

Smashing single cells into k -mer sketches

This manuscript ([permalink](#)) was automatically generated from [czbiohub/primate-brain-organoid-paper@e74a513](#) on November 5, 2019.

Authors

- Venkata Naga Pranathi Vemuri

 [0000-0002-5748-9594](#) ·  [pranathivemuri](#) ·  [pranuvemuri](#)

Data Sciences Platform, Chan Zuckerberg Biohub

Abstract

Single-cell RNA-sequencing is a powerful technology for identifying novel and known cell types, however its power is limited to organisms with well-annotated genomes. We demonstrate the utility of using annotation-agnostic methods which quantify cell-cell similarity using k -mer profiles. We benchmark a few methods and demonstrate the utility of converting cell types from mouse to human and back, and compare to using purely 1:1 mapped orthologous genes.

Introduction

There are a predicted 8.7 million Eukaryotic species on earth [1], yet only 14% (1,233,500) have been catalogued and 0.001% (9,449) have genomes deposited in the National Center for Biotechnology Information Genome Assembly [2]. And yet, the genome sequence is not enough. To truly understand the diversity of life on this planet, we need to determine not just the DNA blueprints of life, but understand the instantiation of the DNA, the cell types of the species. While sequencing DNA gives a quantitative measure of the nucleotide differences, it does not inform the functional strategies that change with DNA sequence. As new species can be defined by a new cell type. For example, the existence of a single cell type, the Cnidocyte [3], a stinging cell of a single-celled biological weapon, defines the phylum Cnidaria. Thus, entire clades, not only species, can be defined by the introduction of an additional cell type or state.

Novel organizations of existing cell states can also define cell types. For example, the development of genitalia in amniotes, while using similar cell types, ultimately uses a different physical organization of cell types to generate genitalia in mammals compared to reptiles [4]

Determining common gene ancestry (“orthology”) is a difficult problem. Many approaches exist [5].

Determining common ancestry of cell types (“orthologous cell types”) [6] is an additional difficult problem. Comparative transcriptomics begins with finding a common feature set for embedding molecular profiles across divergent species into a common space. Many researchers take the approach of using one-to-one orthologous genes [Cite: brawand2011, CCA, LIGER, Scanorama, basically all the single cell “alignment” packages], others use clusters of orthologous groups [8], others map reads onto a common genome derived from whole-genome alignment [cite: recent primate brain paper from Barbara Treutlein], or map onto native genomes [9] and re-annotate using a tool such as Comparative Annotation Toolkit [10].

k -mers have been proposed for comparing single cells [11] as they are a fast, simple way to create cell-cell similarities. However, the work so far has focused on using annotated organisms and not cross-species analyses.

Methods

Methods go here.

We used ENSEMBL version 97.

Results

To determine whether short segments of sequences could detect gene orthologues, we k -merized orthologous genes derived from the ENSEMBL version 97 [12] COMPARA database [13] (Fig.~). We compared human protein sequences to orthologous chimpanzee, mouse, (orangutan, bonobo, gorilla, macaque, opossum, platypus, chicken) protein sequences, as these are species used in [14]. As a background, we randomly chose 10 non-orthologous genes relative to the human gene. In addition to k -merizing the protein-coding sequence, we also re-encoded the protein-coding sequence into Dayhoff [15] and hydrophobic-polar encodings [16], show in Table~.

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 `bbdbfecdac`, and HP-encoded to `phpphhpppp`.

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

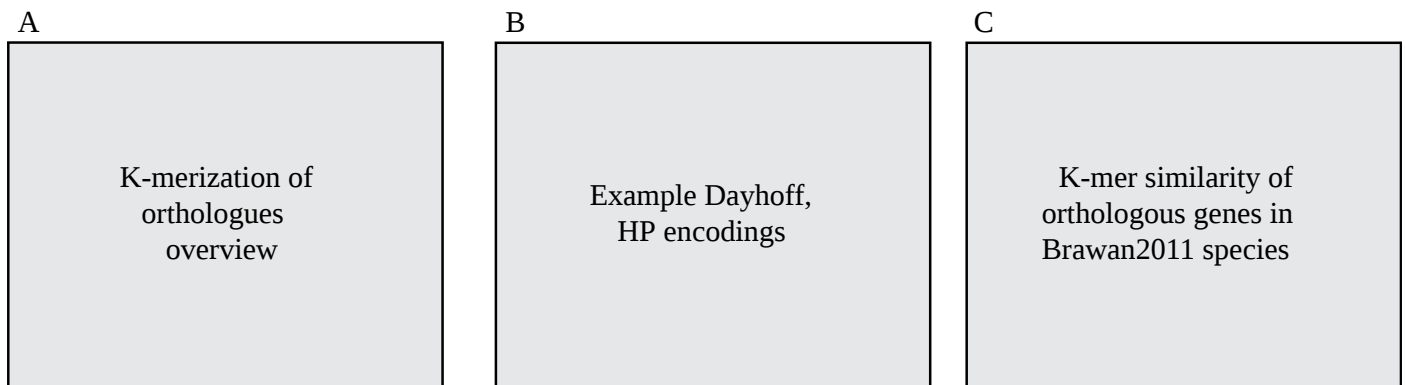


Figure 1: Figure 1.

We found that, consistent with previous knowledge, that 1:1 orthologues had higher k -mer similarities as determined by the Jaccard Index.

Additionally, more recently diverged genes had higher k -mer similarity as well.

A

Extract coding sequence using human proteins,
k-merize coding sequence and minhash, com
pare samples/cells across minhashes

B

kNN graph of
Brawand2011
MinHashes

C

kNN graph of
Brawand2011
Gene expression

D

K-mers driving
similarity in
brawand2011

E

Are the k-mers from
unmapped reads or
unannotated genes?

Figure 2: Figure 2.

Discussion

Conclusions and future directions go here.

References

1. How Many Species Are There on Earth and in the Ocean?

Camilo Mora, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, Boris Worm
PLoS Biology (2011-08-23) <https://doi.org/fpr4z8>
DOI: [10.1371/journal.pbio.1001127](https://doi.org/10.1371/journal.pbio.1001127) · PMID: [21886479](https://pubmed.ncbi.nlm.nih.gov/21886479/) · PMCID: [PMC3160336](https://pubmed.ncbi.nlm.nih.gov/PMC3160336/)

2. Genome List - Genome - NCBI <https://www.ncbi.nlm.nih.gov/genome/browse/>

3. Cnidocyte

Wikipedia

(2019-10-20) <https://en.wikipedia.org/w/index.php?title=Cnidocyte&oldid=922123291>

4. A relative shift in cloacal location repositions external genitalia in amniote evolution

Patrick Tschopp, Emma Sherratt, Thomas J. Sanger, Anna C. Groner, Ariel C. Aspiras, Jimmy K. Hu, Olivier Pourquié, Jérôme Gros, Clifford J. Tabin
Nature (2014-11-05) <https://doi.org/f6sg74>
DOI: [10.1038/nature13819](https://doi.org/10.1038/nature13819) · PMID: [25383527](https://pubmed.ncbi.nlm.nih.gov/25383527/) · PMCID: [PMC4294627](https://pubmed.ncbi.nlm.nih.gov/PMC4294627/)

5. Standardized benchmarking in the quest for orthologs

Adrian M Altenhoff, Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, Benjamin Linard, Cécile Pereira, ... Christophe Dessimoz
Nature Methods (2016-04-04) <https://doi.org/f3rpzx>
DOI: [10.1038/nmeth.3830](https://doi.org/10.1038/nmeth.3830) · PMID: [27043882](https://pubmed.ncbi.nlm.nih.gov/27043882/) · PMCID: [PMC4827703](https://pubmed.ncbi.nlm.nih.gov/PMC4827703/)

6. The origin and evolution of cell types

Detlev Arendt, Jacob M. Musser, Clare V. H. Baker, Aviv Bergman, Connie Cepko, Douglas H. Erwin, Mihaela Pavlicev, Gerhard Schlosser, Stefanie Widder, Manfred D. Laubichler, Günter P. Wagner
Nature Reviews Genetics (2016-11-07) <https://doi.org/f9b62x>
DOI: [10.1038/nrg.2016.127](https://doi.org/10.1038/nrg.2016.127) · PMID: [27818507](https://pubmed.ncbi.nlm.nih.gov/27818507/)

7. How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology

John C. Marioni, Detlev Arendt
Annual Review of Cell and Developmental Biology (2017-10-06) <https://doi.org/ggb632>
DOI: [10.1146/annurev-cellbio-100616-060818](https://doi.org/10.1146/annurev-cellbio-100616-060818) · PMID: [28813177](https://pubmed.ncbi.nlm.nih.gov/28813177/)

8. The COG database: a tool for genome-scale analysis of protein functions and evolution

R. L. Tatusov
Nucleic Acids Research (2000-01-01) <https://doi.org/fr3ggz>
DOI: [10.1093/nar/28.1.33](https://doi.org/10.1093/nar/28.1.33) · PMID: [10592175](https://pubmed.ncbi.nlm.nih.gov/10592175/) · PMCID: [PMC102395](https://pubmed.ncbi.nlm.nih.gov/PMC102395/)

9. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution

Alex A. Pollen, Aparna Bhaduri, Madeline G. Andrews, Tomasz J. Nowakowski, Olivia S. Meyerson, Mohammed A. Mostajo-Radji, Elizabeth Di Lullo, Beatriz Alvarado, Melanie Bedolli, Max L. Dougherty, ... Arnold R. Kriegstein
Cell (2019-02) <https://doi.org/gfvgzr>
DOI: [10.1016/j.cell.2019.01.017](https://doi.org/10.1016/j.cell.2019.01.017) · PMID: [30735633](https://pubmed.ncbi.nlm.nih.gov/30735633/) · PMCID: [PMC6544371](https://pubmed.ncbi.nlm.nih.gov/PMC6544371/)

10. Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation

Ian T. Fiddes, Joel Armstrong, Mark Diekhans, Stefanie Nachtweide, Zev N. Kronenberg, Jason G. Underwood, David Gordon, Dent Earl, Thomas Keane, Evan E. Eichler, ... Benedict Paten

Genome Research (2018-06-08) <https://doi.org/gdpg7n>
DOI: [10.1101/gr.233460.117](https://doi.org/10.1101/gr.233460.117) · PMID: [29884752](https://pubmed.ncbi.nlm.nih.gov/29884752/) · PMCID: [PMC6028123](https://pubmed.ncbi.nlm.nih.gov/PMC6028123/)

11. K-mer counting with low memory consumption enables fast clustering of single-cell sequencing data without read alignment

Christina Huan Shi, Kevin Y. Yip

bioRxiv (2019-08-02) <https://www.biorxiv.org/content/10.1101/723833v1>

DOI: [10.1101/723833](https://doi.org/10.1101/723833)

12. Ensembl 2018

Daniel R Zerbino, Premanand Achuthan, Wasii Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, ... Paul Flicek

Nucleic Acids Research (2017-11-16) <https://doi.org/gcwg6r>

DOI: [10.1093/nar/gkx1098](https://doi.org/10.1093/nar/gkx1098) · PMID: [29155950](https://pubmed.ncbi.nlm.nih.gov/29155950/) · PMCID: [PMC5753206](https://pubmed.ncbi.nlm.nih.gov/PMC5753206/)

13. Ensembl comparative genomics resources

Javier Herrero, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, Stephen M. J. Searle, Ridwan Amode, Simon Brent, ... Paul Flicek

Database (2016) <https://doi.org/ggb9tv>

DOI: [10.1093/database/bav096](https://doi.org/10.1093/database/bav096) · PMID: [26896847](https://pubmed.ncbi.nlm.nih.gov/26896847/) · PMCID: [PMC4761110](https://pubmed.ncbi.nlm.nih.gov/PMC4761110/)

14. The evolution of gene expression levels in mammalian organs

David Brawand, Magali Soumillon, Anamaria Necșulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, ... Henrik Kaessmann

Nature (2011-10) <https://doi.org/fcvk54>

DOI: [10.1038/nature10532](https://doi.org/10.1038/nature10532) · PMID: [22012392](https://pubmed.ncbi.nlm.nih.gov/22012392/)

15. Atlas of protein sequence and structure

Margaret O Dayhoff

National Biomedical Research Foundation. (1969)

16. Physical biology of the cell

Rob Phillips, Julie Theriot, Jane Kondev, Hernan Garcia

Garland Science (2012)