INCOMPLETE DRAFT: RNA –k-mers–> Protein

This manuscript (permalink) was automatically generated from czbiohub/kmerslay-paper@094e600 on March 7, 2020.

Authors

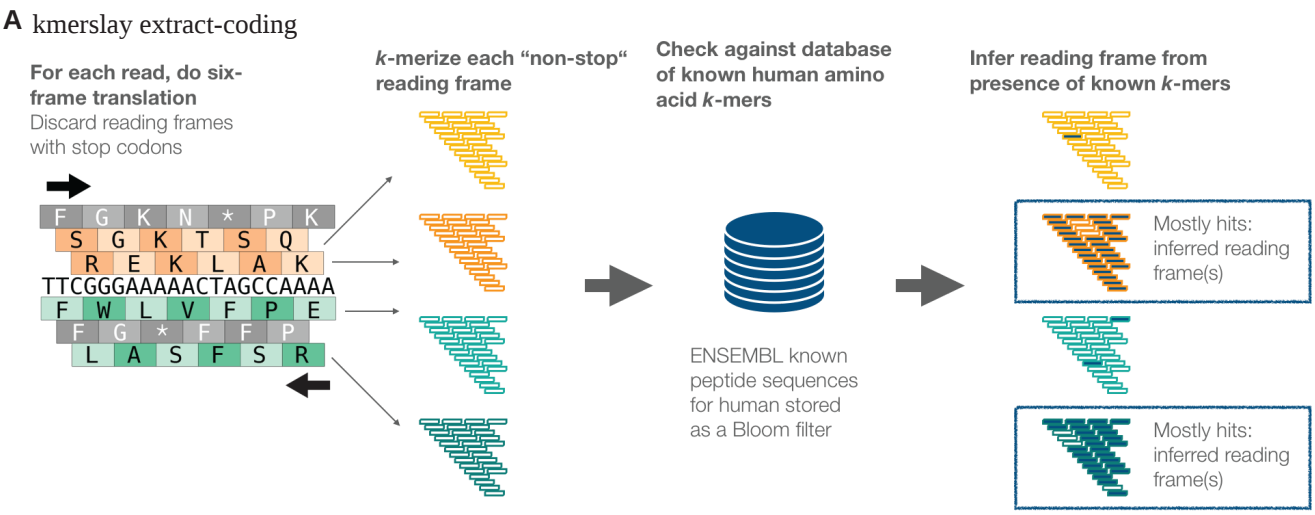
- **Olga Borisovna Botvinnik**
  0000-0003-4412-7970 · olgabot · olgabot
  Data Sciences Platform, Chan Zuckerberg Biohub

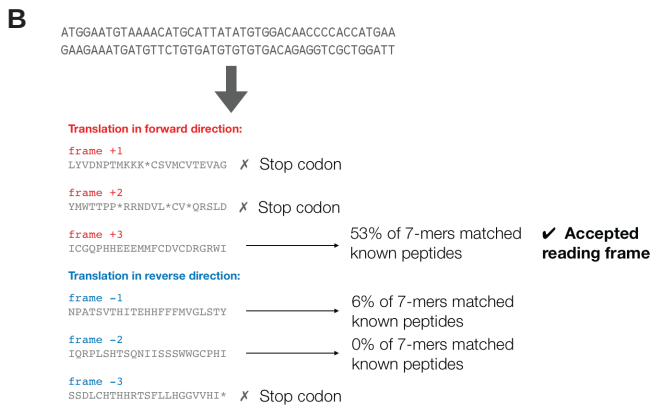
- **N. Tessa Pierce**
  0000-0002-2942-5331 · bluegenes
  Department of Population Health and Reproduction, University of California, Davis · Funded by NSF 1711984

Abstract

RNA-sequencing is a widely used measure of cellular state, and it is often used as proxy for estimating abundance of protein-coding molecules. However, the tools to estimate protein-coding sequence from RNA-seq reads don't allow for extraction of protein-coding sequences from distantly related species. We present `kmerslay`, a suite of k-mer based tools to extract protein-coding sequences from RNA-seq reads using reduced amino acid alphabets to allow for flexibility in proteome evolution. `kmerslay` can also perform all-by-all k-mer similarity of sequences in both amino acid and nucleotide reduced alphabets to find clusters of similar sequences which can then be used for homology inference. By enabling extraction of protein-coding sequences across a wide variety of species, `kmerslay` spearheads analysis of non-model organisms and their contribution to the evolution of life.

Outline

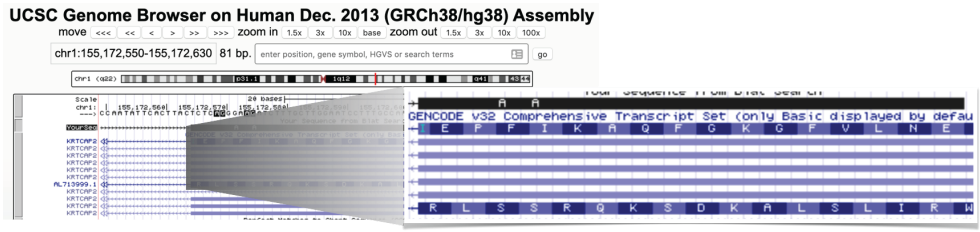




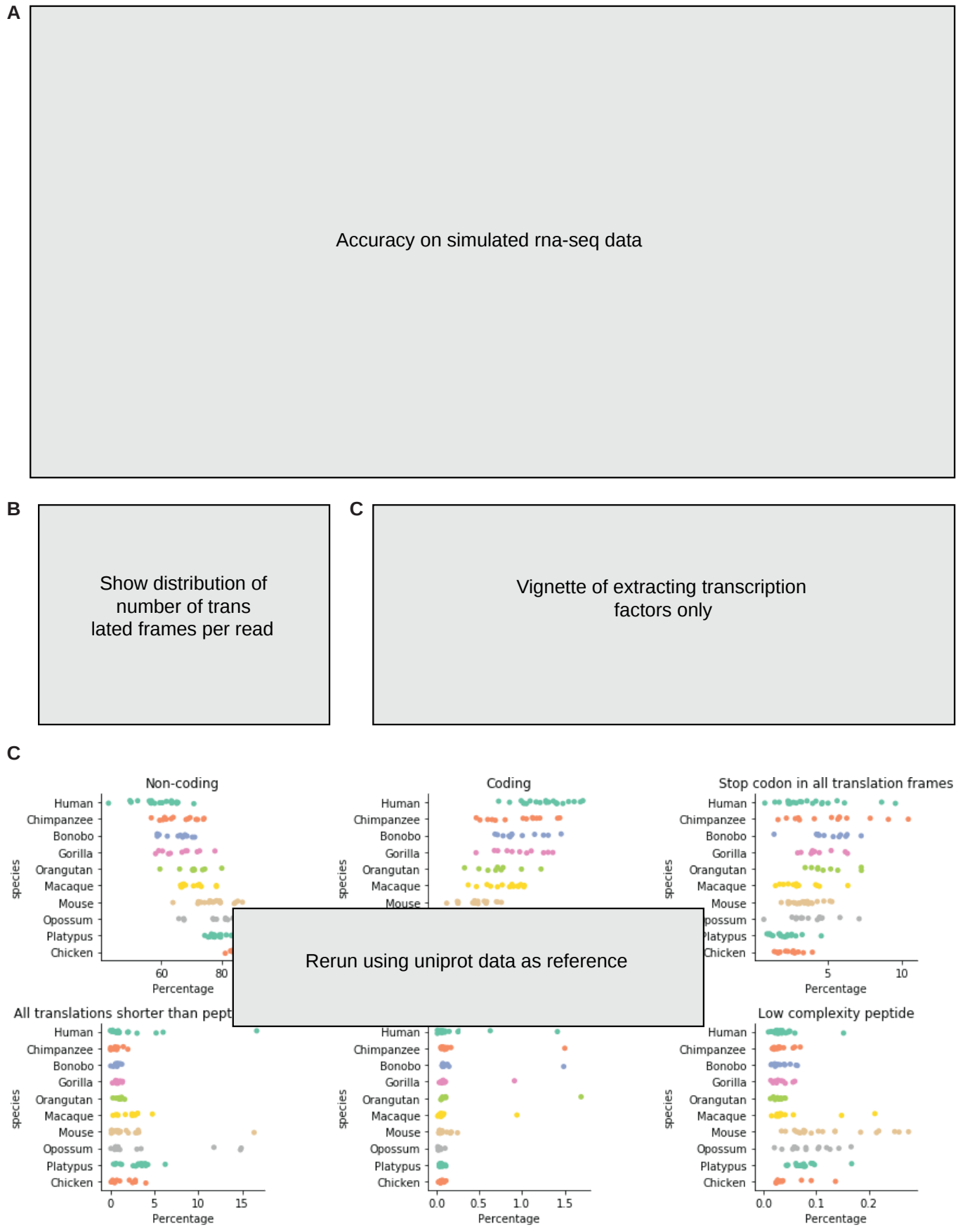
C

```
>A00111:133:H3VGJDSXX:1:2362:27805:31454 2:N:0:TTTGACAGGCTG+TCATTACATCAT
CTCGGTGGATGAGGCCAGATGCAAAGAGAGCCAACAGGAGGCAGAGGAGAATCTCAGGAAAAATCTTTGCTTGGAATCCTTTGCCAAAGACAAGATTCTC

>A00111:133:H3VGJDSXX:1:2362:27805:31454 2:N:0:TTTGACAGGCTG+TCATTACATCAT translation_frame: -1 jaccard: 1.0
ENLVFGKGFQAKIFPEILLCLLLALFASGLIHR
>A00111:133:H3VGJDSXX:1:2362:27805:31454 2:N:0:TTTGACAGGCTG+TCATTACATCAT translation_frame: -2 jaccard: 0.5185185185185185
RILSLAKDSKQRFFLRFSSASCWLSLHLASSTE
```

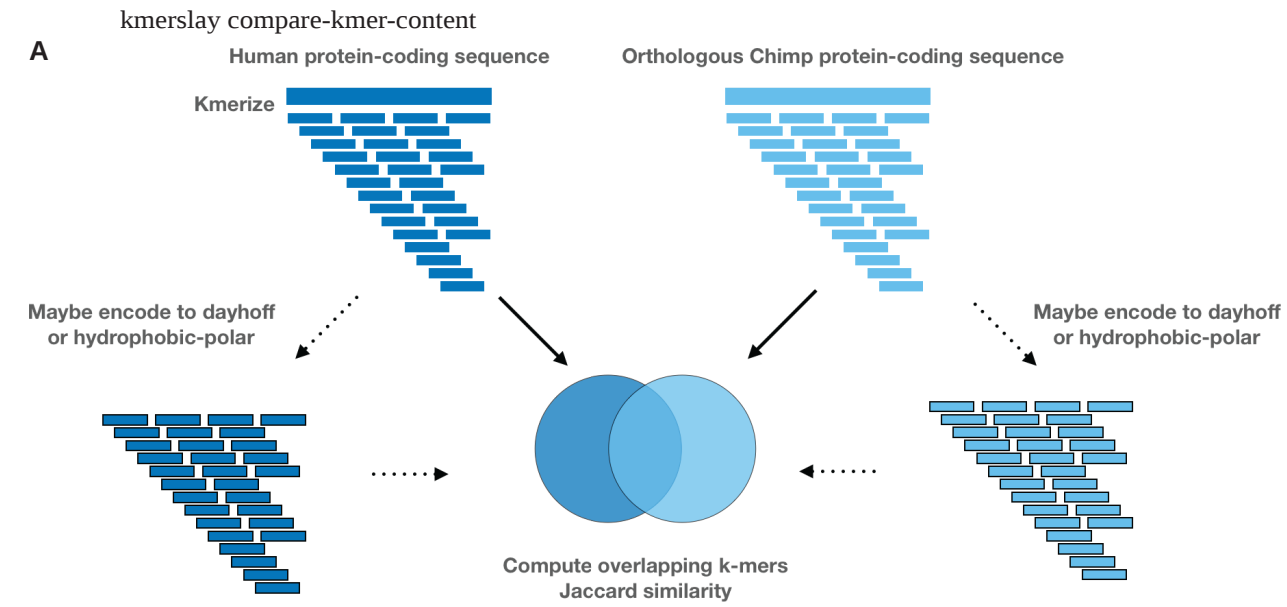


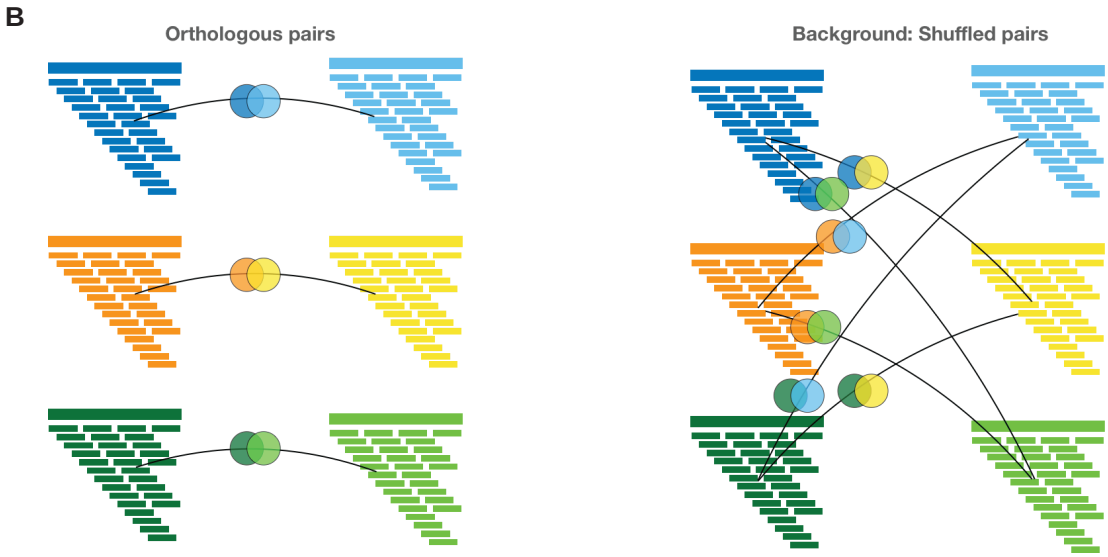
Overview of `kmerslay extract-coding` **A.** First, each read is translated into all six possible protein-coding translation frames. Next, reading frames with stop codons are eliminated. Each protein-coding frame is k -merized, then the fraction of k-mers which appear in the known protein-coding database is computed. Frames which contain a fraction of coding frames exceeding the threshold are inferred to be putatively protein-coding. **B.** Worked example of an RNA-seq read with a single putatitive reading frame. **C.** Worked example of an RNA-seq read with multiple reading frames, and a UCSC genome browser shot of the read showing that both reading frames are present in the annotation.



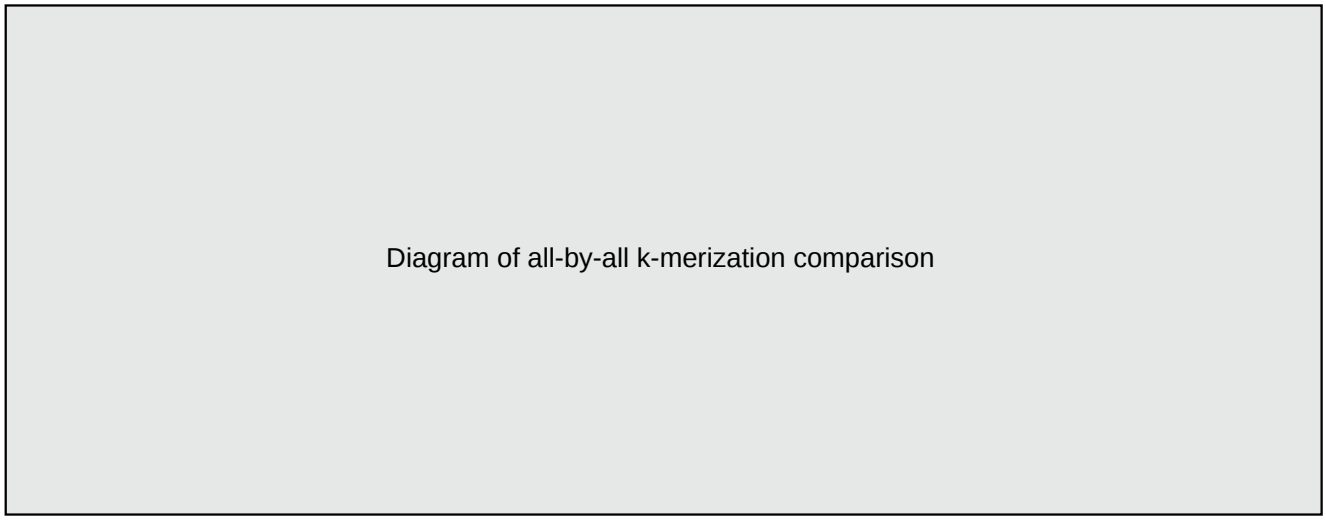
Applications of `kmerslay extract-coding`. **A.** We simulated RNA-seq data using Opisthokonta species from the Quest for Orthologs dataset for true positive protein-coding RNAs, reads completely contained within intergenic, intronic, and UTR sequences as true positive noncoding RNAs, and reads partially overlapping a coding and noncoding region as an adversarial test set. We then predicted protein-coding sequences and computed false positive and false negative rates. False Positive coding reads were found to be … False negative noncoding reads were found to be … **B.** Number of putative protein-coding sequences per read. **C.** This method could also be used to extract only reads whose

putative protein-coding sequences are transcription factors. **D.** We ran `kmerslay extract-coding` on the five tissues and nine species from the Brawand 2011 dataset.





C



Overview of `kmerslay compare-kmer-content` **A.** Protein sequences are k -merized by converting into a bag of words using a sliding window of size k, potentially re-encoded to a lossy alphabet, and then their fraction of overlapping k-mers is computed into a Jaccard similarity. **B.** One option for `kmerslay compare-kmer-content` is to specify a pair

of sequence files, and compute a background of k-mer similarty using randomly shuffled pairs. **C.** Another option for `kmerslay compare-kmer-content` is to do an all-by-all k-mer similarity comparison.

Applications of `kmerslay compare-kmer-content`. **A.** We used `kmerslay compare-kmer-content` on pairs of orthologous protein sequences between humans and the remaining Opisthokonta species in the Quest for Orthologs dataset. x-axis, k -mer size, y-axis, mean difference. **B.** False positive calls by `kmerslay compare-kmer-content` are either paralogs or read-through protein products. **C.** We applied `kmerslay compare-kmer-content` to ... to find putative orthologs. We found ... the accuracy was ...

Dayhoff and HP alphabets

Table 1: Dayhoff and hydrophobic-polar encodings are a reduced amino acid alphabet allowing for permissive cross-species sequence comparisons. For example, the amino acid sequence `SASHAFIERCE` would be Dayhoff-encoded to `bbbdbfecdac`, and HP-encoded to `phpphhhpppp`, as below.

Amino acid	Property	Dayhoff	Hydrophobic-polar (HP)
C	Sulfur polymerization	a	p
A, G, P, S, T	Small	b	A, G, P: h
			S,T: p
D, E, N, Q	Acid and amide	c	p
H, K, R	Basic	d	p
I, L, M, V	Hydrophobic	e	h
F, W, Y	Aromatic	f	h

```
protein20: SASHAFIERCE
dayhoff6:  bbbdbfecdac
hp2:       phpphhhpppp
```

All implemented alphabets (with citations, not as nicely organized)

Citation	Alphabet	Amino acid groups
Phillips, R., Kondev, J., Theriot, J., & Garcia, H. (2012). Physical Biology of the Cell. Garland Science.	hp2	AFGILMPVWY CDEHKNQRST
Peterson, E. L., Kondev, J., Theriot, J. A., & Phillips, R. (2009). Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics, 25(11), 1356–1362. http://doi.org/10.1093/bioinformatics/btp164	gbmr4	G ADKERNTSQ YFLIVMCWH P
Dayhoff, M. O., & Eck, R. V. (1968). Atlas of Protein Sequence Structure, 1967-68.	dayhoff6	AGPST HRK DENQ FWY ILMV C

Citation	Alphabet	Amino acid groups
This paper	botvinnik8	AG DE RK NQ ST FY LIV CMWHP
Hu, X., & Friedberg, I. (2019). SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier. GigaScience, 8(10), 309–12. http://doi.org/10.1093/gigascience/giz118	aa9	G AST KR EQ DN CFILMVY W H P
Peterson, E. L., Kondev, J., Theriot, J. A., & Phillips, R. (2009). Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics, 25(11), 1356–1362. http://doi.org/10.1093/bioinformatics/btp164	sdm12	G A D KER N TSQ YF LIVM C W H P
Peterson, E. L., Kondev, J., Theriot, J. A., & Phillips, R. (2009). Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics, 25(11), 1356–1362. http://doi.org/10.1093/bioinformatics/btp164	hsdm17	G A D KE R N T S Q Y F LIV M C W H P
Dayhoff, M. O., & Eck, R. V. (1968). Atlas of Protein Sequence Structure, 1967-68.	protein20	G A D E K R N T S Q Y F L I V M C W H P

Introduction

Methods

Results

Discussion

Bold text

Semi-bold text

Centered text

Right-aligned text

Italic text

Combined *italics and **bold***

~~Strikethrough~~

1. Ordered list item
2. Ordered list item
    a. Sub-item
    b. Sub-item

        i. Sub-sub-item
3. Ordered list item
    a. Sub-item

- List item
- List item
- List item

subscript: H_2O is a liquid

superscript: 2^{10} is 1024.

unicode superscripts^{0123456789}

unicode subscripts$_{0123456789}$

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to editing and version control.

Line break without starting a new paragraph by putting
two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

A heading centered on its own printed page

Horizontal rule:

`Heading 1`'s are recommended to be reserved for the title of the manuscript.

`Heading 2`'s are recommended for broad sections such as *Abstract*, *Methods*, *Conclusion*, etc.

`Heading 3`'s and `Heading 4`'s are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah

Link with text

Link with hover text

Link by reference

Citations

Citation by DOI [1].

Citation by PubMed Central ID [2].

Citation by PubMed ID [3].

Citation by Wikidata ID [4].

Citation by ISBN [5].

Citation by URL [6].

Citation by tag [7].

Multiple citations can be put inside the same set of brackets [1,5,7]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [2,3,7,8].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure 1

Figure 2

Figure 3

Figure 4

Table 1

Equation 1

Equation 2

Quotes and code

> Quoted text

> Quoted block of text
>
> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

Code `in the middle` of normal text, aka `inline code`.

Code block with Python syntax highlighting:

```
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

Figures

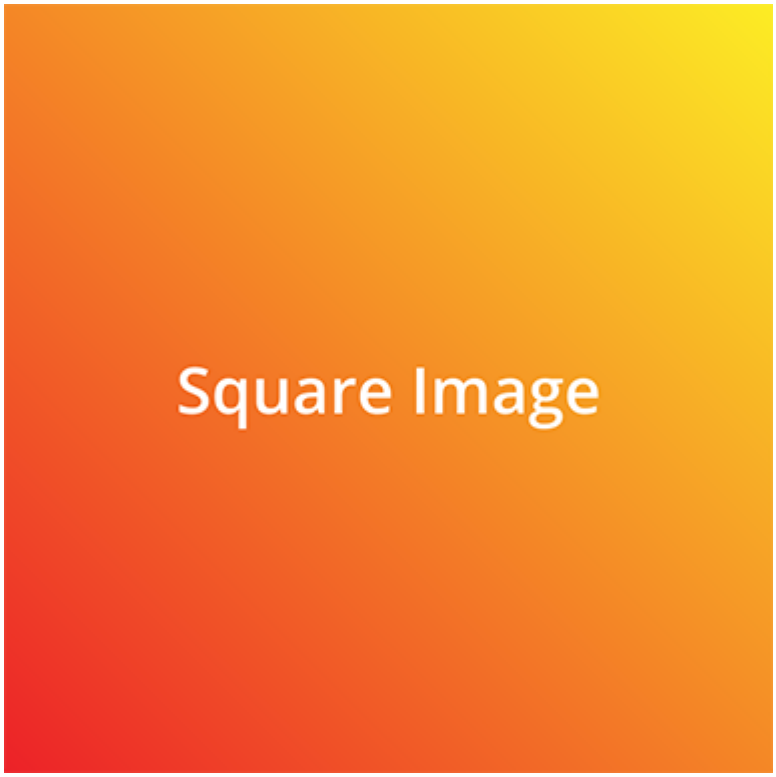


Figure 1: A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.

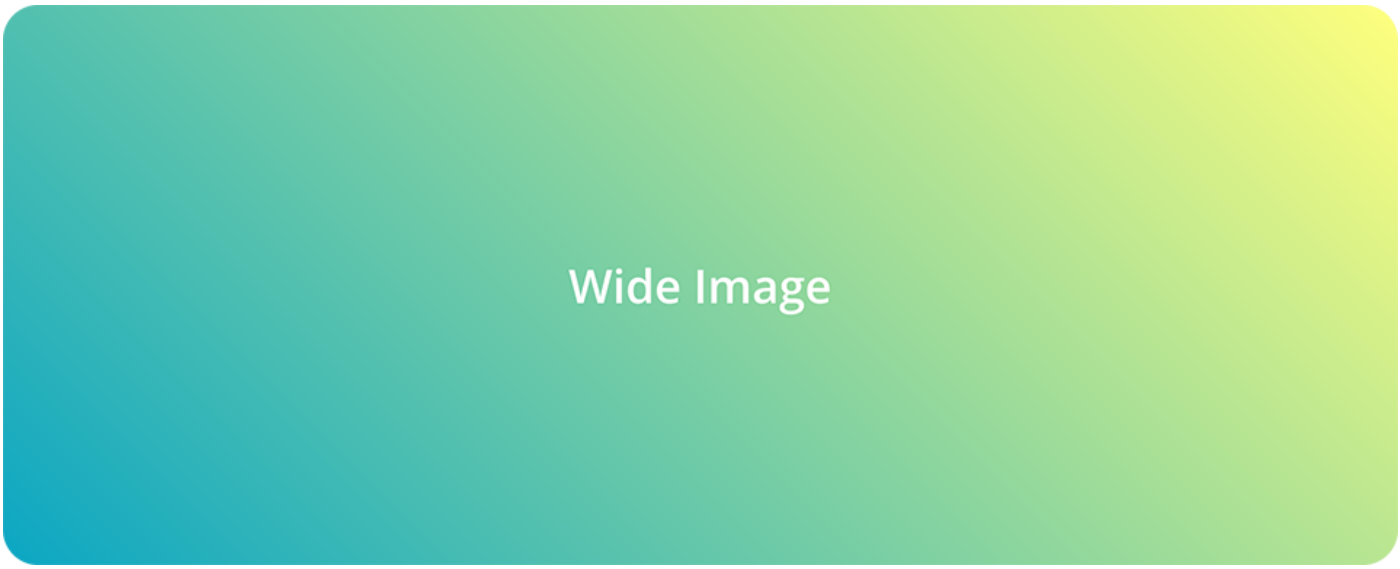


Figure 2: An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.

Figure 3: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.

Figure 4: A vector `.svg` image loaded from GitHub. The parameter `sanitize=true` is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image.

Tables

Table 1: A table with a top caption and specified relative column widths.

Bowling Scores	**Jane**	**John**	**Alice**	**Bob**
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

Table 2: A table too wide to fit within page.

	Digits 1-33	**Digits 34-66**	**Digits 67-99**	**Ref.**
pi	3.14159265358979323 846264338327950	28841971693993751 0582097494459230	78164062862089986 2803482534211706	`piday.org`
e	2.71828182845904523 536028747135266	24977572470936999 5957496696762772	40766303535475945 7138217852516642	`nasa.gov`

Table 3: A table with merged cells using the `attributes` plugin.

	Colors	
Size	**Text Color**	**Background Color**
big	blue	orange
small	black	white

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \tag{1}$$

An equation too long to fit within page:

$$x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 \tag{2}$$

Special

⚠ **WARNING** *The following features are only supported and intended for* `.html` *and* `.pdf` *exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as* `.docx`*.*

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot `attributes` plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Available background colors for text, images, code, banners, etc:

`white` `lightgrey` `grey` `darkgrey` `black` `lightred` `lightyellow` `lightgreen` `lightblue` `lightpurple` `red` `orange` `yellow` `green` `blue` `purple`

Using the Font Awesome icon set:

✔ ? ★ 🔔 ⊗ •••

Light Grey Banner
useful for *general information* - manubot.org

Blue Banner
useful for *important information* - manubot.org

Light Red Banner
useful for *warnings* - manubot.org

Supplementary

References

1. **Sci-Hub provides access to nearly all scholarly literature**
   Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene
   eLife (2018-03-01) https://doi.org/ckcj
   DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

2. **Reproducibility of computational workflows is automated using continuous analysis**
   Brett K Beaulieu-Jones, Casey S Greene
   Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
   DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMC6103790

3. **Bitcoin for the biological literature.**
   Douglas Heaven
   Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888
   DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

4. **Plan S: Accelerating the transition to full and immediate Open Access to scientific publications**
   cOAlition S
   (2018-09-04) https://www.wikidata.org/wiki/Q56458321

5. **Open access**
   Peter Suber
   MIT Press (2012)
   ISBN: 9780262517638

6. **Open collaborative writing with Manubot**
   Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
   Manubot (2020-01-14) https://greenelab.github.io/meta-review/

7. **Opportunities and obstacles for deep learning in biology and medicine**
   Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, … Casey S. Greene
   Journal of The Royal Society Interface (2018-04-04) https://doi.org/gddkhn
   DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

8. **Open collaborative writing with Manubot**
   Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
   PLOS Computational Biology (2019-06-24) https://doi.org/c7np
   DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653