INCOMPLETE DRAFT: Smashing single cells into k-mer sketches

This manuscript (permalink) was automatically generated from czbiohub/orthology-free-comparative-transcriptomics-paper@9ad7a8b on December 17, 2019.

Authors

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

- **Venkata Naga Pranathi Vemuri**
  0000-0002-5748-9594 · pranathivemuri · pranuvemuri
  Data Sciences Platform, Chan Zuckerberg Biohub

- **Phoenix Aja Logan**
  0000-0003-4581-0552 · phoenixAja · phoenixlogan
  Data Sciences Platform, Chan Zuckerberg Biohub

- **Saba Nafees**
  0000-0002-3292-7703 · snafees · sabanafeesTTU
  Data Sciences Platform, Chan Zuckerberg Biohub; Department of Biological Sciences, Texas Tech University; Department of Mathematics & Statistics, Texas Tech University

Abstract

Single-cell RNA-sequencing is a powerful technology for identifying novel and known cell types. However, the task of identifying new and novel cell types across species is nontrivial, especially when one or more of the species have poorly annotated genomes. Thus, there is an unmet need to quantitatively compare single-cell transcriptomes across species, without the need for a reference genome. To this end, we have developed a genome-agnostic method to compare molecular profiles using a lossy encoding on k-mers from putative protein-coding RNA-seq reads. We benchmark the annotation-agnostic methods on a bulk comparative transcriptomic dataset consisting of nine species and six tissues, showing that we can recapitulate the results as using only reads mapping to 1:1 orthologs of protein-coding genes, and we are able to resolve ... which can only be seen by using the reference-free k-mer method. We then show that k-mers can also be used for comparing transcriptomes built from long read sequencing, by comparing the cell-cell similarity nearest neighbor graphs built on k-mers from short reads and long reads from the same cells in a primate brain organoid system. We find that using k-mers on short reads has the advantage of resolving ... in comparison to using read counts from 1:1 gene orthologs, while long reads provide additional information in the form of ... Using k-mers, we were able to resolve cell types X in the primate brain organoid dataset, which was hidden using read counts alone. Thus, we have show the reference-free methods using the k-mers from single cells is a novel, annotation-agnostic method for comparing cells across species that is capable of identifying cell states unique to a particular organism, helping to build the cell type evolution tree of life.

Introduction

There are a predicted 8.7 million Eukaryotic species on earth [1], yet only 14% (1,233,500) have been catalogued and 0.000002% (200/8,700,000 = 2.3e-08) have genomes present in ENSEMBL Assemblies (as of ENSEMBL 98 – September 2019 release) [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 from DNA modifications due to speciation events . As new species can be defined by a new cell type. For example, the existence of a single cell type, the stinging cell called a "Cnidocyte" [3], 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. However, it is unclear how many examples there are of this, and how it is possible to find cell types that are unique to one species. Thus, we aim to develop a computational method to compare cell types across species that is reference-agnostic and can find cell types that are novel and present in only one organism.

Organizations of existing cell states can also define novel organismal structures. For example, different physical organizations of similar cell types generate different genitalia in amniotes when comparing mammals to reptiles [4]

Determining common gene ancestry ("orthology") is a difficult problem.

Many approaches exist, reviewed by [5,6,7]. Generally, the approaches are structured in this way: (1) find orthologous groups of genes, (2) build gene trees, (3) build species trees, and (4) assign orthologs, as described in a recent approach (Orthofinder) [8]. In this approach, we are not interested in exactly reconstructing the species or gene trees, but rather inferring function based on cell type transcriptomes. Instead of exactly building the gene trees, we subset the protein-coding sequences into peptide words, and re-encode to lossy peptide encodings.

Determining common ancestry of cell types ("orthologous cell types") [9,10] 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 [11,12,13,14,15,16], others use clusters of orthologous groups [19], others map reads onto a common genome derived from whole-genome alignment [20,21], or map onto native genomes [22] and re-annotate using a tool such as Comparative Annotation Toolkit [23].

Annotating one dataset's cell types from another can be performed using random forest models trained on the original dataset [24], using correlation between cell gene expression profiles as in Cell BLAST [25], locality-sensitive hashing of bit vectors of gene expression as in CellFishing.jl [26], or using a cell type hierarchy as in Garnett [27] and OnClass [28].

k-mers have been proposed for comparing single cells [29] as they are a fast, simple way to create cell-cell similarities. k-mers have also been used for orthologous gene detection [30] However, the work so far has focused on using annotated organisms and not cross-species analyses.

Reduced amino acid alphabets have been previously used to speed up database searches [31,32], protein fold prediction [33,34], and homology recognition [35]. We aim to find "orthologous reads" across species' transcriptomes. By representing each species' transcriptome as the set of k-mers, we can unbiasedly compare transcriptomes in an orthologous space without the need for knowing the orthologous genes ahead of time, or even the need for a reference genome. Additionally, we do not need to reduce the signal to only the genes with a 1:1 orthologous match.

Methods

Methods go here.

Experimental

Primate brain organoid protocols

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

Single-cell capture of primate brain organoids

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

Long read library prep

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

Short read library prep

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

Sequencing

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

Computational

k-mer comparison of orthologous genes

We used ENSEMBL version 97. We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

Extraction of putative coding reads from RNA-seq

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

$bam2fasta$ conversion

The `.bam` file generated by the Drop-seq [36] pipeline for the different primates in this study are in the order of 6-12 GB. The Drop-seq `.bam` files so obtained can attribute to few limitations as discussed below. Firstly, loading them in memory all at once would require a lot of RAM depending on how the program will allocate memory for different data typed tags in the `.bam` file. Secondly, if Drop-seq data is not accompanied by a barcodes file to filter the `.bam` file on, it would mean we would have to recursively go through the alignments in the bam file and deduce alignments with higher quality and combine sequences with already exisiting barcodes. This would need a look up

dictionary to be updated as it loops through the alignments in the `.bam` file and would search the look up dictionary as it updates the barcodes. In conclusion, this is a very memory intensive process that seemed to fail on even machines with 2TB RAM.

Hence we propose a method that could work on a computer with lesser RAM and not cause computer hangups. We released an open source pypi package for the same [37]. The package contains solution for the above discussed problem by sharding the `.bam` file into chunks of smaller `.bam` files and stores them in the machine's temporary folder, e.g. `/tmp`. The chunk size of the `.bam` file is a tunable parameter that can be accessed with `--line_count`; by default it is 1500 alignment lines. This process is done serially by iterating through the alignments in the `.bam` file, using `pysam`, a Python wrapper around samtools [38]. Now we employ a MapReduce [39] approach to the temporary `.bam` files to obtain all the reads per cell barcode in a `.fasta` file. In the "Map" step, we distribute the computation i.e parsing the barcode, determining the quality of the read, and if alignment is not duplicated, in parallel across multiple processes on the temporary shards of `.bam` files. These bam shards create temporary `.fasta` files that contain for each read: the cell barcode, unique molecular identifier (UMI), and the aligned sequence. There might be a cell barcode that would be present in different chunks of these sharded `.bam` files. As a result we would have multiple temporary `.fasta` files for the same barcodes. We implemented a method to find the unique barcodes based on these temporary `.fasta` file names and then assigning each of the unique barcodes all the temporary barcode `.fasta` files created by different `.bam` shards in a dictionary. In the "Reduce" step, we concatenate of strings of temporary `.fasta` file names, hence its memory consumption is less than it would be if appending to a list. These temporary `.fasta` files are then combined to one `.fasta` file per barcode by concatenating all the sequences obtained from different `.fasta` files. The concatenation of all sequences for each of the unique barcodes is also then parallelized to use multiple processes. For each of the cell barcodes, there is an option to obtain valid cell barcodes, based on the UMI count per cell barcode. For our datasets we have set the minimum number of UMIs per cell barcode to 1000, a common threshold. The minimum number of UMIs per cell barcode can be customized with the flag `--min-umi-per-barcode`. The computational resources and time taken for processing is as shown in Table [1].

Table 1: Human primate species bam file here is from a brain organoid for human

Primate	BAM file size(GB)	Time(hrs)	RAM(GB)	Processes
Human	12	7	16	32
Orangutan	9	4	16	32
Chimp	9	4	16	32

This method primarily gives us time performance improvement. It reduces time from days or just process running out of memory to hours. Depending on the size of `.bam` file and resources of the cluster/computer it can be further reduced.

Prefer DOI for references, but for Biorxiv use the URL. DOI example: [5]. Biorxiv example: [8]. Multiple citations per line example: [5,8].

Results

Figure 1 – k-mers are sufficient to detect orthologous genes

To determine whether short segments of sequences could detect gene orthologues, we k-merized orthologous genes derived from the ENSEMBL version 97 [40] COMPARA database [41] (Figure [1]). We compared human protein sequences to orthologous chimpanzee, mouse, (orangutan, bonobo,

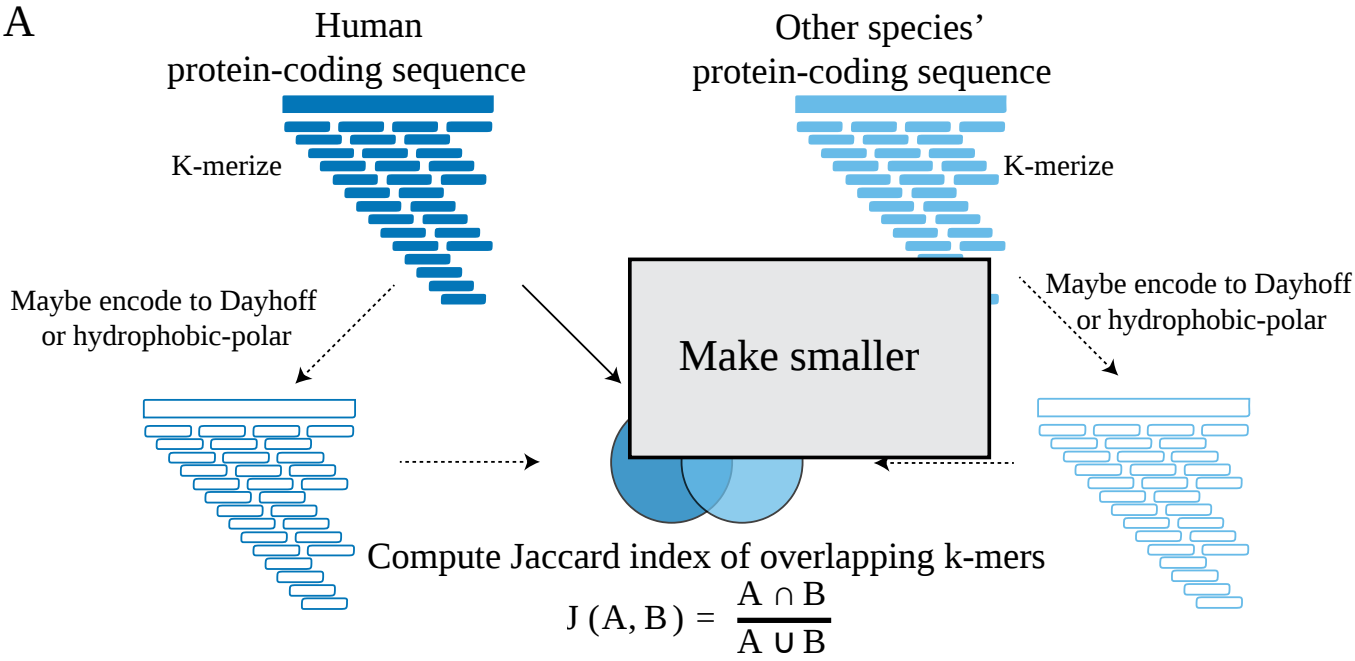
gorilla, macaque, opossum, platypus, chicken) protein sequences, as these are species used in [11]. As a background, we randomly chose 10 non-orthologous genes relative to the human gene. In addition to k-merizing the protein-coding sequence, we also re-encoded the protein-coding sequence into a six-letter Dayhoff alphabet [42], a nine-letter encoding [30], and a two-letter hydrophobic-polar encodings [43,44], show in Table [3].

Table 2: 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`.

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

k-mer size may be tuned to find an “optimal” length of protein domains across the tree of life. Protein domain lengths follow a power law distribution where proteins with more domains, have shorter domains, whereas proteins with fewer domains, have fewer but longer domains [45,46].

A



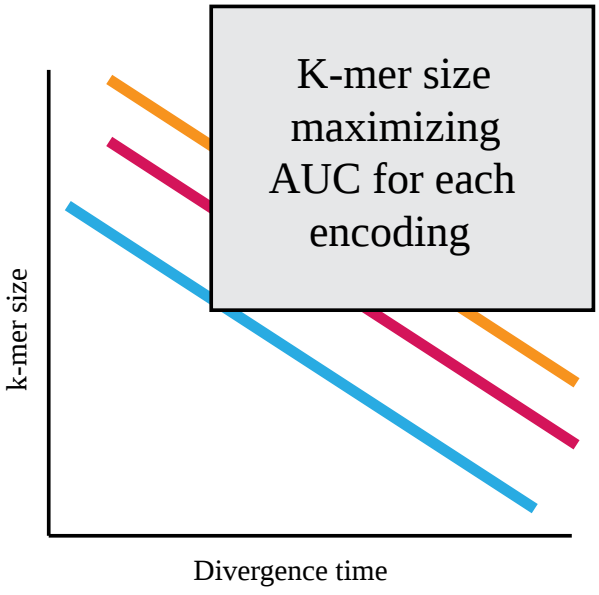
B

C

Update to Brawand2011 species, show diff between background

Update to Brawand2011 species

D



E

Placeholder

Figure 1: **A.** Overview of k-mer comparison of orthologous genes. The protein-coding sequence of each pair of known orthologs is k-merized, potentially encoded as Dayhoff or Hydrophobic polar, and then the Jaccard index (the intersection divided by the union) is computed on the k-mers. **B.** Jaccard similarity of orthologous genes in Dayhoff-encoded k-mer space relative to humans in eight species. x-axis, k-mer size; y-axis, Jaccard index. **C.**

We found that, consistent with previous knowledge, that 1:1 orthologues had higher k-mer similarities as determined by the Jaccard Index. This approach is similar to SwiftOrtho [30], a k-mer based orthology relationship finder.

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

Outline

- Kmers can approximate orthologies
  - Jaccard similarity of orthologues is higher than non-orthologues
  - Benchmarking using https://orthology.benchmarkservice.org/cgi-bin/gateway.pl
  - Finding orthologues
    - Gold standard
      - ENSEMBL COMPARA
      - Quest for Orthologs consortium, Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., et al. (2016). Standardized benchmarking in the quest for orthologs. Nature Methods, 13(5), 425–430. http://doi.org/10.1038/nmeth.3830 [5]
    - Orthologous groups/Conserved Domain Database [47]

Figure 2 – k-mers from lossily-encoded putative protein-coding reads faithfully pull out reads from protein-coding genes within amniotes

A

B

C

Figure 2: Figure 2.

- Overview of kmermaid pipeline
  - Comparison of tissue across species
    - Partition reads to coding/noncoding bins
    - MinHash the Dayhoff-encoded coding sequences
    - Jaccard similarity on the MinHashes
- Which reads are found to have coding features but didn't map to the genome?
- Do these features map to novel genes or gene fusions?
- Kmers can find correct reading from of RNA-seq reads

- Human peptides → human, chimp, bonobo, orangutan, gorilla, macaque, mouse, opossum, playtpus, chicken RNAseq from Brawand2011 data
- Comparison to other methods: RNASamba [48]

Figure 3 – k-mers can pull out only reads from transcription factors and Amniotes can be compared on the MinHashes of their protein-coding sequences

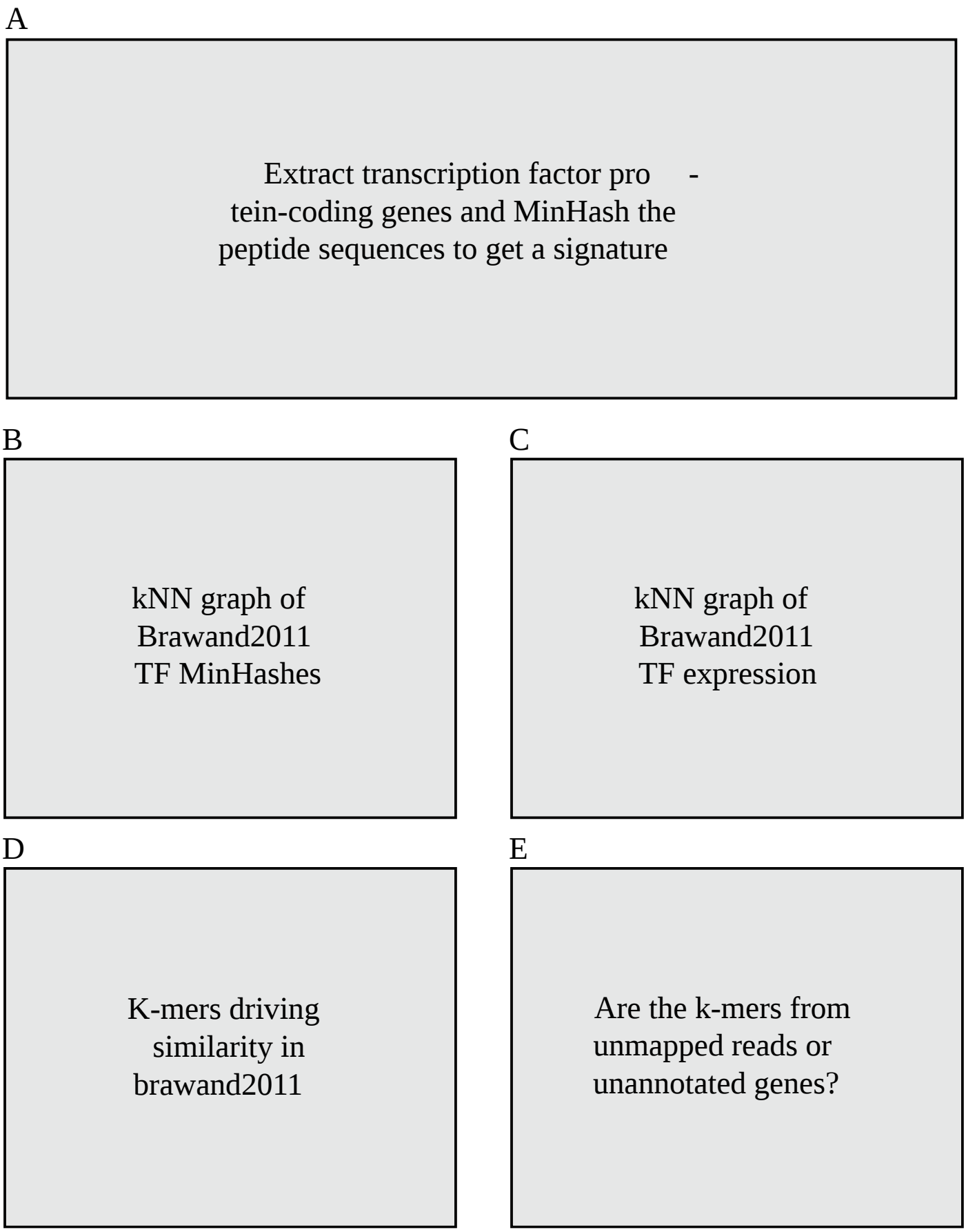


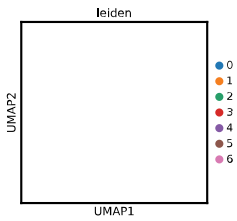
Figure 3: Figure 3.

- Kmers can find only transcription factor reads of TFs from RNA-seq reads
  - Human peptides → human, chimp, bonobo, orangutan, gorilla, macaque, mouse, opossum, playtpus, chicken RNAseq from Brawand2011 data

Figure 4 – k-mers can compare short and long read datasets in primate brain organoids

A

B



kNN/UMAP of
primate organoid
long reads
MinHashes

Figure 4: **A.** Overview of experimental system. Poly-A RNA molecules from single cells from primate brain organoids were captured using the Dolomite system, where molecule received a cell barcode and molecular barcode, was reverse transcribed and primed for full-length cDNA. Then the library was split for sequencing on the Illumina or PacBio platforms. For the Illumina platform, the library was first sheared to be compatible with the platform. **B**. UMAP of short read gene counts from Human (left), Chimp (middle), and Orangutan (right) organoids. **C.** UMAP of short read MinHashes with ksize=33, dayhoff encoding and a log2sketchsize=14. **D**. UMAP of long read MinHashes with ksize=33, dayhoff encoding, andl log2sketchsize=14. **E.** Examples of short reads with ambiguous coding sequence resolved by long reads. **F.** UMAP on short-read gene counts of 1:1 orthologous transcription factor genes across all species. **G.** UMAP on short-read MinHashes of all k-mers that match human transcription factor protein-coding peptide sequences. **H.** UMAP on long-read MinHashes of all k-mers that match human transcription factor protein-coding peptide sequences.

References for Primate Brain development

- Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., et al. (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. eLife, 7, D635. http://doi.org/10.7554/eLife.32332 [49]
- Mazin, P. V., Jiang, X., Fu, N., Han, D., Guo, M., Gelfand, M. S., & Khaitovich, P. (2018). Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques. Rna, 24(4), 585–596. http://doi.org/10.1261/rna.064931.117 [50]
- Xiong, J., Jiang, X., Ditsiou, A., Gao, Y., Sun, J., Lowenstein, E. D., et al. (2018). Predominant patterns of splicing evolution on human, chimpanzee and macaque evolutionary lineages. Human Molecular Genetics, 27(8), 1474–1485. http://doi.org/10.1093/hmg/ddy058 [51]

Discussion

Conclusions and future directions go here.

Some potential references

Gene expression evolution through duplications

- Farre, D., & Alba, M. M. (2010). Heterogeneous Patterns of Gene-Expression Diversification in Mammalian Gene Duplicates. Molecular Biology and Evolution, 27(2), 325–335. http://doi.org/10.1093/molbev/msp242 [52]
- Thornton, J. W., & DeSalle, R. (2000). Gene family evolution and homology: genomics meets phylogenetics. Annual Review of Genomics and Human Genetics, 1(1), 41–73. http://doi.org/10.1146/annurev.genom.1.1.41 [53]
- Farre, D., & Alba, M. M. (2010). Heterogeneous Patterns of Gene-Expression Diversification in Mammalian Gene Duplicates. Molecular Biology and Evolution, 27(2), 325–335. http://doi.org/10.1093/molbev/msp242 [52]

Taxa-restricted genes

- Human-specific genes in fetal neocortex Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., et al. (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. eLife, 7, D635. http://doi.org/10.7554/eLife.32332 [49]
- Insects – Santos, M. E., Le Bouquin, A., Crumière, A. J. J., & Khila, A. (2017). Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. Science, 358(6361), 386–390. http://doi.org/10.1126/science.aan2748 [54]

Correlated evolution of celltypes?

- Liang, C., Musser, J. M., Cloutier, A., Prum, R. O., & Wagner, G. P. (2018). Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. Genome Biology and Evolution, 10(2), 538–552. http://doi.org/10.1093/gbe/evy016 [55]

Cell type homology

- Thornton, J. W., & DeSalle, R. (2000). Gene family evolution and homology: genomics meets phylogenetics. Annual Review of Genomics and Human Genetics, 1(1), 41–73. http://doi.org/10.1146/annurev.genom.1.1.41 [53]
- Tschopp, P., & Tabin, C. J. (2017). Deep homology in the age of next-generation sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1713), 20150475–8. http://doi.org/10.1098/rstb.2015.0475 [56]
- Hejnol, A., & Lowe, C. J. (2015). Embracing the comparative approach: how robust phylogenies and broader developmental sampling impacts the understanding of nervous system evolution.

Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1684), 20150045–16. http://doi.org/10.1098/rstb.2015.0045 [57]
- Santos, M. E., Le Bouquin, A., Crumière, A. J. J., & Khila, A. (2017). Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. Science, 358(6361), 386–390. http://doi.org/10.1126/science.aan2748 [54]
- Mammalian decidual cell

Cell type evolution

- Erkenbrack, E. M., Maziarz, J. D., Griffith, O. W., Liang, C., Chavan, A. R., Nnamani, M. C., & Wagner, G. P. (2018). The mammalian decidual cell evolved from a cellular stress response. PLOS Biology, 16(8), e2005594–27. http://doi.org/10.1371/journal.pbio.2005594 [58]

Supplemental Methods

Table 3: 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`.

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

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 · PMID: 21886479 · PMCID: PMC3160336

2. **Archives: Table of assemblies** https://uswest.ensembl.org/info/website/archives/assembly.html

3. **Animal Phylogeny and Its Evolutionary Implications**
Casey W. Dunn, Gonzalo Giribet, Gregory D. Edgecombe, Andreas Hejnol
Annual Review of Ecology, Evolution, and Systematics (2014-11-23) https://doi.org/gfkjbd
DOI: 10.1146/annurev-ecolsys-120213-091627

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 · PMID: 25383527 · PMCID: PMC4294627

5. **Standardized benchmarking in the quest for orthologs**
Adrian M AltenhoffBrigitte 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 · PMID: 27043882 · PMCID: PMC4827703

6. **Gearing up to handle the mosaic nature of life in the quest for orthologs**
Kristoffer Forslund, Cecile Pereira, Salvador Capella-Gutierrez, Alan Sousa da Silva, Adrian Altenhoff, Jaime Huerta-Cepas, Matthieu Muffato, Mateus Patricio, Klaas Vandepoele, Ingo Ebersberger, …
Bioinformatics (2017-08-30) https://doi.org/gc2cbq
DOI: 10.1093/bioinformatics/btx542 · PMID: 28968857 · PMCID: PMC5860199

7. **Advances and Applications in the Quest for Orthologs**
Natasha Glover, Christophe Dessimoz, Ingo Ebersberger, Sofia K Forslund, Toni Gabaldón, Jaime Huerta-Cepas, Maria-Jesus Martin, Matthieu Muffato, Mateus Patricio, Cécile Pereira, … Paul D Thomas
Molecular Biology and Evolution (2019-06-27) https://doi.org/ggcncc
DOI: 10.1093/molbev/msz150 · PMID: 31241141 · PMCID: PMC6759064

8. **OrthoFinder: phylogenetic orthology inference for comparative genomics**
David M. Emms, Steven Kelly
bioRxiv (2019-04-24) https://www.biorxiv.org/content/10.1101/466201v2
DOI: 10.1101/466201

9. **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 · PMID: 27818507

10. **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 · PMID: 28813177

11. **The evolution of gene expression levels in mammalian organs**
David Brawand, Magali Soumillon, Anamaria Necsulea, 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 · PMID: 22012392

12. **Integrating single-cell transcriptomic data across different conditions, technologies, and species**
Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, Rahul Satija
Nature Biotechnology (2018-04-02) https://doi.org/gc87v6
DOI: 10.1038/nbt.4096 · PMID: 29608179 · PMCID: PMC6700744

13. **Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity**
Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, Evan Z. Macosko
Cell (2019-06) https://doi.org/gf3m3v
DOI: 10.1016/j.cell.2019.05.006 · PMID: 31178122 · PMCID: PMC6716797

14. **Fast Batch Alignment of Single Cell Transcriptomes Unifies Multiple Mouse Cell Atlases into an Integrated Landscape**
Jong-Eun Park, Krzysztof Polański, Kerstin Meyer, Sarah A. Teichmann
Cold Spring Harbor Laboratory (2018-08-21) https://doi.org/ggfk7d
DOI: 10.1101/397042

15. **Panoramic stitching of heterogeneous single-cell transcriptomic data**
Brian Hie, Bryan Bryson, Bonnie Berger
Cold Spring Harbor Laboratory (2018-07-17) https://doi.org/gfv9k8
DOI: 10.1101/371179

16. **Fast, sensitive and accurate integration of single-cell data with Harmony**
Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, Soumya Raychaudhuri
Nature Methods (2019-11-18) https://doi.org/dfcg
DOI: 10.1038/s41592-019-0619-0 · PMID: 31740819 · PMCID: PMC6884693

17. **Integrative inference of brain cell similarities and differences from single-cell genomics**
Joshua Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, Evan Macosko
Cold Spring Harbor Laboratory (2018-11-02) https://doi.org/gfgr7b
DOI: 10.1101/459891

18. **Wiring together large single-cell RNA-seq sample collections**
Nikolas Barkas, Viktor Petukhov, Daria Nikolaeva, Yaroslav Lozinsky, Samuel Demharter, Konstantin Khodosevich, Peter V. Kharchenko
bioRxiv (2018-11-02) https://www.biorxiv.org/content/10.1101/460246v1
DOI: 10.1101/460246

19. **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/fr3gqz
DOI: 10.1093/nar/28.1.33 · PMID: 10592175 · PMCID: PMC102395

20. **Single-cell genomic atlas of great ape cerebral organoids uncovers human-specific features of brain development**
Sabina Kanton, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fatima Sanchis Calleja, Leila Sidow, Jonas Fleck, Patricia Guijarro, Dingding Han, … J. Gray Camp
bioRxiv (2019-06-27) https://www.biorxiv.org/content/10.1101/685057v1
DOI: 10.1101/685057

21. **Organoid single-cell genomic atlas uncovers human-specific features of brain development**
Sabina Kanton, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchís-Calleja, Patricia Guijarro, Leila Sidow, Jonas Simon Fleck, Dingding Han, … J. Gray Camp
Nature (2019-10) https://doi.org/dcws
DOI: 10.1038/s41586-019-1654-9 · PMID: 31619793

22. **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 · PMID: 30735633 · PMCID: PMC6544371

23. **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 · PMID: 29884752 · PMCID: PMC6028123

24. **A Single Cell Transcriptomic Atlas Characterizes Aging Tissues in the Mouse**
The Tabula Muris Consortium, Angela Oliveira Pisco, Aaron McGeever, Nicholas Schaum, Jim Karkanias, Norma F. Neff, Spyros Darmanis, Tony Wyss-Coray, Stephen R. Quake
bioRxiv (2019-11-18) https://www.biorxiv.org/content/10.1101/661728v2
DOI: 10.1101/661728

25. **Cell BLAST: Searching large-scale scRNA-seq databases via unbiased cell embedding**
Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang Yang, Ge Gao
bioRxiv (2019-04-04) https://www.biorxiv.org/content/10.1101/587360v2
DOI: 10.1101/587360

26. **CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA sequencing**
Kenta Sato, Koki Tsuyuzaki, Kentaro Shimizu, Itoshi Nikaido
Genome Biology (2019-02-11) https://doi.org/ggfk7f
DOI: 10.1186/s13059-019-1639-x · PMID: 30744683 · PMCID: PMC6371477

27. **Supervised classification enables rapid annotation of cell atlases**
Hannah A. Pliner, Jay Shendure, Cole Trapnell
Nature Methods (2019-09-09) https://doi.org/ggfk7c
DOI: 10.1038/s41592-019-0535-3 · PMID: 31501545 · PMCID: PMC6791524

28. **Unifying single-cell annotations based on the Cell Ontology**
Sheng Wang, Angela Oliveira Pisco, Jim Karkanias, Russ B. Altman
bioRxiv (2019-10-20) https://www.biorxiv.org/content/10.1101/810234v1
DOI: 10.1101/810234

29. **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

30. **SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier**
Xiao Hu, Iddo Friedberg
GigaScience (2019-10-01) https://doi.org/ggcr5x
DOI: 10.1093/gigascience/giz118 · PMID: 31648300 · PMCID: PMC6812468

31. **Fast databank searching with a reduced amino-acid alphabet**
Claudine Landès, Jean-Loup Risler
Bioinformatics (1994) https://doi.org/cvrjmw
DOI: 10.1093/bioinformatics/10.4.453 · PMID: 7804879

32. **RAPSearch: a fast protein similarity search tool for short reads**
Yuzhen Ye, Jeong-Hyeon Choi, Haixu Tang
BMC Bioinformatics (2011-05-15) https://doi.org/dgt5rt
DOI: 10.1186/1471-2105-12-159 · PMID: 21575167 · PMCID: PMC3113943

33. **Simplified amino acid alphabets for protein fold recognition and implications for folding**
Lynne Reed Murphy, Anders Wallqvist, Ronald M. Levy
Protein Engineering, Design and Selection (2000-03) https://doi.org/bdtngh
DOI: 10.1093/protein/13.3.149 · PMID: 10775656

34. **Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment**
Eric L. Peterson, Jané Kondev, Julie A. Theriot, Rob Phillips
Bioinformatics (2009-04-07) https://doi.org/btqmnp
DOI: 10.1093/bioinformatics/btp164 · PMID: 19351620 · PMCID: PMC2732308

35. **Local homology recognition and distance measures in linear time using compressed amino acid alphabets**
R. C. Edgar
Nucleic Acids Research (2004-01-02) https://doi.org/ckg5d4
DOI: 10.1093/nar/gkh180 · PMID: 14729922 · PMCID: PMC373290

36. **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets**
Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, … Steven A. McCarroll
Cell (2015-05) https://doi.org/f7dkxv
DOI: 10.1016/j.cell.2015.05.002 · PMID: 26000488 · PMCID: PMC4481139

37. **bam2fasta: tool for converting a bam file to fastas**
Pranathi Vemuri
https://github.com/pranathivemuri/bam2fasta

38. **The Sequence Alignment/Map format and SAMtools**
H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin,
Bioinformatics (2009-06-08) https://doi.org/ff6426
DOI: 10.1093/bioinformatics/btp352 · PMID: 19505943 · PMCID: PMC2723002

39. **MapReduce**
Jeffrey Dean, Sanjay Ghemawat
Communications of the ACM (2008-01-01) https://doi.org/br3wxw
DOI: 10.1145/1327452.1327492

40. **Ensembl 2018**
Daniel R Zerbino, Premanand Achuthan, Wasiu 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 · PMID: 29155950 · PMCID: PMC5753206

41. **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 · PMID: 26896847 · PMCID: PMC4761110

42. **Atlas of protein sequence and structure**
Margaret O Dayhoff
National Biomedical Research Foundation. (1969)

43. **Physical biology of the cell**
Rob Phillips, Julie Theriot, Jane Kondev, Hernan Garcia
Garland Science (2012)

44. **Theory for the folding and stability of globular proteins**
Ken A. Dill
Biochemistry (1985-03-12) https://doi.org/fnj5k7
DOI: 10.1021/bi00327a032 · PMID: 3986190

45. **The organization of domains in proteins obeys Menzerath-Altmann's law of language**
Khuram Shahzad, Jay E. Mittenthal, Gustavo Caetano-Anollés
BMC Systems Biology (2015-08-11) https://doi.org/f7s6rb
DOI: 10.1186/s12918-015-0192-9 · PMID: 26260760 · PMCID: PMC4531524

46. **Length Variations amongst Protein Domain Superfamilies and Consequences on Structure and Function**
Sankaran Sandhya, Saane Sudha Rani, Barah Pankaj, Madabosse Kande Govind, Bernard Offmann, Narayanaswamy Srinivasan, Ramanathan Sowdhamini
PLoS ONE (2009-03-31) https://doi.org/bz4bqf
DOI: 10.1371/journal.pone.0004981 · PMID: 19333395 · PMCID: PMC2659687

47. **Identifiers.org < EMBL-EBI** https://www.ebi.ac.uk/miriam/main/collections/MIR:00000119

48. **RNAsamba: coding potential assessment using ORF and whole transcript sequence information**
Antonio P. Camargo, Vsevolod Sourkov, Marcelo F. Carazzolle
Cold Spring Harbor Laboratory (2019-04-28) https://doi.org/ggdtxk
DOI: 10.1101/620880

49. **Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex**
Marta Florio, Michael Heide, Anneline Pinson, Holger Brandl, Mareike Albert, Sylke Winkler, Pauline

Wimberger, Wieland B Huttner, Michael Hiller
eLife (2018-03-21) https://doi.org/gc678k
DOI: 10.7554/elife.32332 · PMID: 29561261 · PMCID: PMC5898914

50. **Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques**
Pavel V. Mazin, Xi Jiang, Ning Fu, Dingding Han, Meng Guo, Mikhail S. Gelfand, Philipp Khaitovich
RNA (2018-01-23) https://doi.org/gctq4n
DOI: 10.1261/rna.064931.117 · PMID: 29363555 · PMCID: PMC5855957

51. **Predominant patterns of splicing evolution on human, chimpanzee and macaque evolutionary lineages**
Jieyi Xiong, Xi Jiang, Angeliki Ditsiou, Yang Gao, Jing Sun, Elijah D Lowenstein, Shuyun Huang, Philipp Khaitovich
Human Molecular Genetics (2018-02-14) https://doi.org/gc2v79
DOI: 10.1093/hmg/ddy058 · PMID: 29452398

52. **Heterogeneous Patterns of Gene-Expression Diversification in Mammalian Gene Duplicates**
D. Farre, M. M. Alba
Molecular Biology and Evolution (2009-10-12) https://doi.org/dxrtmd
DOI: 10.1093/molbev/msp242 · PMID: 19822635

53. **GENEFAMILYEVOLUTION ANDHOMOLOGY: Genomics Meets Phylogenetics**
Joseph W. Thornton, Rob DeSalle
Annual Review of Genomics and Human Genetics (2000-09) https://doi.org/bjp5pm
DOI: 10.1146/annurev.genom.1.1.41 · PMID: 11701624

54. **Taxon-restricted genes at the origin of a novel trait allowing access to a new environment**
M. Emília Santos, Augustin Le Bouquin, Antonin J. J. Crumière, Abderrahman Khila
Science (2017-10-19) https://doi.org/gcgjbs
DOI: 10.1126/science.aan2748 · PMID: 29051384

55. **Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes**
Cong Liang, Jacob M Musser, Alison Cloutier, Richard O Prum, Günter P Wagner
Genome Biology and Evolution (2018-01-23) https://doi.org/gc69v9
DOI: 10.1093/gbe/evy016 · PMID: 29373668 · PMCID: PMC5800078

56. **Deep homology in the age of next-generation sequencing**
Patrick Tschopp, Clifford J. Tabin
Philosophical Transactions of the Royal Society B: Biological Sciences (2017-02-05) https://doi.org/gfzpbg
DOI: 10.1098/rstb.2015.0475 · PMID: 27994118 · PMCID: PMC5182409

57. **Embracing the comparative approach: how robust phylogenies and broader developmental sampling impacts the understanding of nervous system evolution**
Andreas Hejnol, Christopher J. Lowe
Philosophical Transactions of the Royal Society B: Biological Sciences (2015-12-19) https://doi.org/ggcd2m
DOI: 10.1098/rstb.2015.0045 · PMID: 26554039 · PMCID: PMC4650123

58. **The mammalian decidual cell evolved from a cellular stress response**
Eric M. Erkenbrack, Jamie D. Maziarz, Oliver W. Griffith, Cong Liang, Arun R. Chavan, Mauris C.

Nnamani, Günter P. Wagner
PLOS Biology (2018-08-24) https://doi.org/gd5b9s
DOI: 10.1371/journal.pbio.2005594 · PMID: 30142145 · PMCID: PMC6108454