

# Smashing single cells into $k$ -mer sketches

This manuscript ([permalink](#)) was automatically generated from [czbiohub/primate-brain-organoid-paper@aae4c14](#) 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

- **Zebinisa Mirvoxid qizi Mirakbarova**

 [0000-0003-1164-7814](#) ·  [mzebinisa](#) ·  [ZebinisaM](#)

Data Sciences Platform, Chan Zuckerberg Biohub; Laboratory of Biotechnology. Centre for Advanced Technologies,  
Ministry of Innovative Development, Republic of Uzbekistan; Institute of Biophysics and Biochemistry, National  
University of Uzbekistan

## 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 0.000002% ( $200/8,700,000 = 2.3\text{e-}08$ ) 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.

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 [Cite: brawand2011, CCA, LIGER, Scanorama, basically all the single cell “alignment” packages], others use clusters of orthologous groups [11], 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 [12] and re-annotate using a tool such as Comparative Annotation Toolkit [13].

*k*-mers have been proposed for comparing single cells [14] 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.

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 [15] 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 existing 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 [16]. 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` [17]. Now we employ a MapReduce [18] 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

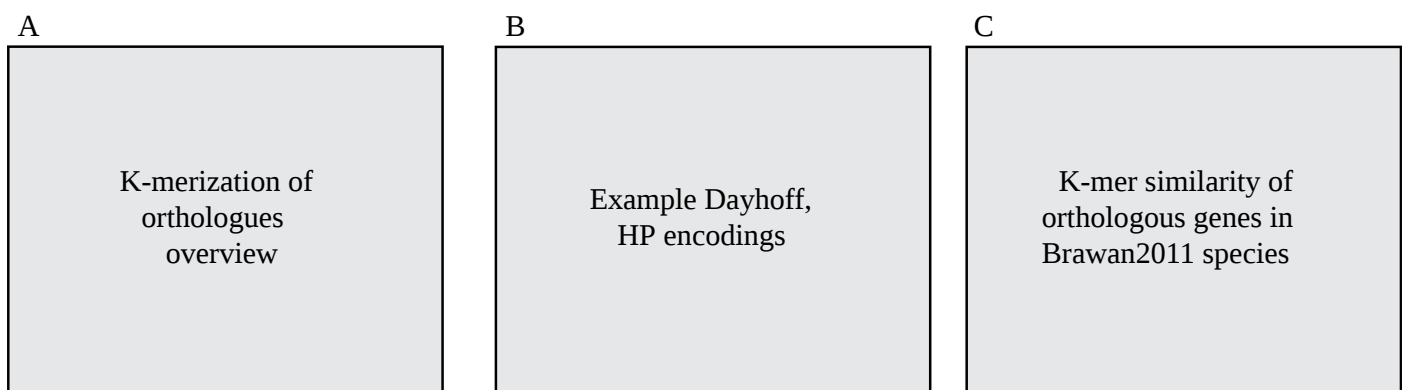
To determine whether short segments of sequences could detect gene orthologues, we  $k$ -merized orthologous genes derived from the ENSEMBL version 97 [19] COMPARA database [20] (Figure [1]). We compared human protein sequences to orthologous chimpanzee, mouse, (orangutan, bonobo, gorilla, macaque, opossum, platypus, chicken) protein sequences, as these are species used in [21]. 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 [22] and hydrophobic-polar encodings [23], show in Table [2].

**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 `bbdbdfecdac`, and HP-encoded to `phpphhpppp`.

Amino acid	Property	Dayhoff	Hydrophobic-polar (HP)
------------	----------	---------	------------------------

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 [[24](#),[25](#)].



**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.

## Outlines

### Figure 1 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 [26]
- Kmers can find correct reading frame of RNA-seq reads
  - Human peptides → human RNAseq
  - Human peptides → chimp RNAseq
  - Human peptides → mouse RNAseq
- Kmers can find only transcription factor reads of TFs from RNA-seq reads
  - Human TFs → human RNAseq
  - Human TFs → chimp RNAseq
  - Human TFs → mouse RNAseq
- 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

## Figure 2 outline

### 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> [27]
- 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> [28]
- 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> [29]

### No need for 1:1 orthology

### 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> [30]
- 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> [31]
- 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> [30]

### 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> [27]
- 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> [32]

Don't need to normalize gene expression counts since we just have presence/absence

Full transcript analyses

- Brain Organoid – Droplet + PacBio data

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> [33]

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> [31]
- 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> [34]
- Hejnal, 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> [35]
- 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> [32]
- 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> [36]

### Figure 3 – long evolutionary distances with HP encoding?

Metazoan body plan formation

- Early development in Cnidarians/Hydra [37,38]
- sponges and others [38]
- planaria [39,40],
- drosophila [41]
- zebrafish [42,43,44],
- mouse [45]



#### **Figure 4 – What features are k-mers able to pick up that mapping doesn't?**

- 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?

### **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. 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](https://doi.org/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](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. 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](https://doi.org/10.1093/bioinformatics/btx542) · PMID: [28968857](https://pubmed.ncbi.nlm.nih.gov/28968857/) · PMCID: [PMC5860199](https://pubmed.ncbi.nlm.nih.gov/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/ggcncx>  
DOI: [10.1093/molbev/msz150](https://doi.org/10.1093/molbev/msz150) · PMID: [31241141](https://pubmed.ncbi.nlm.nih.gov/31241141/) · PMCID: [PMC6759064](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/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](https://doi.org/10.1038/nrg.2016.127) · PMID: [27818507](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1146/annurev-cellbio-100616-060818) · PMID: [28813177](https://pubmed.ncbi.nlm.nih.gov/28813177/)

### 11. 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/)

### 12. 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/)

### 13. 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/)

### 14. 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)

### 15. 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](https://doi.org/10.1016/j.cell.2015.05.002) · PMID: [26000488](https://pubmed.ncbi.nlm.nih.gov/26000488/) · PMCID: [PMC4481139](https://pubmed.ncbi.nlm.nih.gov/PMC4481139/)

### 16. bam2fasta: tool for converting a bam file to fastas

Pranathi Vemuri

<https://github.com/pranathivemuri/bam2fasta>

### 17. 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](https://doi.org/10.1093/bioinformatics/btp352) · PMID: [19505943](https://pubmed.ncbi.nlm.nih.gov/19505943/) · PMCID: [PMC2723002](https://pubmed.ncbi.nlm.nih.gov/PMC2723002/)

### 18. MapReduce

Jeffrey Dean, Sanjay Ghemawat

*Communications of the ACM* (2008-01-01) <https://doi.org/br3wxw>

DOI: [10.1145/1327452.1327492](https://doi.org/10.1145/1327452.1327492)

### 19. 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](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/)

## 20. 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/)

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

## 22. Atlas of protein sequence and structure

Margaret O Dayhoff

*National Biomedical Research Foundation*. (1969)

## 23. Physical biology of the cell

Rob Phillips, Julie Theriot, Jane Kondev, Hernan Garcia

*Garland Science* (2012)

## 24. 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](https://doi.org/10.1186/s12918-015-0192-9) · PMID: [26260760](https://pubmed.ncbi.nlm.nih.gov/26260760/) · PMCID: [PMC4531524](https://pubmed.ncbi.nlm.nih.gov/PMC4531524/)

## 25. 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](https://doi.org/10.1371/journal.pone.0004981) · PMID: [19333395](https://pubmed.ncbi.nlm.nih.gov/19333395/) · PMCID: [PMC2659687](https://pubmed.ncbi.nlm.nih.gov/PMC2659687/)

## 26. Identifiers.org < EMBL-EBI <https://www.ebi.ac.uk/miriam/main/collections/MIR:00000119>

## 27. 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](https://doi.org/10.7554/elife.32332) · PMID: [29561261](https://pubmed.ncbi.nlm.nih.gov/29561261/) · PMCID: [PMC5898914](https://pubmed.ncbi.nlm.nih.gov/PMC5898914/)

## 28. 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](https://doi.org/10.1261/rna.064931.117) · PMID: [29363555](https://pubmed.ncbi.nlm.nih.gov/29363555/) · PMCID: [PMC5855957](https://pubmed.ncbi.nlm.nih.gov/PMC5855957/)

## 29. 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](https://doi.org/10.1093/hmg/ddy058) · PMID: [29452398](https://pubmed.ncbi.nlm.nih.gov/29452398/)

**30. 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](https://doi.org/10.1093/molbev/msp242) · PMID: [19822635](https://pubmed.ncbi.nlm.nih.gov/19822635/)

**31. 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](https://doi.org/10.1146/annurev.genom.1.1.41) · PMID: [11701624](https://pubmed.ncbi.nlm.nih.gov/11701624/)

**32. 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](https://doi.org/10.1126/science.aan2748) · PMID: [29051384](https://pubmed.ncbi.nlm.nih.gov/29051384/)

**33. 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](https://doi.org/10.1093/gbe/evy016) · PMID: [29373668](https://pubmed.ncbi.nlm.nih.gov/29373668/) · PMCID: [PMC5800078](https://pubmed.ncbi.nlm.nih.gov/PMC5800078/)

**34. 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/gfzpbj>

DOI: [10.1098/rstb.2015.0475](https://doi.org/10.1098/rstb.2015.0475) · PMID: [27994118](https://pubmed.ncbi.nlm.nih.gov/27994118/) · PMCID: [PMC5182409](https://pubmed.ncbi.nlm.nih.gov/PMC5182409/)

**35. 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](https://doi.org/10.1098/rstb.2015.0045) · PMID: [26554039](https://pubmed.ncbi.nlm.nih.gov/26554039/) · PMCID: [PMC4650123](https://pubmed.ncbi.nlm.nih.gov/PMC4650123/)

**36. 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](https://doi.org/10.1371/journal.pbio.2005594) · PMID: [30142145](https://pubmed.ncbi.nlm.nih.gov/30142145/) · PMCID: [PMC6108454](https://pubmed.ncbi.nlm.nih.gov/PMC6108454/)

**37. Stem cell differentiation trajectories in Hydra resolved at single-cell resolution**

Stefan Siebert, Jeffrey A. Farrell, Jack F. Cazet, Yashodara Abeykoon, Abby S. Primack, Christine E. Schnitzler, Celina E. Juliano

*Science* (2019-07-25) <https://doi.org/ggcd2p>

DOI: [10.1126/science.aav9314](https://doi.org/10.1126/science.aav9314) · PMID: [31346039](https://pubmed.ncbi.nlm.nih.gov/31346039/)

**38. Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq**

Arnau Sebé-Pedrós, Baptiste Saudemont, Elad Chomsky, Flora Plessier, Marie-Pierre Mailhé, Justine

Renno, Yann Loe-Mie, Aviezer Lifshitz, Zohar Mukamel, Sandrine Schmutz, ... Heather Marlow  
*Cell* (2018-05) <https://doi.org/gdkxf5>  
DOI: [10.1016/j.cell.2018.05.019](https://doi.org/10.1016/j.cell.2018.05.019) · PMID: [29856957](https://pubmed.ncbi.nlm.nih.gov/29856957/)

**39. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea***

Christopher T. Fincher, Omri Wurtzel, Thom de Hoog, Kellie M. Kravarik, Peter W. Reddien  
*Science* (2018-04-19) <https://doi.org/ggacd2n>  
DOI: [10.1126/science.aag1736](https://doi.org/10.1126/science.aag1736) · PMID: [29674431](https://pubmed.ncbi.nlm.nih.gov/29674431/) · PMCID: [PMC6563842](https://pubmed.ncbi.nlm.nih.gov/PMC6563842/)

**40. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics**

Mireya Plass, Jordi Solana, F. Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J. Theis, Christine Kocks, Nikolaus Rajewsky  
*Science* (2018-04-19) <https://doi.org/gfgtbt>  
DOI: [10.1126/science.aag1723](https://doi.org/10.1126/science.aag1723) · PMID: [29674432](https://pubmed.ncbi.nlm.nih.gov/29674432/)

**41. The *Drosophila* embryo at single-cell transcriptome resolution**

Nikos Karaikos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, Robert P. Zinzen  
*Science* (2017-08-31) <https://doi.org/gbvp2z>  
DOI: [10.1126/science.aan3235](https://doi.org/10.1126/science.aan3235) · PMID: [28860209](https://pubmed.ncbi.nlm.nih.gov/28860209/)

**42. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis**

Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, Alexander F. Schier  
*Science* (2018-04-26) <https://doi.org/gdggtf>  
DOI: [10.1126/science.aar3131](https://doi.org/10.1126/science.aar3131) · PMID: [29700225](https://pubmed.ncbi.nlm.nih.gov/29700225/) · PMCID: [PMC6247916](https://pubmed.ncbi.nlm.nih.gov/PMC6247916/)

**43. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution**

James A. Briggs, Caleb Weinreb, Daniel E. Wagner, Sean Megason, Leonid Peshkin, Marc W. Kirschner, Allon M. Klein  
*Science* (2018-04-26) <https://doi.org/gdgg4d>  
DOI: [10.1126/science.aar5780](https://doi.org/10.1126/science.aar5780) · PMID: [29700227](https://pubmed.ncbi.nlm.nih.gov/29700227/) · PMCID: [PMC6038144](https://pubmed.ncbi.nlm.nih.gov/PMC6038144/)

**44. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo**

Daniel E. Wagner, Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, Allon M. Klein  
*Science* (2018-04-26) <https://doi.org/gdgg3>  
DOI: [10.1126/science.aar4362](https://doi.org/10.1126/science.aar4362) · PMID: [29700229](https://pubmed.ncbi.nlm.nih.gov/29700229/) · PMCID: [PMC6083445](https://pubmed.ncbi.nlm.nih.gov/PMC6083445/)

**45. The single-cell transcriptional landscape of mammalian organogenesis**

Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, ... Jay Shendure  
*Nature* (2019-02) <https://doi.org/gfvqww>  
DOI: [10.1038/s41586-019-0969-x](https://doi.org/10.1038/s41586-019-0969-x) · PMID: [30787437](https://pubmed.ncbi.nlm.nih.gov/30787437/) · PMCID: [PMC6434952](https://pubmed.ncbi.nlm.nih.gov/PMC6434952/)