INCOMPLETE DRAFT: Identification of *de novo* orthologous genes from comparative single-cell RNA-seq transcriptomics

This manuscript (permalink) was automatically generated from czbiohub/de-novo-orthology-paper@c49b752 on March 23, 2020.

Authors

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

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

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

- **Saba Nafees**
  iD 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

- **Jim Karkanias**
  iD 0000-0002-8057-6055 · jkensai · jkarkanias
  Data Sciences Platform, Chan Zuckerberg Biohub

Abstract

We introduce `kmermaid`, a novel computational method for identifying orthologous cell types discovering *de novo* orthologous genes across species. As `kmermaid` skips both traditional alignment and gene orthology assignment it can a) be applied to transcriptomes from organisms with no or poorly annotated genomes, b) predicts protein-coding sequences from raw RNA-seq reads, and c) identify putative functions of protein sequences contributing to shared cell types. By enabling analyses across divergent species' transcriptomes in an orthology-, genome- and gene annotation-agnostic manner, `kmermaid` illustrates the potential of non-model organisms in building the cell type evolutionary tree of life.

Identifying orthologous genes across species remains an open problem. We show how orthologous genes can be identified directly from RNA-seq reads of tissue and cell types that are shared across species.

Single-cell RNA-sequencing is a powerful technology for identifying cell types in a variety of species. However, the task of identifying even known cell types in species with poorly annotated genomes is nontrivial, as 99.999% of the predicted 8.7 million Eukaryotic species on Earth have no submitted genome assembly [1,2] and identifying orthologous genes, which remains an open problem [3,4].

Typesetting math: 100% et need to quantitatively compare single-cell transcriptomes across species, without the need for orthologous gene mapping, gene annotations, or a reference genome. Short, k-long sequence substrings, or k-mers, have been proposed for clustering single cells [5] and here we implemented k-mers from putatitvely translated RNA-seq reads with reduced amino acid alphabets [**???**,6,6,7,8], to find shared cell types across species, and further identify *de novo* orthologous genes by querying the predicted protein sequences to a reference database. This method relies solely on divergence time between species, which we show can be estimated from RNA-seq nucleotide k-mers (Supplemental Figure [**???**]). We benchmark the genome-agnostic method on the Quest for Orthologs Opisthokonta dataset, showing that k-mers from reduced amino acid alphabets are sufficient to estimate orthology. Using human amino acid sequences, we show that one can extract putative protein-coding reads from 239 Opisthokonta species in ENSEMBL, and present the best k-mer size and alphabet for different divergence times. We first apply this method on a bulk comparative transcriptomic dataset consisting of nine amniote species and six tissues [11], showing that we achieve similar clustering results as using only reads mapping to 1:1 orthologs or Hierarchical Orthologous Groups (HOGs) [12,13,14] of protein-coding genes, but are able to resolve … which can only be seen by using the k-mer method. We further demonstrate the utility of this method by comparing transcriptomes from organisms diverged by approximately 676 million years [15]: a single-cell atlas of a model organism, mouse from *Tabula Muris Senis* [16], and bulk RNA-seq from *Botryllus schlosseri* [17], a colonial tunicate which exhibits cell populations similar to the myeloid immune lineage. Across this evolutionary distance, only XX 1:1 orthologous genes exist as found by … and XX HOGs via orthologous matrix (OMA) [18,19] We show that the myeloid-like cells from *B. schlosseri* not only cluster with the myeloid immune cells from *Tabula Muris Senis*, we also find *de novo* orthologous genes, such as … We find that using k-mers has the advantage of resolving … in comparison to using read counts from 1:1 gene orthologs. Using k-mers, we were able to resolve cell types … , which was hidden using read counts alone. Thus, we have shown the reference-free method 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 evolutionary tree of life.

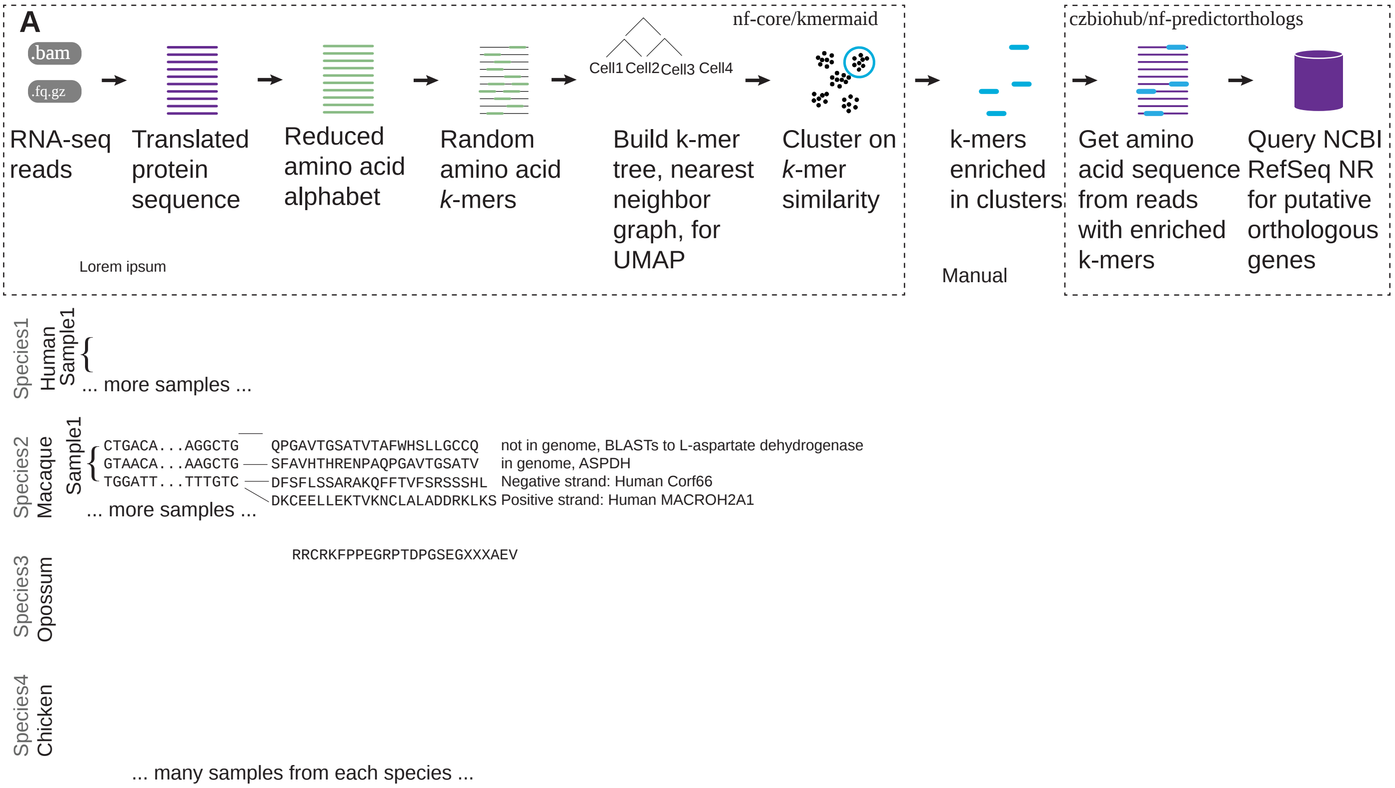

A
nf-core/kmermaid
czbiohub/nf-predictorthologs
.bam
.fq.gz
RNA-seq reads
Translated protein sequence
Reduced amino acid alphabet
Random amino acid k-mers
Cell1 Cell2 Cell3 Cell4
Build k-mer tree, nearest neighbor graph, for UMAP
Cluster on k-mer similarity
k-mers enriched in clusters
Get amino acid sequence from reads with enriched k-mers
Query NCBI RefSeq NR for putative orthologous genes
Lorem ipsum
Manual
Species1
Human
Sample1
... more samples ...
Species2
Macaque
Sample1
CTGACA...AGGCTG
GTAACA...AAGCTG
TGGATT...TTTGTC
QPGAVTGSATVTAFWHSLLGCCQ
SFAVHTHRENPAQPGAVTGSATV
DFSFLSSARAKQFFTVFSRSSSHL
DKCEELLEKTVKNCLALADDRKLKS
not in genome, BLASTs to L-aspartate dehydrogenase
in genome, ASPDH
Negative strand: Human Corf66
Positive strand: Human MACROH2A1
... more samples ...
RRCRKFPPEGRPTDPGSEGXXXAEV
Species3
Opossum
Species4
Chicken
... many samples from each species ...

Figure 1: A. Overview of the `kmermaid` pipeline. (**a**, **b**, **c**) `kmermaid` consists of a protein-coding prediction phase (**a**) that is invoked by the command `khtools extract_coding`, a k-mer sketch computation phase (**b**) invoked by the command `sourmash sketch`, a signature similarity comparison phase (**c**) invoked by the command `sourmash compare`, and an optional database-creation phase (**d**) invoked by the command `sourmash index`. The coding

Typesetting math: 100%

e components: (1) six-frame translation, removal of stop-codon frames, and subsequent k-

merization of RNA-sequencing reads; (2) a degenerate protein alphabet which allows for protein-coding detection from a wide variety of species; (3) a bloom filter containing known protein-coding sequences from a well annotated organism; and (4) computation of the Jaccard index of translated RNA-seq reading frames. The sketch computation phase involves randomly subsetting the degenerate peptide k-mers using a MinHash algorithm. The sketch comparison phase consists of computing the Jaccard intersection of MinHashed degenerate peptide k-mers between all pairs of samples.

To determine whether short segments of sequences could detect gene orthologues, we k-merized orthologous genes derived from the ENSEMBL version 97 [20] COMPARA database [21] (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 [22], a nine-letter encoding [9], and a two-letter hydrophobic-polar encodings [23,24], show in Table [1].

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 [9], a k-mer based orthology relationship finder.

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

Across tissues of the same time from the Brawand 2011 [11] dataset, we extracted protein-coding sequences, generated dayhoff signatures of k-mer size length 12, extracted hashes and thus k-mers shared by samples from the same tissue, went back to the original protein sequence, and searched NCBI RefSeq NR for potential proteins. For each sample, we observed that shared k-mers appeared in 1:1 orthologous genes XX% of the time 1:many orthologs YY% of the time, many:many orthologs ZZ% of the time, in genes not known to be orthologs AA% of the time, in unannotated regions AA% of the time, in multimapped reads BB% of the time, and in unmapped reads CC% of the time. Overall, we observed XX de novo orthologs in each tissue. We removed genes that were already known to be orthol

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 [25]
        - Orthologous groups/Conserved Domain Database [26]

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 [27]

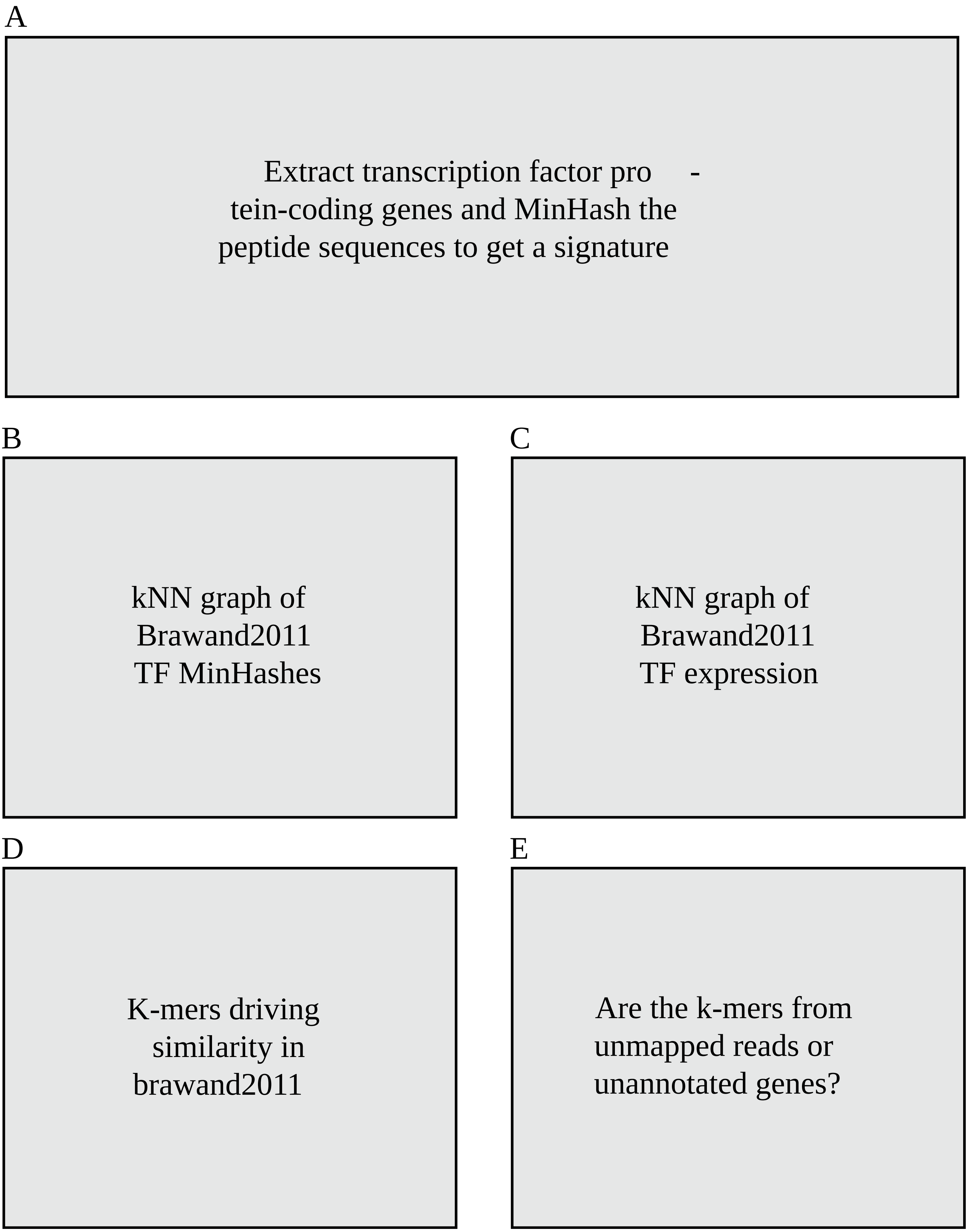

A
Extract transcription factor protein-coding genes and MinHash the peptide sequences to get a signature
B
kNN graph of Brawand2011 TF MinHashes
C
kNN graph of Brawand2011 TF expression
D
K-mers driving similarity in brawand2011
E
Are the k-mers from unmapped reads or unannotated genes?

Figure 3: Figure 3

Typesetting math: 100%

- 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

`kmermaid` implements the concept of lightweight orthology assignment using k-mers to the problem of cross-species RNA-seq analyses and achieves unprecedented speed of analysis. By removing the orthology inference step, `kmermaid` opens up the possibilty of finding shared and divergent tissue and cell types across a broad range of species, paving the way for evolutionary analyses of cell types across species. `kmermaid` can be used in *de novo* setting for non-model organisms, finding similar cell types within an organism, or finding similar cell types relative to a reference organism, without the need for a reference genome or transcriptome. The memory usage of `kmermaid` is quite low, using only 50MB for extracting coding sequences and 50MB for assigning protein k-mer signatures. As the number of RNA-seq datasets, especially single-cell RNA-seq datasets continues to grow, we expect `kmermaid` to be widely used for identifying cell types in non-model organisms.

`kmermaid` is free and open-source software and is available as Supplementary Data and at http://github.com/czbiohub/kmermaid and as a scalable Nextflow workflow at http://github.com/nf-core/nf-kmermaid.

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 [28]
- 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 [29]
- 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 [28]

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 [30]
- 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 [31]

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 [32]

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. 46/annurev.genom.1.1.41 [29]

- 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 [33]
- 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 [34]
- 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 [31]
- 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 [35]

In summary, we developed a method to identify both known cell types in a non-model organism using a reference atlas from another organism, without the need for a genome or gene annotation from the non-model organism. This method can be used to combine single-cell cell atlases from well-annotated, model organisms, with sequencing data from poorly annotated non-model organisms, to directly find homologous cell types and orthologous genes. By eliminating read alignment and orthologous gene mapping, `kmermaid` enables comparison of transcriptomes of the remaining 99.999% Eukaryotic species on Earth without submitted genome assemblies, with the cell atlases of a handful of model organisms to identify shared and novel cell types, and *de novo* identify orthologous genes. By identifying homologous cell types across a broad variety of species, we come closer to an understanding of the evolution of genes, cells, and thus life itself.

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: [25]. Biorxiv example: [36]. Multiple citations per line example: [25,36].

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: [25]. Biorxiv example: [36]. Multiple citations per line example: [25,36].

Long read library prep

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [25]. Biorxiv example: [36]. Multiple citations per line example: [25,36].

Short read library prep

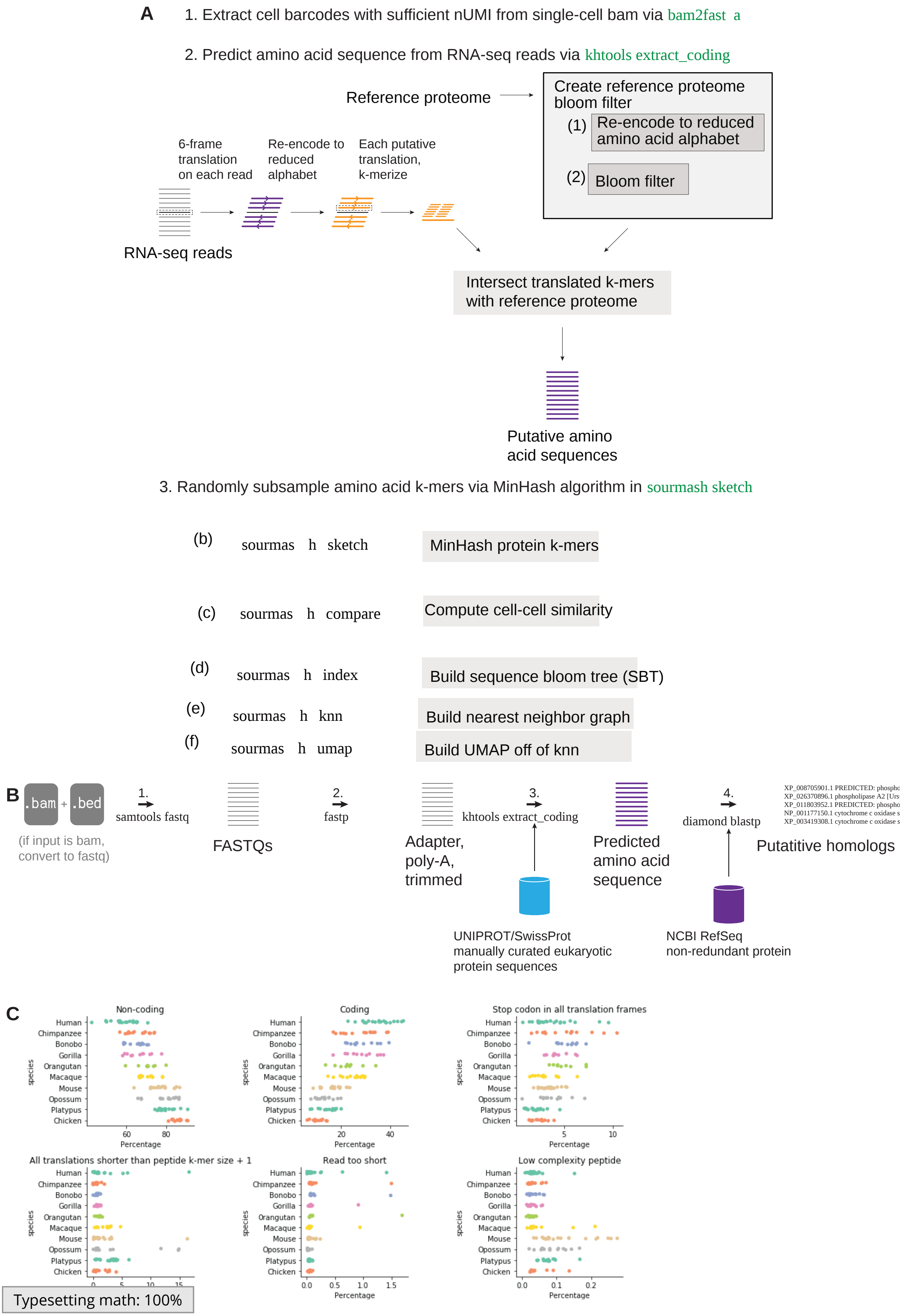
We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [25]. Biorxiv example: [36]. Multiple citations per line example: [25,36].

Sequencing

We did things. One sentence per line. Prefer DOI for references, but for Biorxiv use the URL. DOI example: [25]. Biorxiv example: [36]. Multiple citations per line example: [25,36].

Computational

Typesetting math: 100%


A
1. Extract cell barcodes with sufficient nUMI from single-cell bam via bam2fast a
2. Predict amino acid sequence from RNA-seq reads via khtools extract_coding
Reference proteome
Create reference proteome bloom filter
(1) Re-encode to reduced amino acid alphabet
(2) Bloom filter
6-frame translation on each read
Re-encode to reduced alphabet
Each putative translation, k-merize
RNA-seq reads
Intersect translated k-mers with reference proteome
Putative amino acid sequences
3. Randomly subsample amino acid k-mers via MinHash algorithm in sourmash sketch
(b) sourmas h sketch
MinHash protein k-mers
(c) sourmas h compare
Compute cell-cell similarity
(d) sourmas h index
Build sequence bloom tree (SBT)
(e) sourmas h knn
Build nearest neighbor graph
(f) sourmas h umap
Build UMAP off of knn
B
.bam + .bed
(if input is bam, convert to fastq)
1. samtools fastq
FASTQs
2. fastp
Adapter, poly-A, trimmed
3. khtools extract_coding
UNIPROT/SwissProt manually curated eukaryotic protein sequences
Predicted amino acid sequence
4. diamond blastp
NCBI RefSeq non-redundant protein
XP_008705901.1 PREDICTED: phospho
XP_026370896.1 phospholipase A2 [Urs
XP_011803952.1 PREDICTED: phospho
NP_001177150.1 cytochrome c oxidase s
XP_003419308.1 cytochrome c oxidase s
Putatitive homologs
C
Non-coding
Coding
Stop codon in all translation frames
All translations shorter than peptide k-mer size + 1
Read too short
Low complexity peptide
species
Human
Chimpanzee
Bonobo
Gorilla
Orangutan
Macaque
Mouse
Opossum
Platypus
Chicken
Percentage
60
80
20
40
5
10
0
15
0.0
0.5
1.0
1.5
0.1
0.2

Typesetting math: 100%

A. Overview of `nf-core/kmermaid` pipeline. 1. If input is bam, extract per-cell sequences. 2. Predict amino acid sequence of each RNA-seq read using `khtools extract-coding`. 3. Randomly subsample amino acid k-mers via MinHash using `sourmash sketch`. 4. Compare all k-mer sketches to one another using `sourmash compare` to compute cell-cell Jaccard similarities. 5. Build sequence bloom tree using `sourmash index`. 6. Build k-nearest neighbor graph using sequence bloom tree. 7. Build UMAP off of KNN. **B.** Overview of `czbiohub/nf-predictorthologs` pipeline for prediction of homologous genes from sequences. 1. If input is bam, must also have a convert bam reads to raw fastq files using the `samtools fastq` subcommand (samtools version 1.9). If input is fastqs, go directly to second step. 2. Trim adapters, poly-A, polyG using the `fastp` tool. 3. Predict protein-coding sequence using khtools extract_coding, using conservative UniProt/SwissProt manually curated database as examples of known protein-coding sequences, for most stringent definition of protein-coding. 4. Query predicted protein in permissive NCBI RefSeq non-redundant protein database for most complete search query. **C.** Example of predicting protein-coding sequence using Brawand2011 RNA-seq data, and human proteome as the reference. x-axis, percentage of reads falling into that category, y-axis, the species which the reads are from.

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: [25]. Biorxiv example: [36]. Multiple citations per line example: [25,36].

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: [25]. Biorxiv example: [36]. Multiple citations per line example: [25,36].

Supplemental Methods

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`.

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. **Genome List - Genome - NCBI** https://www.ncbi.nlm.nih.gov/genome/browse

3. **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

4. **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

5. **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

6. **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

7. **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

8. **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

9. **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

10. **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

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. **Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs**
Adrian M. Altenhoff, Manuel Gil, Gaston H. Gonnet, Christophe Dessimoz
PLoS ONE (2013-01-14) https://doi.org/ggkv2j
DOI: 10.1371/journal.pone.0053786 · PMID: 23342000 · PMCID: PMC3544860

13. **Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees**
B. Boeckmann, M. Robinson-Rechavi, I. Xenarios, C. Dessimoz
Briefings in Bioinformatics (2011-07-07) https://doi.org/c78rwm
DOI: 10.1093/bib/bbr034 · PMID: 21737420 · PMCID: PMC3178055

14. **Big data and other challenges in the quest for orthologs**
E. L. L. Sonnhammer, T. Gabaldon, A. W. Sousa da Silva, M. Martin, M. Robinson-Rechavi, B. Boeckmann, P. D. Thomas, C. Dessimoz,
Bioinformatics (2014-07-26) https://doi.org/f6ntvb
DOI: 10.1093/bioinformatics/btu492 · PMID: 25064571 · PMCID: PMC4201156

15. **TimeTree :: The Timescale of Life**http://timetree.org/

16. **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

17. **Complex mammalian-like haematopoietic system found in a colonial chordate**
Benyamin Rosental, Mark Kowarsky, Jun Seita, Daniel M. Corey, Katherine J. Ishizuka, Karla J. Palmeri, Shih-Yu Chen, Rahul Sinha, Jennifer Okamoto, Gary Mantalas, … Ayelet Voskoboynik
Nature (2018-12) https://doi.org/gfkzvm
DOI: 10.1038/s41586-018-0783-x · PMID: 30518860 · PMCID: PMC6347970

18. **Assigning confidence scores to homoeologs using fuzzy logic**
Natasha M. Glover, Adrian Altenhoff, Christophe Dessimoz
PeerJ (2019-01-11) https://doi.org/ggkv2k
DOI: 10.7717/peerj.6231 · PMID: 30648004 · PMCID: PMC6330999

19. **Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference**
Clément-Marie Train, Natasha M Glover, Gaston H Gonnet, Adrian M Altenhoff, Christophe Dessimoz
Bioinformatics (2017-07-12) https://doi.org/ggkv2h
DOI: 10.1093/bioinformatics/btx229 · PMID: 28881964 · PMCID: PMC5870696

20. **Ensembl 2018**
Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish
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

21. **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

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. **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

25. **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

26. https://www.ebi.ac.uk/miriam/main/collections/MIR:00000119

27. **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

28. **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

29. **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

30. **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
Typesetting math: 100% 32332 · PMID: 29561261 · PMCID: PMC5898914

31. **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

32. **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

33. **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

34. **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

35. **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

36. **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