 2021; Veríssimo *et al.*, 2013; Vervier *et al.*, 2019; Sprouffske and Wagner, 2016; Petzoldt, 2022; Vaas *et al.*, 2013; Garre *et al.*, 2022; Cuevas and Edwards, 2017; Cuevas *et al.*, 2016; Osthege *et al.*, 2022). Parametric analyses fit a mathematical model of microbial population growth to observed data, then extract the fitted parameter values to quantify traits. This approach can be powerful, but it has drawbacks for many users. First, users must explicitly or implicitly choose a model of growth, making error-prone mathematical assumptions about the form the growth data should take. Moreover, for accurate results users must laboriously verify that their data meet the assumptions of the model and that the fitting process converged on an appropriate solution. Finally, some dynamics do not fit any known models of microbial growth, leaving researchers with few to no options to analyze such growth data.

Given these drawbacks, in recent years some groups have developed software to analyze microbial growth data with non-parametric or model-free approaches (Table S1) (Wirth and Funk, 2023; Swain *et al.*, 2016; Fernandez-Ricaud *et al.*, 2016; Jung *et al.*, 2015; Olsen *et al.*, 2010; Sprouffske and Wagner, 2016; Petzoldt, 2022; Vaas *et al.*, 2013; Cuevas and Edwards, 2017; Cuevas *et al.*, 2016; Osthege *et al.*, 2022; Midani *et al.*, 2021; Hall *et al.*, 2014; Tonner *et al.*, 2020, 2017; Hemmerich *et al.*, 2017). These analyses make no specific mathematical assumptions about the form of the growth data, instead extracting parameters of interest directly from the data itself or from non-parametrically smoothed transformations of the data. However, the vast majority of these tools can only quantify a few traits of interest.

To address the limited capacity of existing growth curve analysis software to both wrangle data and carry out model-free analyses, here I present my newly-released R package, gcplyr. gcplyr is built in R, a popular scripting language for scientific data analysis and visualization. gcplyr can import growth curve data from a wide range of plate readers, and facilitates model-free analyses to extract a number of traits, including lag time, growth rate and doubling time, carrying capacity, area under the curve, and more. Notably, the model-free analysis implemented by gcplyr enables the characterization of dynamics that do not fit traditional mathematical models of microbial growth, and therefore were difficult or impossible to analyze using existing software. Examples of these dynamics include diauxic shifts

(Blazanin, Moore, *et al.*, 2023) and growth in the presence of antagonists [e.g. bacteriophages (Blazanin, Vasen, *et al.*, 2023)]. Additionally, gcplyr provides a framework for data organization that is flexible and extendable, making it easy for users to run custom analyses or integrate gcplyr workflows with other R packages.

Implementation

gcplyr is an open-source R package (R Core Team, 2022) that can be installed with R's built-in install.packages function from the central CRAN repository (https://CRAN.R-project.org/package=gcplyr), or from GitHub (https://github.com/mikeblazanin/gcplyr). gcplyr has been written to minimize the number of external dependencies needed for installation.

Usage

gcplyr is usable with a basic working knowledge of the R coding language. A comprehensive manual and tutorials are included with the gcplyr installation and available online (https://mikeblazanin.github.io/gcplyr/). Within R, gcplyr functions make it easy to import data and metadata, merge them with experimental design information, smooth and calculate derivatives as necessary, and analyze curves to produce summary growth curve metrics (Fig 1). gcplyr was designed to easily integrate growth curve analyses with other common R tasks, including plotting with ggplot2, data transformation with tidyverse, and statistical analysis.

Inputs

Growth curve data is typically output by plate reader software as block-shaped (also called 'matrix') or wide-shaped (also called 'column-wise') (Fig 1). gcplyr can import any tabular file containing block-shaped or wide-shaped data, with user-set arguments enabling functionality regardless of where the data is located in the input file. This is an especially useful advance for block-shaped data, where each timepoint is saved separately, and which only one recently-released tool can parse (Csibra and Stan, 2023). Additionally, user-set arguments can designate metadata located in input files to be incorporated into the resulting data.

Reshaping data

After reading input files into R, gcplyr extends the functionality of tidyr (Wickham, Vaughan, *et al.*, 2023) to reshape data into a 'tidy' format (also known as 'long' format) (Wickham) ideal for subsequent visualization and analysis steps. Tidy data have all observations in a single column, with each unique datapoint with its own row, and additional columns specifying the timepoint, well, and any other information. Tidy data is the best layout for most analyses (Wickham), is consistent with requirements of data repositories like Dryad, and is the expected input for a number of popular R packages (Wickham *et al.*, 2019). By transforming data into tidy data, gcplyr makes it easy for users to also visualize their data using ggplot2 (Wickham, 2016), manipulate their data using dplyr (Wickham, François, *et al.*, 2023), or apply any of the other tidyverse packages to their data (Wickham *et al.*, 2019).

Incorporating experimental designs

gcplyr can import experimental design information from user-supplied spreadsheet files, or users can use gcplyr functions to directly generate experimental designs in R. Once imported or generated, these designs can be exported to files or easily merged with imported growth curve data.

Smoothing data

Unlike fitting-based approaches, the model-free analyses that gcplyr implements can be sensitive to experimental noise in growth curve data. To smooth noise in raw density data, gcplyr implements several well-established smoothing functions, including moving average, moving median, loess (Savitzky and Golay, 1964; Cleveland and Devlin, 1988; Cleveland, 1979; R Core Team, 2022), and GAM (Trevor Hastie and Robert Tibshirani, 1986; Wood, 2003, 2017, 2011). Smoothing functions are tuned by user-set smoothness parameters. The package documentation discusses best practices for setting smoothness parameters and smoothing data.

Calculating derivatives

Model-free calculation of many growth curve metrics depends on identifying features in the derivatives of density data. gcplyr can calculate both the population rate of growth (derivative) and cellular growth rate (per-capita derivative), using the slope of density over time. However, such derivatives can be sensitive to experimental noise, even if density data has been smoothed. To address this, gcplyr calculates derivatives by fitting a linear regression on a rolling window of multiple points (Hall *et al.*, 2014), with user-set parameters defining the width of the window. The package documentation discusses best practices for calculating derivatives and setting their smoothing parameters.

Characterizing microbial growth

Finally, gcplyr quantifies traits of interest by computing metrics from density and derivatives data. Many of these metrics are quantified by identifying peaks or valleys in the density or derivatives (e.g. maximum growth rate, doubling time, diauxie, carrying capacity), using a local extrema-finding algorithm implemented in gcplyr. Other metrics are quantified by identifying threshold-crossing events in the density or derivatives. Finally, some metrics are quantified using purpose-built functions that implement previously-established methods (e.g. lag time (Swinnen *et al.*, 2004; Yates and Smotzer, 2007) or area under the curve).

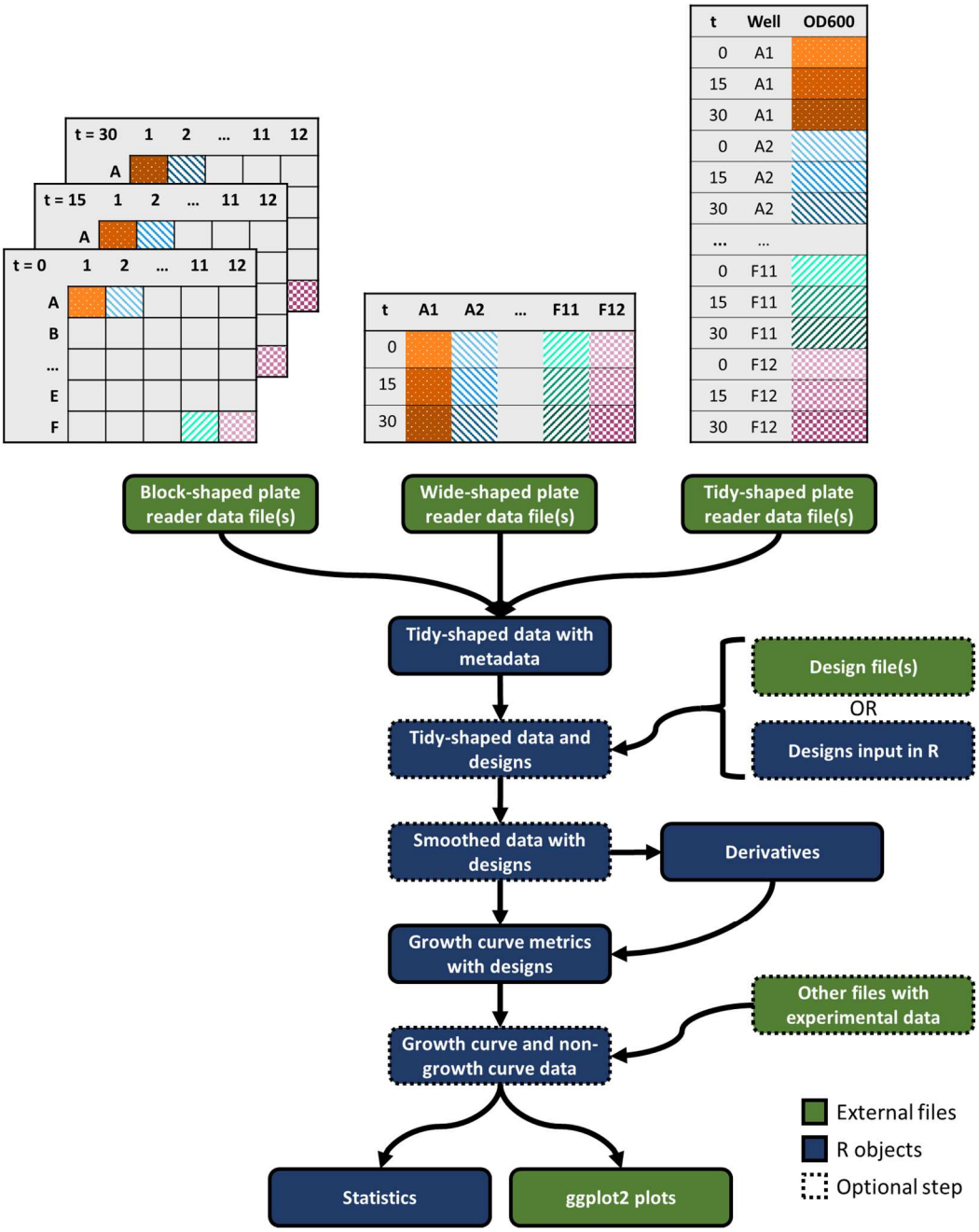

Fig 1. Workflow to use gcplyr to analyze microbial growth curve data. Growth curve data are commonly found in one of three layouts: block-shaped, wide-shaped, or tidy-shaped. Block-shaped data are organized to match the physical layout of the multi-well plate. Wide-shaped data contain one column for each well from a plate. Tidy data contain all the observations in a single column. gcplyr functions import data files regardless of the data layout, reshape them into tidy-shaped data, then merge them with experimental design information. Data and designs can then be smoothed and have derivatives calculated to extract growth curve metrics. Metrics can be easily merged with other non-growth curve experimental data before statistical analyses and visualization.

Discussion

High-throughput microbial growth data has necessitated computational tools for data wrangling and analysis. Here I introduced gcplyr, a new R package built specifically to address this need.

First, gcplyr wrangles and reshapes growth curve data. gcplyr recognizes that there are common layouts for growth curve data (Fig 1), and it flexibly imports such files. During data reorganization, gcplyr employs a tidy data framework, providing a number of benefits over other data organization approaches:

1. It enables users to integrate as many pieces of experimental design information as desired.
2. It enables users to leverage the general-use functions in gcplyr and other packages to generate custom analyses to identify unique features in their data.
3. It enables users to easily integrate their growth curve analyses with existing visualization (e.g. ggplot2) and statistics packages in R.
4. It enables users to merge growth curve data analyses with other sources of data.

gcplyr is the first package designed for scriptable data wrangling in the R programming ecosystem (Table S1). This provides high customizability for users, and enables them to easily incorporate gcplyr with other scripted analyses. In contrast, many existing growth curve analysis tools have little to no data wrangling capability (Table S1), requiring users to reformat files manually. Some packages do provide data reshaping via non-R user interfaces, including bletl via Python scripting (Osthege *et al.*, 2022), and Parsley (Csibra and Stan, 2023), QurvE (Wirth and Funk, 2023), and AUDIT (Coutin *et al.*, 2020) via graphical user interfaces.

gcplyr is also the first package in R that is designed for scriptable non-parametric analyses for a wide array of microbial growth traits (Table S1). Many existing tools quantify traits of interest by fitting parametric mathematical models of microbial growth to growth curve data, including biogrowth (Garre *et al.*, 2022), growthcurver (Sprouffske and Wagner, 2016) and AUDIT (Coutin *et al.*, 2020). Some tools provide non-parametric analyses through non-R interfaces, including the Python packages bletl (Osthege *et al.*, 2022), phenom (Tonner *et al.*, 2020), and AMiGA (Midani *et al.*, 2021), and the graphical user interface-based QurvE (Wirth and Funk, 2023). And some scriptable R packages provide non-parametric estimation of specific traits, including area under the curve by growthcurver (Sprouffske and Wagner, 2016), and maximum growth rate by growthrates (Petzoldt, 2022). gcplyr builds on these approaches to quantify many different traits non-parametrically, including maximum growth rate, lag time, area under the curve, carrying capacity, diauxie, and more.

All growth curve analyses, including the model-free and non-parametric analyses implemented by gcplyr, have limitations. First, analyses rely critically on the input data. Most frequently, growth curves are collected by automated spectrophotometers, where optical density is used as a proxy for microbial cell density [although see (Mackie *et al.*, 2014) for an alternate proxy for cell density]. However, optical density has limited sensitivity to detect low concentrations of microbial cells, and typically saturates before maximum concentrations of cells are reached (Biesta-Peters *et al.*, 2010). These limitations constrain any analysis of optical density-based growth curves, but may especially affect non-parametric analyses. Non-parametric analyses can also be more sensitive to measurement error than parametric methods. To account for this, gcplyr includes smoothing methods that can reduce the effects of noise, although smoothing is not a panacea for measurement error. Finally, non-parametric approaches require

more fine-grained temporal resolution for precise quantification of microbial growth traits. In all, these limitations highlight the relative strengths and weaknesses of parametric and non-parametric analyses. Parametric methods can be the best approach when population dynamics closely follow known mathematical models of microbial growth and users can validate that model fitting converged on appropriate solutions, while the non-parametric analyses enabled by gcplyr enable flexible quantification of traits in a broader range of conditions.

Growth curves are a widespread experimental approach in the microbial sciences. However, until now, tools capable of wrangling and doing model-free analysis of growth curve data were limited. By enabling these steps, gcplyr streamlines high-quality model-free analysis for a wide range of applied and fundamental research on microbial growth.

Acknowledgements

Thanks to Paul Turner for support during this project, to Jyot Antani, Alita Burmeister, Noah Houpt, and Jordan Lewis for feedback on drafts of this paper, and to all the gcplyr alpha and beta testers.

References

Biesta-Peters,E.G. *et al.* (2010) Comparison of Two Optical-Density-Based Methods and a Plate Count Method for Estimation of Growth Parameters of Bacillus cereus. *Applied and Environmental Microbiology*, **76**, 1399–1405.

Blazanin,M., Moore,J.P., *et al.* (2023) Fight not flight: parasites drive the bacterial evolution of resistance, not avoidance. 2023.04.29.538831.

Blazanin,M., Vasen,E., *et al.* (2023) Theoretical validation of growth curves for quantifying phage-bacteria interactions. *bioRxiv*.

Bukhman,Y.V. *et al.* (2015) Modeling Microbial Growth Curves with GCAT. *Bioenergy Research*, **8**, 1022–1030.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, **74**, 829–836.

Cleveland,W.S. and Devlin,S.J. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, **83**, 596–610.

Coutin,N.P.J. *et al.* (2020) Interactively AUDIT Your Growth Curves with a Suite of R Packages. *G3*, **10**, 933–943.

Csibra,E. and Stan,G.-B. (2023) Parsley: a web app for parsing data from plate readers. *Bioinformatics*, btad733.

Cuevas,D.A. *et al.* (2016) Elucidating genomic gaps using phenotypic profiles. *F1000Research*, **3**, 1–28.

Cuevas,D.A. and Edwards,R.A. (2017) PMAnalyzer: A new web interface for bacterial growth curve analysis. *Bioinformatics*, **33**, 1905–1906.

Fernandez-Ricaud,L. *et al.* (2016) PRECOG: A tool for automated extraction and visualization of fitness components in microbial growth phenomics. *BMC Bioinformatics*, **17**, 249.

Garre,A. *et al.* (2022) biogrowth: Modelling of Population Growth.

Hall,B.G. *et al.* (2014) Growth rates made easy. *Molecular Biology and Evolution*, **31**, 232–238.

Hemmerich,J. *et al.* (2017) Automated growth rate determination in high-throughput microbioreactor systems. *BMC Research Notes*, **10**, 617.

Huang,L. (2014) IPMP 2013 — A comprehensive data analysis tool for predictive microbiology. *International Journal of Food Microbiology*, **171**, 100–107.

Jung,P.P. *et al.* (2015) Protocols and Programs for High-Throughput Growth and Aging Phenotyping in Yeast. *PLoS ONE*, **10**, e0119807.

Liu,Y. *et al.* (2021) Microrisk Lab: An Online Freeware for Predictive Microbiology. *Foodborne Pathogens and Disease*, **18**, 607–615.

Mackie,A.M. *et al.* (2014) Biolog Phenotype MicroArrays for Phenotypic Characterization of Microbial Cells. In, Paulsen,I.T. and Holmes,A.J. (eds), *Environmental Microbiology: Methods and Protocols*, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 123–130.

Midani,F.S. *et al.* (2021) AMiGA: Software for Automated Analysis of Microbial Growth Assays. *mSystems*, **6**, e00508-21.

Monod,J. (1949) The Growth of Bacterial Cultures. *Annual Review of Microbiology*, **3**, 371–394.

Olsen,B. *et al.* (2010) YODA: Software to facilitate high-throughput analysis of chronological life span, growth rate, and survival in budding yeast. *BMC Bioinformatics*, **11**, 141.

Osthege,M. *et al.* (2022) bletl - A Python package for integrating BioLector microcultivation devices in the Design-Build-Test-Learn cycle. *Engineering in Life Sciences*, **22**, 242–259.

Petzoldt,T. (2022) growthrates: Estimate Growth Rates from Experimental Data.

R Core Team (2022) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Savitzky,A. and Golay,M.J. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, **36**, 1627–1639.
Sprouffske,K. and Wagner,A. (2016) Growthcurver: An R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics*, **17**, 17–20.
Swain,P.S. *et al.* (2016) Inferring time derivatives including cell growth rates using Gaussian processes. *Nat Commun*, **7**, 13766.
Swinnen,I.A.M. *et al.* (2004) Predictive modelling of the microbial lag phase: a review. *International Journal of Food Microbiology*, **94**, 137–159.
Tonner,P.D. *et al.* (2020) A Bayesian non-parametric mixed-effects model of microbial growth curves. *PLOS Computational Biology*, **16**, e1008366.
Tonner,P.D. *et al.* (2017) Detecting differential growth of microbial populations with Gaussian process regression. *Genome Res.*, **27**, 320–333.
Trevor Hastie and Robert Tibshirani (1986) Generalized Additive Models. *Statistical Science*, **1**, 297–310.
Vaas,L.A.I. *et al.* (2013) Opm: An R package for analysing OmniLog® phenotype microarray data. *Bioinformatics*, **29**, 1823–1824.
Veríssimo,A. *et al.* (2013) BGFit: Management and automated fitting of biological growth curves. *BMC Bioinformatics*, **14**.
Vervier,K. *et al.* (2019) CarboLogR: a Shiny/R application for statistical analysis of bacterial utilisation of carbon sources. 695676.
Wickham,H., François,R., *et al.* (2023) dplyr: A Grammar of Data Manipulation.
Wickham,H. (2016) ggplot2: Elegant Graphics for Data Analysis Springer-Verlag New York.
Wickham,H. Tidy Data. *Journal of Statistical Software*, **10**.
Wickham,H., Vaughan,D., *et al.* (2023) tidyr: Tidy Messy Data.
Wickham,H. *et al.* (2019) Welcome to the tidyverse. *Journal of Open Source Software*, **4**, 1686.
Wirth,N.T. and Funk,J. (2023) QurvE: Robust and User-Friendly Analysis of Growth and Fluorescence Curves.
Wood,S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, **73**, 3–36.
Wood,S.N. (2017) Generalized Additive Models: An Introduction with R 2nd ed. Chapman and Hall/CRC.
Wood,S.N. (2003) Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, **65**, 95–114.
Yates,G.T. and Smotzer,T. (2007) On the lag phase and initial decline of microbial growth curves. *Journal of Theoretical Biology*, **244**, 511–517.

Table S1. gcplyr and other available microbial growth curve data wrangling and analysis tools.

Software	Used via	Param. analysis	Non-param. analysis	Scriptable	Imports	Data reshaping	Incorporate designs	Plotting	Smoothing	Citation
gcplyr	R package	No	Yes	Yes	Block, wide, tidy	Yes, into wide or tidy	Yes	With ggplot2	Yes	This paper
growthcurver		Yes	Minimally	Yes	Wide	No	No	Built-in	No	(Sprouffske and Wagner, 2016)
growthrates			Minimally	Yes	Tidy	No	No	With ggplot2	Yes	(Petzoldt, 2022)
biogrowth			No	Yes	Wide	No	No	Built-in	No	(Garre *et al.*, 2022)
opm			Minimally	Yes	Specialized formats	Yes, into opm-specific class	Yes	Built-in	Yes	(Vaas *et al.*, 2013)
QurvE	GUI or R package	Yes	Yes	Yes	Wide, specialized formats	From specialized formats into wide	Yes	Built-in	Yes	(Wirth and Funk, 2023)
AUDIT	GUI	Yes	No	Yes, in part	Tidy, specialized formats	From specialized formats into tidy	Yes	Built-in, with ggplot2	Yes	(Coutin *et al.*, 2020)
Parsley		No	No	No	Block, wide	Yes, into tidy	Yes	No	No	(Csibra and Stan, 2023)
bletl	Python package	Minimally	Yes	Yes	Specialized formats	Yes	Yes	Built-in	Yes	(Osthege *et al.*, 2022)
AMiGA	Python package or command line	No	Yes	Some	Wide	No	Yes	Built-in	Yes	(Midani *et al.*, 2021)
fitderiv	GUI or Python package	No	Yes	Some	Wide	No	No	Built-in	Yes	(Swain *et al.*, 2016)
phenom	Python code	No	Yes	Yes	Wide	No	Yes	No	Yes	(Tonner *et al.*, 2020)
B-GREAT			Yes	Yes	Wide	No	Yes	No	Yes	(Tonner *et al.*, 2017)
PMAnalyzer	bash	Yes	Minimally	No	Wide	No	No	Built-in	No	(Cuevas and Edwards, 2017; Cuevas *et al.*, 2016)
GrowthRates	Command line	No	Yes	No	Wide, specialized formats	Yes	No	No	No	(Hall *et al.*, 2014)
-----	MATLAB code	No	Yes	Yes	One well at a time	No	No	No	No	(Hemmerich *et al.*, 2017)
GCAT	GUI	Yes	No	No	Wide	No	Yes	Built-in	No	(Bukhman *et al.*, 2015)
PRECOG		No	Yes	No	Wide	No	No	Built-in	Yes	(Fernandez-Ricaud *et al.*, 2016)
IPMP 2013		Yes	No	No	Wide	No	No	Built-in	No	(Huang, 2014)
GATHODE		No	Yes	No	Wide	No	Minimally	Built-in	Yes	(Jung *et al.*, 2015)
Microrisk Lab		Yes	No	No	Wide	No	No	Built-in	No	(Liu *et al.*, 2021)
CarboLogR		Yes	No	No	Specialized formats	No	Yes	Built-in	No	(Vervier *et al.*, 2019)
YODA	Webpage GUI	No	Yes	No	Wide	No	No	No	No	(Olsen *et al.*, 2010)
BGFit		Yes	No	No	Wide	No	No	Built-in	No	(Veríssimo *et al.*, 2013)