Appendix B: The GOA TOOLs paper

B.1 Over 1,700 bioinformatics projects ranked

In 2018, Russell et al. defined the impact of individual open-source bioinformatics projects on the research community [183]. They searched for peer-reviewed research papers describing bioinformatics projects hosted on GitHub, finding 1,720 such projects. They found that the bottom 28% (482) of the papers were not cited by any articles in PubMed Central, 40% (688) were cited by 1 to 5 articles, and that 32% (550) of the papers have more than five citations.

Russell et al. highlighted 23 projects out of 1,720 as being particularly high profile and crucial in the research community. One of the open-source projects featured in this thesis, GOA TOOLs, sits among the top 1% of high-profile projects, such as samtools [40] and bedtools [173], based on the parameters identified by Russell of being associated with a impactful project, as shown in Figures B.1 and B.2.

GOA TOOLs is a Python library and tool suite for managing gene ontology (GO) terms and running gene ontology enrichment analyses (GOEA) and is freely available at https://github.com/tanghaibao/goatools. GO terms describe the biological functions of a gene product.

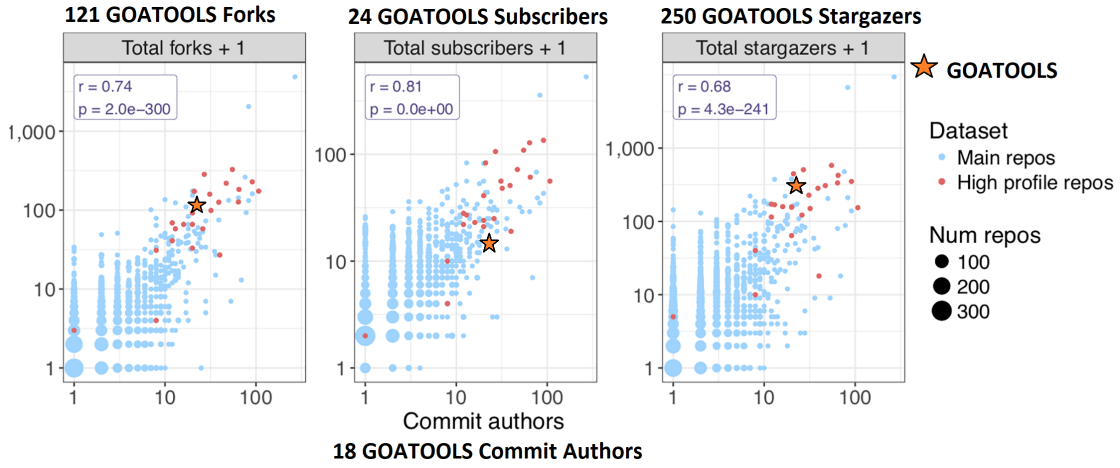

Figure B.1: **GOA TOOLs is a top project based on three parameters from data collected by GitHub.** Russell found that the three GitHub features shown here correlate to an impactful bioinformatics project. The x and y axes for each of the three panels are in a log 10 scale, resulting in over several hundred projects in the lower left-hand corner and only tens of high-profile projects in the upper right-hand corner. GOA TOOLs is comfortably placed in the upper right-hand corner. A "fork" occurs in GitHub when another researcher makes a copy of a project. A "watcher" is a researcher who signs up for email when there are project discussions. A "stargazer" is a researcher who has signaled that they like the project by "starring" it in GitHub. This figure was adapted with permission from Figure 4 in Russell et al. [183] using GOA TOOLs GitHub data from 2018, the same publication year of Russell's paper.

Repo Name	Lang.	Mb	Stars	Forks	Watchers
biopython	Python	49	1761	914	144
samtools	C	12	762	385	107
bedtools2	C	47	526	210	64
igv	Java	410	342	162	48
GOATOOLS	Python	34	250	121	24
bowtie2	C++	153	240	82	28
vcftools	C++	1	225	96	27
bcftools	C	14	211	122	48
cufflinks	C++	491	193	101	42
htsjdk	Java	72	184	209	50
tophat	C++	11	80	38	19

Figure B.2: **GOA TOOLS is in the top 1% of bioinformatics projects.** GOA TOOLs sits comfortably among the top 23 of 1,720 bioinformatics projects, such as samtools [40], bedtools [173], and tophat [103], when using the criteria determined by Russell et al. to correlate to an impactful bioinformatics project. Important criteria include the size of the project, the number of GitHub stars given by the research community, and other GitHub features such as forks and watchers. The data for this table are from the supplemental data of Russell et al. [183]. This table uses GOA TOOLs GitHub data from 2018, the same publication year of Russell's paper.

B.2 Researcher contributions and community interest

Researchers show their appreciation for an open-source project on GitHub by awarding it stars. The rate of researcher appreciation of the GOA TOOLs project increased substantially between the beginning of the project in January 2010 (when it was awarded its first star) and the publication date of this thesis in June 2021 (with 432 stars).

Table B.1 shows that GOA TOOLs has three distinct appreciation periods divided by two events. The first event in 2016 marks the beginning of my adding functionality to GOA TOOLs for the paper while writing the paper (Table B.1: b). The second event is the publication of the paper in the journal *Scientific Reports* in 2018 [106] (Table B.1: c).

Table B.1: **The rate of researcher appreciation rose 1,000% following D.V. Klopfenstein's contributions.** The letters a, b, c, and d in the start time and end time columns indicate consequential dates in the GOA TOOLs project. There are three rating periods in the table. The GOA TOOLs project was awarded 2.47 stars per 100 days before my major code contributions. The rate increased by 500% from 2.47 to 12.42 when I began committing new functionality for the GOA TOOLs paper. The rate increased 1,000% from 2.47 to 25.07 after the publication of the paper.

Stars per 100 days	Start time	End time	Time period description
2.47	Jan 2010 (a)	Jan 2016 (b)	Beginning of project to paper proposal
12.42	Jan 2016 (b)	Jul 2018 (c)	Paper proposal to paper publication
25.07	Jul 2018 (c)	Jun 2021 (d)	Paper publication to June 2021

The rate of appreciation increased fivefold from 2.47 to 12.42 stars per 100 days after I proposed to the project's owner, Dr. Haibao Tang, that I write a

peer-reviewed research paper centering on GOA TOOLs and began adding the functionality needed for the paper. The rate of appreciation doubled from 12.42 to 25.07 stars per 100 days after the publication of the paper for a total tenfold increase the original appreciation rating of 2.47.

Figure B.3 shows researcher interest in GOA TOOLs (top panel) along with my contributions (bottom panel). In the top panel, researcher interest is shown using short vertical blue lines to represent the date when researchers added stars to the GOA TOOLs project. The three rating periods are divided by the dashed magenta lines (Figure B.3: b and c).

The bottom panel shows my additions of code, called commits, as the principal contributor out of a total of 19 GOA TOOLs contributors. Plots of the code contributions from all 19 contributors are available on the GOA TOOLs website (https://github.com/tanghaibao/goatools/graphs/contributors).

The GitHub stars rose from 55 (Figure B.3: b) to 169 (Figure B.3: c) solely due to researchers noticing my new functionality in GitHub. There was no announcement of any improvements using any channels such as papers or social media posts. Researchers noticed the new functionality and immediately responded by quintupling the rate of stars awarded.

Dr. Tang agreed to my adding the functionality for the paper in January 2016. My first major commits in support of the paper (Figure B.3: U) show my first contributions of the new functionality to GOA TOOLs that was created for this thesis.

For example, my advisor wanted the multiple GOEA *p*-values to be corrected using the Benjamini-Hochberg multiple test correction, which was not available in GOA TOOLs before 2016. Consequently, I added support for

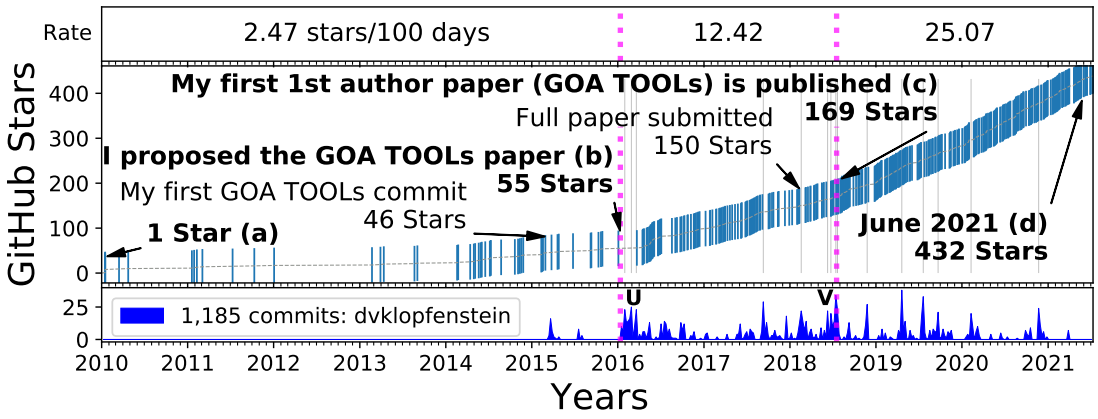

Figure B.3: **The amount of researcher interest in GOA TOOLs rose rapidly following the creation of commits by D. V. Klopfenstein starting in 2016.** The top panel shows the interest of researchers in GOA TOOLs. The bottom panel shows the code contributions from the principal lead, D. V. Klopfenstein. The text and arrows in the top panel describe key project dates. The dotted magenta lines divide the figure into three time periods. The bold text ending with the letters a, b, c, and d mark consequential events (Table B.1). Each short vertical blue line in the top panel represents a date when a researcher added a star to GOA TOOLs, illustrating their appreciation of the project. The light gray lines in the top panel show the top 15 weeks of commit activity out of over 1,100 total commits by D. V. Klopfenstein. The x-axis is time in months and years and is shared by both panels. The y-axis of the top panel is the total number of GitHub stars given to the project, which in spring 2021 was over 400. The bottom panel shows new features contributed to GOA TOOLs by its top contributor. The number of commits appears on the y-axis. The letters U and V indicate periods of commit activity just after I suggested writing the paper (U) and just before the paper was published (V).

Benjamini-Hochberg, which is available through the statsmodels Python package [189], as well as access to all multiple test corrections available through the statsmodels package.

My collaborators wanted the GOEA results in an Excel spreadsheet format. I added the functionality (Figure B.3: U) while improving it by setting the default cell (column) widths to researcher-friendly values. For example, the column width is wider for lengthy columns like "name" and shorter for columns like "*p*-value."

The numerous contributions that I submitted in the first quarter of 2016 were followed by a jump in researcher interest as seen in the major acceleration of stars awarded to GOA TOOLs in the second quarter of 2016.

Because researchers expressed desire for additional functionality in GO plotting and because I used my augmented GO plotting to create figures for the paper, I released the improved GO plotting before publication, as shown by the spikes of commit activity in 2017 and early 2018 before the paper was published. I delayed submitting the novel functionality, which included an innovative novel GO grouping, until just before publication (Figure B.3: V).

After publication, I continued to add new functionalities to GOA TOOLs that were desired by the research community as expressed in the issues section of the project. Examples of new functionalities include robust support for the evidence codes which describe the proof (evidence) that a biological function is correctly attributed to the gene product to which it is annotated. Examples of evidence codes include the functionality determined by "experimental evidence" or "computationally determined." I also added support for using optional relationships between GO terms such as "part of" and "regulates" to augment the required relationship of "is a."

B.3 GOEA stochastic simulations

Dr. Tang suggested that I create original stochastic GOEA simulations and interpret the results in the paper to make the paper more captivating.

The stochastic GOEA simulation code and commits, results, data, and figures represent a project separate from the GOA TOOLs project and are freely available in the GitHub repository, (https://github.com/dvklopfenstein/goatools_simulation).

I began the simulation project by creating a more straightforward set of simulations, which include the Benjamini-Hochberg and Bonferroni multiple test corrections, to establish the proof-of-concept of the stochastic simulations; architect baseline implementations of simulation code; and explore methods to visualize the resulting statistical data in figures.

I built upon my successful deployment of the multiple test correction stochastic simulations to create the more complicated GOEA simulations. The simulation project uses the GOA TOOLs project as a prerequisite for the GOEA simulations. I am the sole contributor to the simulation repository.